Selective Sampling for Sensor Type Classification in Buildings

Jing Ma[1], Dezhi Hong[2], Hongning Wang[1]
[1]University of Virginia, Charlottesville, VA, USA 22904
[2]University of California San Diego, La Jolla, CA, USA 92093
jm3mr@virginia.edu,dehong@ucsd.edu,hw5x@virginia.edu

ABSTRACT

A key barrier to applying any smart technology to a building is the requirement of locating and connecting to the necessary resources among the thousands of sensing and control points, i.e., the metadata mapping problem. Existing solutions depend on exhaustive manual annotation of sensor metadata — a laborious, costly, and hardly scalable process. To reduce the amount of manual effort required, this paper presents a multi-oracle selective sampling framework to leverage noisy labels from information sources with unknown reliability such as existing buildings, which we refer to as weak oracles, for metadata mapping. This framework involves an interactive process, where a small set of sensor instances are progressively selected and labeled for it to learn how to aggregate the noisy labels as well as to predict sensor types.

Two key challenges arise in designing the framework, namely, weak oracle reliability estimation and instance selection for querying. To address the first challenge, we develop a clustering-based approach for weak oracle reliability estimation to capitalize on the observation that weak oracles perform differently in different groups of instances. For the second challenge, we propose a disagreement-based query selection strategy to combine the potential effect of a labeled instance on both reducing classifier uncertainty and improving the quality of label aggregation. We evaluate our solution on a large collection of real-world building sensor data from 5 buildings with more than 11, 000 sensors of 18 different types. The experiment results validate the effectiveness of our solution, which outperforms a set of state-of-the-art baselines.

KEYWORDS

Selective sampling, label aggregation, sensor type classification, smart buildings

1 INTRODUCTION

Hundreds of organizations have participated in the Better Buildings Initiative [10] to reduce the energy footprints of commercial buildings; and the investment in smart building technologies has soared from 1 billion to 19 billion dollars since 2012 [36]. Despite being effective in reducing energy consumption, these technologies [2, 5, 13, 16] are adopted still in less than 20% of the buildings [42]. A key barrier to applying any smart technology to a building is the requirement of locating and connecting to the necessary resources among the thousands of sensing and control points. For example, an application that monitors and controls the temperature of a room needs to access the temperature sensor and the temperature setpoint of the room. Doing so requires the capability to interpret the context of the sensors including their type, location, etc, often referred to as the *metadata*. Unfortunately, this metadata is historically not designed for automated machine parsing and exists in disparate formats across buildings. For example, Table 1 shows a few examples of room temperature sensor names from different buildings: the concept of temperature is encoded with various distinct phrases – `Temperature`, `Temp`, `RMT`, `ART`, and `ZN-T`. The metadata thus requires significant manual effort to parse, and it often takes weeks. This manual process is fundamentally not scalable, and calls for automated mapping solutions.

Table 1: Examples of sensor names of temperature sensors in different buildings.

Building	Sensor Name
A	RM511A Zone Temp 3
	RMI1071 Space Temperature Local
B	SDH_SF1_R282_RMT
	SDB_KETI_413_temperature
C	SODA1C600A_ART
	SODA1R438__ART
D	EBU3B.RM-B215..ZN-T
	EBU3B.Unknown..ZN-T
E	AP_M.RM-B301.ZN-T
	RM-5441.ZN-T

While industry-wide standards [1, 3] provide a common ground for creating metadata in new buildings, legacy buildings still take a significant share of the market and require manual work for metadata interpretation and conversion. Various solutions have thus been proposed to parse and extract the metadata, including their type [4, 18, 19, 22, 37], location [20, 23], and relationships with each other [24, 33, 38]. While a few [18, 39] focus on supervised approaches that require a considerable set of training examples, the majority [4, 7, 22, 25] is built upon a semi-supervised technique – active learning. Such solutions involve progressively selecting a small set of representative examples and acquiring their labels[1]. These methods have achieved promising results and can significantly reduce the required manual effort. However, they all fundamentally rely on the availability of infallible experts, such as a building manager, to provide correct labels for a small set of sensor metadata. Not only is it almost never possible not to make mistakes as a human, but also the cost of employing a human expert is prohibitive.

In this work, we seek to relax the strict dependence on an infallible human labeler as explored by the aforementioned work, and explore the value of imperfect information sources for metadata extraction. A key intuition is that, aside from resorting to a costly human annotator, there are often abundant cost-free information sources, such as classifiers trained on another labeled building, which we can leverage as *weak oracles* to annotate the sensors in

[1]For example, the *label* for "`RM511A Zone Temp 3`" would be *temperature sensor.*

978-1-7281-5497-8/20/$31.00 ©2020 IEEE
DOI 10.1109/IPSN48710.2020.00028

the target building. The term oracle refers to an information source; and leveraging such low-cost weak oracles helps reduce the dependence and burden on perfect human annotators. Technically, it is a branch of *multi-oracle selective sampling* problem [11], which extends conventional active learning by leveraging weak oracles aside from an omnipotent expert.

In practice of metadata mapping, each weak oracle is likely to be good at recognizing only certain types of sensors in a new building. For example, buildings in similar geographical regions might exhibit similar patterns in the readings of sensors measuring ambient temperature or light level, while in buildings with similar equipment configurations, similar operation patterns might manifest. Therefore, when leveraging different types of weak oracles, we would need to identify their output's trustworthiness in the new building, i.e., what types of sensors they can confidently predict for. However, as we do not know the best matching between weak oracles and sensors in a new building beforehand, it remains a challenge to uncover the underlying groups and further measure each weak oracle's reliability in predicting for each group.

In this paper, we propose an iterative algorithm that synergizes a *strong oracle* (e.g., a human expert) and multiple *weak oracles* (e.g., classifiers trained on existing labeled buildings) to further reduce the manual effort required for extracting metadata. We specifically focus on inferring a key kind of metadata – sensor type (e.g., temperature, co2, airflow volume, etc). In particular, in addition to employing a classical active learning procedure which iteratively selects an example for manual labeling, in each iteration, we also use label from the strong oracle to help estimate each weak oracle's reliability in the building.

While both strong and weak oracles are involved, it is noteworthy that our work is not a trivial combination of these two sources of information. Particularly, we need to address two major challenges. First, each weak oracle might have different predictive capabilities in different groups of sensors. However, existing methods evaluate the performance of weak oracles globally, which limits our use of knowledge of weak oracles about different sensors. To address this challenge, we take a divide-and-conquer approach to estimate each weak oracle's reliability in different clusters of sensors, and identify the clusters on the fly. For each sensor, we then *aggregate* all the noisy labels from weak oracles into one most probable label based on their reliability. Second, the criterion for selecting instances for labeling relies on two different components – the classifier trained using labels from strong oracle and the labels aggregated from weak oracles – whose objectives are often met in isolation. The query to strong oracle usually depends only on the informativeness of a sample considering its features, while label aggregation evaluates the informativeness only based on weak oracles' responses. To reconcile these different targets, we make two key considerations into selecting an instance for labeling — how much the instance benefits the classifier training and also how much it improves the reliability estimation of the weak oracles. In this way, our method distinguishes itself from prior strategies that focus either on only improving classifier learning [8, 11, 12, 17, 22], or on estimating the reliability of weak oracles [35, 40]. As we accumulate more labels from the strong oracle, we can progressively improve our estimate of the weak oracle's reliability, and therefore more effectively combine their strengths across different groups of sensors to obtain more accurate labels. On the other hand, the improved aggregated label in turn boosts classifier training for the new building.

We demonstrate the effectiveness of our approach by evaluating it on a real-world benchmark building dataset [26][2], which contains one-week data for over 11,000 sensors in 5 commercial buildings across 3 college campuses. We conduct extensive experiments and compare our proposed approach with a set of state-of-the-art solutions. The experiment results show that our proposed method performs significantly better on sensor type classification, i.e., with much fewer human labels required to achieve the same level of accuracy. Our main contributions can be summarized as follows:

- We address the problem of building sensor type classification by proposing a selective sampling framework which leverages the noisy labels collected from multiple weak oracles.
- We develop a clustering-based estimation method to identify each weak oracle's reliability in different groups of sensors, which delivers deeper insights about the expertise of weak oracles, and thus benefits label aggregation.
- We explore a disagreement-based strategy to select the most useful instances by jointly considering the representativeness of the instance and the disagreement between the aggregated labels and the classifier's prediction.

2 BACKGROUND AND RELATED WORK

In this section, we introduce the problem of sensor metadata inference in buildings, and the recent advances on this topic. This serves as the basis for our developed solution in this work.

2.1 Metadata Challenge

The context of sensors is usually described in point names, or referred to as *metadata*, which are often a concatenation of abbreviations encoding contextual information. For example, as shown in Table 1, a sensor name `SODA1C600A_ART` conveys: `SOD` - building name, `A1` - air handling unit id, `C600A` - room id, and `ART` - type of measurement. The naming convention and rules used in generating these point names vary from vendor to vendor, thus requiring effort to interpret on a per-building basis. However, point names do not necessarily contain all the information required. Not only because the metadata can be incomplete, such as `EBU3B.Unknown..ZN-T` in Table 1, but also these point names need update, as buildings are upgraded or applications evolve over time.

A lack of capability to automatically parse and interpret sensor context has been long standing in the way of revolutionizing buildings at scale. Commercial smart building solutions have started to prevail (Panoptix, APOGEE, Talisen Technologies, etc), but they still rely on proprietary tools like Niagara to interpret and map the metadata, involving significant manual effort. Moreover, anecdotally, labeling one single point usually takes a few minutes and sub-hundred dollars. Our proposed solution would ideally reduce this manual effort to the minimum.

2.2 Related Work

Despite the existence of standard schemas [1, 3, 6], extracting key contextual information about sensors and actuators in a building,

[2]The dataset and code for reproducing results in this paper are readily available on github.

especially the legacy ones, and mapping it to a schema still remain a laborious manual process. Significant advances have been achieved recently in sensor metadata extraction. The majority of these works build upon active learning to reduce manual effort. Bhattacharya et al. [7] propose to iteratively learn a set of regular expressions to parse sensor names and convert them into a common name space. Hong et al. [22] develop a clustering-based active learning method to select the most informative sensors for classifier training and as well propagate the labels to similar instances. Balaji et al. [4] formulate a similar approach by employing hierarchical clustering to group sensor names and labeling one instance from each group. Koh et al. [25] explore a multi-stage active learning mechanism involving conditional random fields and multilayer perceptron to learn the representation of metadata structure for labeling sensors. These active learning based methods have shown promising results in reducing the required manual labeling effort, yet they rely on an omnipotent expert to provide correct labels for queried examples. In this work, we demonstrate that, aside from the infallible experts, other information sources with unknown reliability, also called *weak oracles*, such as classifiers trained on existing buildings, can also be utilized to help construct an accurate model to extract metadata for a new building. While we particuarly consider type classification in this work, our proposed methodology is complementary to the aforementioned work that extracts information in addition to types [7, 25].

In a related line of research, Hong et al. [21] introduce a transfer learning based technique that adapts knowledge from existing labeled buildings to a new one for classifying sensor types. However, they only estimate the quality of transferred labels once at the beginning of their procedure and then combine these labels to predict the type for a subset of sensors in the new building. By contrast, in our approach, we continuously update the estimate of weak oracles' reliability and also use the labels from weak oracles to facilitate classifier training. The weak oracles benefit our learning process in two aspects by: 1) helping to better estimate the informativeness of instances for selection; 2) providing additional labels for classifier training to complement the trustful but costly strong oracle. As for the instance selection strategy, most active learning methods [4, 8, 22] use the prediction uncertainty about instances from the learnt classifier; we also leverage the weak oracles to measure the informativeness of instances.

The key in utilizing noisy labels from weak oracles depends on the aggregation rule, i.e., how to combine the noisy labels from each weak oracle; and there have been extensive studies in crowdsourcing aiming to effectively infer high-quality labels from noisy responses [41, 43, 46]. For example, weighted majority voting and Bayesian voting [9, 30, 32] are proposed to model the ability of different labelers. In recent years, increasing attention has been paid to combining active learning and these label aggregation methods [14, 44]. But most of these studies do not assume the existence of a strong oracle, and the active selection is only about which weak oracle to use for labeling an instance. Even if they have access to a strong oracle, the obtained labels from the strong oracle are not used to refine the aggregated labels. To further promote the interaction between active learning and label aggregation, we develop a framework which allows online updates of the parameters for weak oracle evaluation.

As noted in the truth discovery work [28], real-world information sources have different domains of expertise and biases, their performances differ on different groups of tasks. Zhang et al. [45] proposes a simple yet effective method to analyze the reliability of labelers by clustering the instances based on their responses. To identify each weak oracle's area of expertise, we also explore clustering-based label aggregation and combine it with the learning process in selective sampling.

3 METHODOLOGY

In practice, aside from the labels returned by reliable domain experts for sensor metadata mapping, there are also abundant information sources to harvest – for example, the labels provided by a group of non-experts, or the labels transferred from other buildings with similar configurations or location. Although these labels do not perfectly apply to a new building, hence considered "noisy", the easy access to such information sources provides a more cost-effective way to extract building metadata.

In this work, we focus on an important category of metadata mapping – sensor type classification. We propose a selective sampling framework (SS) to aggregate labels from a set of weak oracles by iteratively interacting with a strong oracle. The aggregated labels complement the expensive labels from the strong oracle for type classifier training: the framework continuously selects the most "informative" instance to query the strong oracle, and the acquired strong label simultaneously improves weak oracle reliability estimation and type classifier training. In this section, we provide details of the proposed solution framework.

3.1 Overview of the Framework

We first formally define the notations to be used in describing the proposed framework. We have a strong oracle O_0 which can always provide correct sensor types as labels, and M weak oracles $\{O_1, O_2, ..., O_M\}$ with unknown reliability. For weak oracle O_k, its label accuracy is denoted as $w^{(k)} \in [0, 1]$. The strong oracle could be a domain expert, while the weak oracles could be non-experts or even statistical classifiers transferred from other buildings. Let us consider a set of N instances of sensor points $D = \{x_1, x_2, ..., x_N\}$ with J different classes of sensor types. Each sensor point x_i is characterized by its point name string and time series readings, upon which we draw features to represent the point. We will refer to the features created from point name strings as *name features* and those from time series readings as *data features*. The sensor type of x_i is denoted by its true label y_i, while the noisy labels for x_i obtained from weak oracle O_k, also referred to as *responses*, are marked as $r_i^{(k)}$, and we denote $r_i = \{r_i^{(k)} | k = 1, ..., M\}$.

In this work, we assume the labels from weak oracles are free, and the labels from the strong oracle have the same unit cost. We leave the setting where different oracles have different labeling cost as our future work. The goal of the framework is to estimate a classifier $f : f(x) \rightarrow y$ with respect to a training set D_{trn}, so as to maximize the classifier's type classification accuracy while minimizing the cost in creating D_{trn}. We denote the set of instances with labels from the strong oracle as D_s, and the set of instances with no strong labels as $D_u = D - D_s$. Since points with similar name features tend to share the same label, we adopt the label

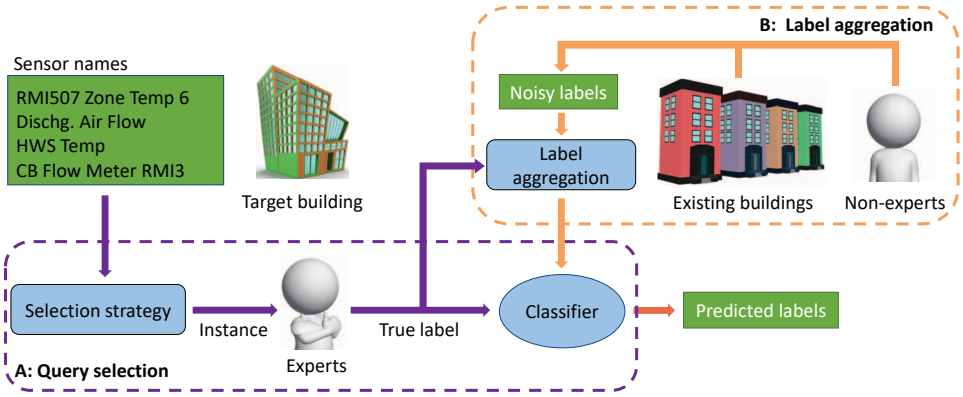

Figure 1: Overview of our selective sampling framework for sensor metadata inference: A) The Query Selection component samples the most informative instances and acquires their true labels from an infallible expert; B) The Label Aggregation component takes advantage of the noisy labels collected from multiple weak oracles to infer the true sensor labels.

propagation technique described in [22] to propagate the labels of instances in D_s to their neighbors in the feature space. As a result, we denote the set of instances in D_u with propagated labels as D_p. The study in [22] shows that the propagated labels have very high accuracy and can thus be regarded as reliable. For each instance, the noisy labels from weak oracles are aggregated to infer the most likely label. The probability of the aggregated label being correct is referred to as the *confidence*. In $D - (D_s \cup D_p)$, the set of instances with aggregated labels from weak oracles above a given confidence threshold is denoted as D_w. The relationship of these notations is shown in Figure 2.

Figure 1 illustrates the workflow of our selective sampling framework, which contains two major components: *query selection* and *label aggregation*. As we assume the weak oracles are free to query, in the query selection component, we only select the most "informative" instances to be labeled by the strong oracle, and query all instances to the weak oracles. The label aggregation component collects noisy labels from weak oracles and infers the true sensor types based on its reliability estimation of these weak oracles. Both the true labels from strong oracle and the aggregated labels from weak oracles are used for training the sensor type classifier.

3.2 Clustering-based Label Aggregation

With access to the noisy labels from weak oracles, an effective label aggregation method is important for integrating these responses and inferring the true sensor types. To this end, we adopt an iterative weighted majority voting (IWMV) method [31] to evaluate the reliability of the weak oracles and infer the true labels from the noisy labels. IWMV is based on a weighted voting scheme, where the aggregated label is calculated by

$$\hat{y}_i = \underset{j \in \{1,\ldots,J\}}{\operatorname{argmax}} \sum_{k=1}^{M} v^{(k)} \mathbf{I}(r_i^{(k)} = j), \tag{1}$$

where $\mathbf{I}(\cdot)$ is an indicator function. IWMV defines a voting weight $v^{(k)}$ for each weak oracle O_k to measure its contribution to label aggregation, and $v^{(k)}$ is computed based on the estimated accuracy

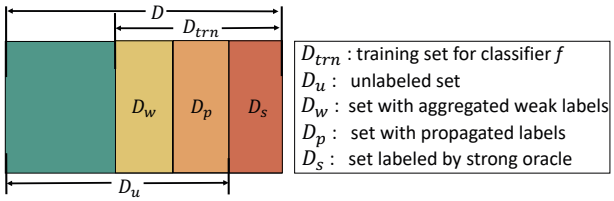

Figure 2: Illustration of different sets of instances.

$w^{(k)}$ of weak oracle O_k. Intuitively, a more accurate weak oracle should have a higher weight in determining the final aggregated label of an instance. The model employs an iterative procedure to estimate the accuracy of weak oracles and infer the true labels. In each step, IWMV predicts the label by Eq. (1), then the maximum likelihood estimation of weak oracle's accuracy is calculated as:

$$\hat{w}^{(k)} = \frac{\sum_{i=1}^{N} \mathbf{I}(r_i^{(k)} = \hat{y}_i)}{\sum_{i=1}^{N} T_{ik}}, \tag{2}$$

where T is the observational matrix, and $T_{ik} = 1$ when weak oracle O_k gives a label for instance x_i, otherwise $T_{ik} = 0$. Based on the weak oracle's accuracy estimated by Eq. (2), the voting weight is computed as:

$$v^{(k)} = J\hat{w}^{(k)} - 1. \tag{3}$$

The process runs iteratively until it converges or reaches the maximum number of iterations.

Based on the above IWMV method, for instance x_i and its associated weak responses $\{r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(M)}\}$, the probability of its true label y_i being j given by the weighted voting can be estimated as:

$$p(y_i = j | x_i) = \frac{\sum_{k' \in Q} v^{(k')}}{\sum_{k=1}^{M} v^{(k)}}, Q = \{k' | r_i^{(k')} = j\}. \tag{4}$$

For the aggregated label $\hat{y}_i$, we define its *confidence* to be $p(y_i = \hat{y}_i | x_i)$ calculated by Eq. (4). It follows that the confidence of the aggregated labels is in the range of $[1/J, 1]$. To improve the accuracy of the aggregated labels, we filter out those whose confidence is

below a *confidence threshold* C; and basically this threshold decides the accuracy-coverage trade-off in the label aggregation component.

As more ground-truth labels are obtained from the strong oracle, D_s and D_p can be utilized for more accurate estimation of weak oracle reliability. In particular, for any $(x_i, y_i = j)$ in $D_s \cup D_p$, we fix $(x_i, \hat{y}_i = j)$ in label aggregation, and set the confidence $p(y_i = j|x_i) = 1$. With such verified information, the estimate of weak oracle reliability in Eq. (2) can be improved with the newly acquired labels, which in turn leads to more accurate aggregated labels in D_w, and therefore improves classifier training.

On the other hand, a common observation is that the reliability of a weak oracle on a group of similar instances tends to be consistent, but it might vary significantly over different groups. For example, if building A is located in a similar region to a target building B but with significantly different configurations of cooling fans, then the weak oracle from building A may achieve 90% accuracy when labeling the light sensors in building B, while might only be 10% accurate in recognizing the fan speed sensors. Therefore, instead of estimating the voting weight for each weak oracle as a whole in Eq. (2), evaluating them on different clusters of sensors could bring deeper insights and improved quality into label aggregation.

Based on the above considerations, we develop a new clustering-based label aggregation method. The clustering Ω on D is initialized by a Dirichlet Process $DP(G_0, \alpha)$ described in [15], where G_0 is a base distribution, and α is a scaling parameter. After generating the clusters, we perform label aggregation on each cluster separately to estimate each weak oracle's reliability. Specifically, we apply Eq. (1), (2) and (3) to estimate every weak oracle O_k's accuracy $w_c^{(k)}$ and voting weight $v_c^{(k)}$ in each cluster c.

However, as our initial clustering is performed only based on the name features of sensors, rather than the ground-truth sensor types, instances that are grouped together by clustering might actually belong to different types. For example, two sensors with similar names `RM108C Zone Temp 3` and `RM108C Zone Temp 3.STP` are likely to be assigned into one cluster, but their types are "temperature" and "temperature set point", respectively. This would hurt the quality of label aggregation: a weak oracle may have high accuracy in recognizing the temperature sensors, but not the set points; and therefore its high weight over temperature sensors in the cluster will mislead the label aggregation for the setpoints in this cluster.

The solution to this problem is to refine the clusters based on the learnt classifier on the fly. If we observe that the learnt classifier assigns multiple distinct labels in a cluster, it is a strong indicator of finer boundaries inside the cluster; thus further sub-clustering is needed. To this end, we measure the *impurity* of a cluster by its class entropy calculated as:

$$H(c) = -\sum_{y \in Y_c} p(y) \log(p(y)). \tag{5}$$

Based on the prediction made by $f(x)$, Y_c is the set of unique labels in cluster c, and $p(y)$ is the proportion of label y in c. The average class entropy of different clusters is computed as:

$$\overline{H} = \frac{\sum_{c \in \Omega} H(c)}{|\Omega|}, \tag{6}$$

where Ω is the set of all current clusters. We thus use this average entropy measure to decide when to further divide the clusters during the selective sampling process. Specifically, we use a threshold r on the change of average class entropy: every time the average class entropy increases by r times than the last time of clustering update, sub-clustering will be performed in each cluster. In particular, k-means (k=$|Y_c|$) will be used to generate sub-clusters in each cluster $c \in \Omega$.

3.3 Disagreement-based Query Selection

In our selective sampling process, the strategy for selecting the most informative example to query the strong oracle is the key to quickly improving the accuracy of the classifier. Built upon the strategy that considers both the informativeness and representativeness of an instance [22], we further consider the potential influence of a selected instance $\hat{x}$ on refining the aggregated labels in D_w. For example, if three weak oracles label a timer as `Humidity`, `Humidity`, `Timer`, respectively, without additional information, the label aggregation component may incorrectly follow the majority and label it as a humidity sensor, and thus assigns high confidence to the first two oracles in the next round. Yet, if we obtain the ground-truth label from the strong oracle and use it to correct the aggregated label from weak oracles, the weak oracles can be re-evaluated and in turn deliver more accurate aggregated labels subsequently.

To measure the informativeness of an instance and the influence it brings to label aggregation, we propose a *disagreement-based* selection strategy. Apart from the classifier $f(x)$ trained on $D_s \cup D_p \cup D_w$, another classifier $f_g(x)$ is trained on $D_s \cup D_p$. In this way, we obtain a relatively "weaker" $f(x)$ trained including information from D_w and a "stronger" $f_g(x)$ trained purely on reliable labels. If they disagree on an instance, the instance should potentially be informative for improving the aggregated labels from weak oracles (e.g., at least one of them is incorrect on this instance). Specifically, in every iteration, by comparing the predictions by the two classifiers, we first identify the candidate set D_c of instances from D_u such that $f(x)$ and $f_g(x)$ give different labels to them, and then compute a disagreement score $d(x_i)$ for each instance $x_i \in D_c$ using the Kullback-Leibler divergence [27] based on the predictions by the two classifiers:

$$d(x_i) = \sum_{j=1}^{J} p_g^i(j) \log\left(\frac{p_g^i(j)}{p_c^i(j)}\right) + \sum_{j=1}^{J} p_c^i(j) \log\left(\frac{p_c^i(j)}{p_g^i(j)}\right), \tag{7}$$

where p_c^i and p_g^i denote the predicted label distribution on x_i by the two classifiers $f(x)$ and $f_g(x)$, respectively.

Another important factor in query selection is the representativeness of the selected instance. Considering that the names of sensors of the same type in a building often share similar sub-strings; as a result, they tend to be "neighbors" in the feature space. We thus find the neighbors of an instance x_i whose Euclidean distance to x_i in the feature space is smaller than a threshold, denoted by $N(x_i)$; and then measure the representativeness of x_i by the number of x_i's neighbors which have different labels predicted by $f(x)$ and $f_g(x)$, namely, $|D_c \cap N(x_i)|$. Consequently, we calculate the informativeness by combining the prediction disagreement score and

Algorithm 1: Selective sampling with clustering-based label aggregation.

Input: sensor points $D = \{x_1, x_2, ..., x_N\}$, budget B, confidence threshold C;
Output: predicted sensor types $Y = \{y_1, ..., y_N\}$;
Initialization: 1) Generate initial clustering Ω with $DP(G_0, \alpha)$;
2) Query all instances in D to M weak oracles and obtain the labels $R = \{r_i^{(k)} | i = 1, ..., N; k = 1, ..., M\}$;
3) Set $D_s = \{\}$, $D_p = \{\}$, $D_u = D$, $D_w = Agg(R, D_s \cup D_p, C, \Omega)$, $iter = 0$;
4) Train the classifier $f(x)$ on D_w, and compute $\overline{H} = \overline{H}_{old}$ following Eq. (6);
while $iter < B$ **do**
  Select an instance $\hat{x} = select(D_u)$ as described in Section 3.3;
  Query O_0 for the true label y for $\hat{x}$;
  $D_s = D_s \cup \{\hat{x}, y\}$, $D_u = D - D_s$;
  Propagate y to the neighbors of $\hat{x}$ and update D_p as described in [22] ;
  Update $D_w = Agg(R, D_s \cup D_p, C, \Omega)$ as described in Section 3.2 ;
  Calculate the average class entropy $\overline{H}$ based on Eq. (6);
  if $\overline{H} > r * \overline{H}_{old}$ **then**
    Update clustering Ω by current predicted labels as described in Section 3.2;
    $\overline{H}_{old} = \overline{H}$;
  end
  Train the classifier $f(x)$ on $D_{trn} = D_s \cup D_p \cup D_w$;
  $iter = iter + 1$;
end

the representativeness of each instance as:

$$score(x_i) = d(x_i) * |D_c \cap N(x_i)| \,. \tag{8}$$

Intuitively, following Eq. (8), instances with high disagreement scores would imply possible errors in the prediction from either $f(x)$ or $f_g(x)$. Therefore, querying such instances provides an opportunity for the classifier or the label aggregation component to correct their mistakes. For the classifier, this strategy discloses its current prediction uncertainty and helps it to improve its estimation. For the label aggregation component, this strategy can guide it to realize the reliability of weak oracles and lead to improved weak labels aggregation.

We shall note that most existing instance selection strategies, such as selection by uncertainty [8], query by committee [17], only consider the immediate effect of selected instance. In other words, they select the next instance to query by estimating the improvement of the classifier trained on $D_s \cup \{\hat{x}\}$, ignoring the change that $\hat{x}$ may bring to D_p or D_w. Noticing this, a prior study [22] proposes an entropy-based selection strategy which first clusters the unlabeled set and then locates the cluster $\hat{c}$ with the highest product of class entropy and cluster size, and finally selects the most "representative" instance from the cluster by estimating the conditional likelihood $p(x|\hat{c})$. By jointly considering the informativeness and representativeness, this entropy-based method has the potential to improve the label quality in D_p. Our disagreement-based selection solution further extends the consideration to D_w; in other words, the potential improvement in reliability estimation of weak oracles. This will best amplify the utility of every obtained ground-truth label; and it is also empirically confirmed in our later evaluations.

Putting it all together: Algorithm 1 summarizes the entire procedure of our proposed selective sampling framework for sensor type classification. The input to our framework includes the point name strings for instance feature construction, a fixed query budget B, and the confidence threshold C. As the weak oracles are free, we query all instances in D to all the weak oracles and obtain their responses R. Agg is the clustering-based label aggregation method described in Section 3.2, which updates the estimate of weak oracle's reliability, and returns instances with aggregated labels whose confidence scores exceed a threshold (but the instances in D_s or D_p are not included). In each iteration, the most informative instance $\hat{x}$ is selected from D_u, following the disagreement-based selection strategy discussed above. Then the ground-truth label y for instance $\hat{x}$ obtained from the strong oracle is propagated to the nearby instances in the feature space, following the label propagation procedure in [22]. With the updated D_s and D_p, the clustering-based label aggregation component is then applied to re-evaluate the weak oracles and update the aggregated labels in D_w. The classifier is re-trained on $D_{trn} = D_s \cup D_p \cup D_w$. The average class entropy is calculated across all the clusters, and once it has increased by r times than the last time of clustering update, we perform sub-clustering to generate finer clusters following the steps in Section 3.2.

Table 2: Key statistics of the evaluation data set, including the number of sensors, name feature dimension, and number of sensor types.

Building	#Sensor	#Name Feature	#Sensor Type
A	661	1501	13
B	1753	284	17
C	1160	135	15
D	4007	492	15
E	3816	511	15

4 EVALUATION

In this section, to demonstrate the effectiveness of our proposed solution, we conduct extensive evaluations based on the data from a large collection of real-world buildings. First, we introduce the dataset and our experiment setup. Then we compare our approach against a suite of related baselines by measuring the accuracy of type classification with a varying annotation budget. In particular, we investigate the effect of clustering-based label aggregation and different query selection strategies on classifier training in a target building, and the robustness of our solution under different configurations of weak oracles.

Table 3: Accuracy of weak oracles (1 through 6) on different target buildings (A through E). Each oracle is a statistical classifier trained on a source building.

	1	2	3	4	5	6
A	0.619 ± 0.392	0.435 ± 0.312	0.285 ± 0.412	0.577 ± 0.392	0.421 ± 0.447	0.299 ± 0.433
B	0.776 ± 0.324	0.356 ± 0.419	0.524 ± 0.381	0.861 ± 0.179	0.582 ± 0.468	0.613 ± 0.197
C	0.948 ± 0.198	0.547 ± 0.473	0.447 ± 0.488	0.917 ± 0.209	0.538 ± 0.483	0.766 ± 0.318
D	0.458 ± 0.246	0.707 ± 0.385	0.440 ± 0.335	0.602 ± 0.479	0.692 ± 0.378	0.506 ± 0.321
E	0.518 ± 0.329	0.583 ± 0.387	0.562 ± 0.390	0.239 ± 0.241	0.661 ± 0.316	0.502 ± 0.328

Table 4: Details of sensor types and corresponding number in each building.

Sensor Type	A	B	C	D	E
co2	16	52	0	7	24
air pressure	142	216	215	0	72
room temp*	159	231	207	238	252
operation status	59	58	41	90	135
setpoint	140	486	229	945	1360
airflow	14	172	9	233	223
hot water supply temp	27	1	1	1	1
hot water return temp	15	1	1	1	1
chilled water supply temp	18	6	10	2	3
chilled water return temp	15	4	9	2	3
supply air temp	20	17	3	3	3
return air temp	6	2	4	3	3
mixed air temp	5	2	3	0	0
occupancy	25	52	0	10	0
vavle position	0	290	10	234	0
power measurement	0	0	0	0	60
control command	0	138	403	2224	1662
fan speed	0	25	15	14	14

*temp stands for temperature

4.1 Experiment Setup

Dataset. We evaluate our framework with sensor data from a collection of real-world buildings, consisting of the point names and one week's time series readings of more than 11,000 sensors of 18 different types from five office buildings. These buildings are located across the US, commissioned by four different vendors with different levels of automation, and were built in different times; they reasonably represent the US office buildings. The length of sensor names varies from 12 to 30, and the time series data is reported every 5 ∼ 15 minutes, depending on the building. Table 2 summarizes the key statistics of the dataset, and Table 4 shows the distribution of sensor types, more details about the dataset can be found in [26].

To extract features from this data, we adopt k-mers [29] as the *name feature* representation of point names, which are all the possible length-k substrings of a point name. In our experiments, the length of k-mers is set to 3 and 4. We count the frequency of k-mers in each point name as the feature value. In all five buildings, *data features* of the time series readings of each sensor point are 44−dimensional statistical summary of 45-minute long sliding windows over the primitive time series, including minimum, maximum, median, variance, skew, etc[3]. The data features are only used for generating weak oracles, as they better generalize across buildings for type classification, according to the findings in [21].

Strong and Weak Oracles Setup. Our framework can take input from non-expert annotators or already trained type classifiers as weak oracles. As a proof-of-concept, in our experiments, we construct classifiers that are trained using *data features* and labels from existing labeled buildings and transfer them to a new building as the weak oracles. For each *target building*, we create classifiers as weak oracles using the data features of sensors in all the other buildings (*source buildings*). Multiple types of classifiers are used to simulate the situation where different weak oracles have different reliability. In particular, we estimate random forest (RF), support vector machines (SVM), and logistic regression (LR) on each source building. In this way we obtain 3×4 weak oracles for each target building, and we randomly select 6 of them to increase the difficulty of selective sampling. Table 3 presents the overall ground-truth accuracy and the standard deviation of the weak oracles on the target buildings for sensor type prediction, which is not disclosed beforehand to any algorithm to be evaluated. For strong oracle, as we assume it is infallible, we directly return the ground-truth label of the selected instance each time the strong oracle is queried.

For these five buildings, accuracy of the weak oracles varies from 28.5% to 94.8%; and even for the same classifier, its classification accuracy varies across target buildings and different types of sensors. This shows the need of accurate estimates of weak oracles' reliability, especially per type of sensors. We adopt logistic regression as the classifier to be estimated for the target building.

4.2 Type Classification Results

To investigate the effectiveness of our framework in utilizing the noisy labels for type classification, we compare our approach with multiple baselines which use distinct ways to leverage the weak supervision, i.e., the information from other buildings. To thoroughly evaluate the performance, we compare the algorithms under a varying query budget, i.e., how many ground-truth labels can be obtained. The macro accuracy across all sensor types is used as the evaluation metric.

First, we briefly introduce all the baseline methods and their settings in the following:

[3]we refer interested readers to [21] for further detail.

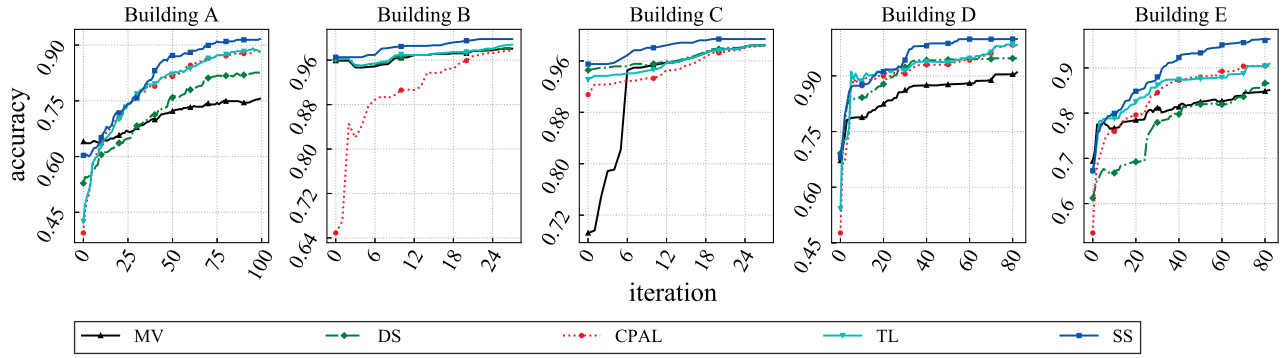

Figure 3: Type classification accuracy of our solution (SS) against various baselines on all five buildings.

- **Clustering-based active learning with label propagation (CPAL).** This is a state-of-the-art active learning method for sensor type classification [22]. Rather than leveraging noisy labels from weak oracles, it propagates obtained ground-truth labels to unlabeled instances as weak supervision.
- **Transfer learning (TL).** This is a state-of-the-art transfer learning solution for type classification [21]. It only leverages the weak oracles for classifier training in the target building. To make it comparable to other solutions, we follow [26] to integrate it with CPAL: the aggregated labels from TL will be used to initialize the classifier for active learning in the target building.
- **Majority voting (MV).** This is the most classical solution to label aggregation: among the noisy labels gathered from weak oracles for each instance, the most frequent label is selected. We break ties arbitrarily. The same as in the TL baseline, we use the aggregated noisy labels from MV to initialize the classifier in CPAL for active learning.
- **Dawid-Skene model (DS).** This is a popularly used method for crowd-sourcing: instead of only estimating the accuracy of weak oracles, DS [9] estimates a confusion matrix for each weak oracle, and the true labels are inferred by an Expectation Maximization algorithm over the confusion matrices of weak oracles. Again, we use its aggregated noisy labels to initialize the classifier in CPAL for active learning.

Figure 3 reports on the comparison between our method (SS) and all baselines across all five buildings. CPAL is the only method that does not leverage weak oracles. In the early stage of model update, there is only a small number of labels from the strong oracle, and thus most of the methods that are augmented with labels from the weak oracles have higher accuracy than CPAL. As more ground-truth labels are acquired, CPAL quickly catches up with MV and DS. The quality of aggregated labels in MV is limited by its untenable assumption that all weak oracles can be equally treated. The DS model, albeit effective in many truth discovery applications, does not adapt well to the building sensor data — the key limitation is that in building domain there are often many types of sensors, but not enough labeled instances in each category. This fact leads the confusion matrix in DS to be too sparse for an accurate estimation of weak oracle reliability.

In most cases, although the weak oracles can benefit classifier training in early stage, errors in the aggregated labels will still degrade the performance significantly in the long term; and a high-quality label aggregation component is thus needed. TL utilizes a weighted ensemble method to aggregate transferred labels. For the coverage-accuracy trade-off, here we set its consistency threshold $\delta = 0.6$ for its best performance [21]. In this setting, TL can collect transferred labels for about 14% instances with accuracy higher than 95% on Building A, while SS can produce labels for 33% instances with about 92% accuracy before querying the strong oracle by setting its confidence threshold C to 0.9. Although slightly lower in accuracy, the improvement of SS in label quantity contributes more to classifier training in the early stage. Furthermore, the accuracy of the aggregated labels in SS can be improved with more rounds of active querying. From Figure 3, SS almost always outperforms all the baselines, achieving higher accuracy at a lower cost in querying. We attribute its improvement to two key factors: the high-quality label aggregation scheme, and the disagreement-based selection strategy which takes the weak oracles into consideration.

Upon further inspection, SS is able to correctly identify most of the examples in the major classes (e.g., room temp or setpoint in Table 1), which suffice many control or monitoring applications. On the other hand, misclassifications mainly occur in minor classes. For instance, mixed air temperature sensors are rare in all the buildings (only 5 examples in building A), and the weak oracles transferred from other buildings also tend to make incorrect predictions for these sensors with high confidence. In this case, the classifier is likely to agree with the aggregated labels, though incorrect, which reduces the possibility to query these sensors to the strong oracle for correction.

To further investigate the label aggregation quality of SS without any active querying, in Figure 4, we present the accuracy and coverage of the aggregated labels before starting the process in Algorithm 1, where the *coverage* of aggregated labels is calculated as $|D_w|/|D|$. For brevity, we only show the results on Building B, but the observations are similar on the other buildings. We notice that when we increase the confidence threshold, we get aggregated labels of better quality, i.e., higher accuracy, but the number of labeled instances drops, i.e., lower coverage. With a proper C to control the trade-off, SS can produce a considerable amount of aggregated labels with encouraging high quality.

We also evaluate the quality of weak oracle accuracy estimation to further verify that the label aggregation in SS benefits from the

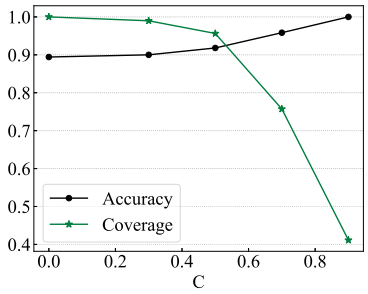

Figure 4: The initial coverage and accuracy of aggregated labels with different confidence threshold C on Building B.

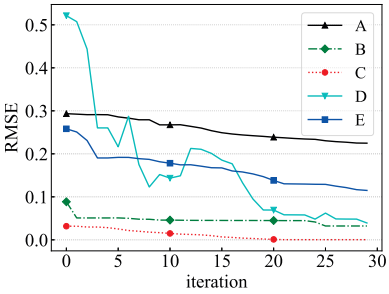

Figure 5: Root-mean-square error of weak oracle accuracy estimation during the selective sampling process on different buildings.

interactive query process. Specifically, we compare the estimated weak oracle accuracy against the ground-truth under root-mean-square error (RMSE) in Figure 5. In general, the estimated accuracy for weak oracles are improved as more true labels are obtained over iterations. This directly contributes to the improved quality of classifier training in the target building, as shown in Figure 3.

4.3 Clustering-based Label Aggregation

To evaluate the effect of the clustering-based label aggregation in our solution, we compare the performance of the learnt classifier under different clustering settings:

- **No clustering.** In this case, no clustering is performed, and the label aggregation component estimates the weak oracles' accuracy globally, i.e., the weight of a weak oracle is the same across all instances in the dataset. But the estimation of weak oracles' accuracy will be updated as more ground-truth labels are acquired after each iteration.
- **Initial clustering.** The clusters are created and fixed before the selective sampling starts. The weak oracle evaluation is performed with respect to this initial cluster setting.
- **Sub-clustering.** After initializing the clusters, we enable them to be further refined as described in Section 3.2. Empirically, we set the sub-clustering threshold $r = 2$.

The performance of clustering settings above is reported in Figure 6. Under the same labeling cost, clustering does help the label aggregation component to better integrate noisy labels. Compared with the no-clustering setting, clustering delves deeper to investigate the response patterns in groups of sensors which share similar metadata.

To better illustrate the fact that the reliability of weak oracles varies drastically across different clusters, Table 5 shows the ground-truth accuracy of the weak oracles in different clusters on the three buildings. For simplicity, we only present the results when 3 clusters are generated for all buildings. As shown in Table 5, the accuracy of weak oracles in buildings varies significantly among different clusters, especially in Building A. For example, on cluster 1 of building A, oracle 1, 4, 5 have the highest reliability, while on cluster 2, their accuracy all drops to below 50%, but oracle 2 and oracle 6 turn out to be more trustful. Similarly, in building E, weak oracle 1 shows 82% accuracy on cluster 1, but only 16% on cluster 2.

Facing this challenge, only the clustering-based method can summarize and derive a relatively *consistent* evaluation on the weak oracles' quality, thus achieving better performance. From Figure 6, the clustering-based method improves the learnt classifier accuracy from 52% to over 60% before any manual label is obtained on building A; and during the entire selective sampling process, the clustering-based method always outperforms the no-clustering one. On top of that, the sub-clustering method generates finer clusters with higher purity, which further enhances classifier's accuracy. Sub-clustering especially benefits label aggregation when initial clustering is misled by the original name features. For example, in building E, many point names share the same prefix, such as `AP_M.RM-1839.ZN-T` and `AP_M.RM-1839.AHTG-STPT`, which are likely to be initially assigned into one cluster, but the performance of weak oracles on these two groups are actually different. As sub-clustering can further refine the clusters with more queries, the two groups can be separated, and so the label aggregation accuracy is improved.

The influence of sub-clustering is less significant on building B and C compared to the case with initial clustering. This is because the data features in building B and C are highly associated with their sensor types, the initial clustering already has high purity and the improvement of further clustering refining tends to be marginal.

4.4 Query Selection Strategy

To verify the effectiveness of our disagreement-based query selection strategy in reducing the labeling cost, we compare it with other selection strategies under the same budget, i.e., number of strong labels to obtain. We consider the following selection strategies:

- **Random selection.** In each iteration, randomly select an instance from the unlabeled set to query.
- **Cluster entropy-based selection.** Hong et al. [22] propose the cluster entropy based selection strategy, which jointly considers the cluster entropy and the size of the cluster where the selected instance belongs to.
- **Confidence-based selection.** This heuristic strategy always selects the instance with the least confidence in the label aggregation component.
- **Disagreement-based selection.** Our disagreement-based selection strategy described in Section 3.3.

Table 5: Ground-truth cluster-level accuracy of weak oracles (1 through 6) in all buildings (A through E) when the number of clusters is fixed to 3. In each cluster of a building, the highest accuracy among all weak oracles is marked in bold.

	A	B	C	D	E
1	0.78, 0.43, **0.60**	0.91, 0.63, **1.00**	**0.71**, **1.00**, **0.98**	0.44, 0.35, 0.47	**0.82**, 0.16, 0.38
2	0.29, 0.91, 0.41	0.07, 0.52, 0.00	0.54, 1.00, 0.06	0.98, 0.99, **0.60**	0.57, 0.47, **0.99**
3	0.02, 0.66, 0.29	0.66, 0.46, 0.00	0.13, 1.00, 0.00	0.08, 0.33, 0.51	0.53, 0.47, 0.99
4	**0.79**, 0.43, 0.55	0.91, **0.79**, 0.93	0.63, 1.00, 0.93	**0.99**, **1.0**, 0.46	0.23, 0.25, 0.22
5	0.78, 0.00, 0.39	**1.00**, 0.24, 0.44	0.54, 1.00, 0.04	0.97, 0.99, 0.58	0.53, **0.72**, 0.99
6	0.02, **0.92**, 0.27	0.61, 0.65, 0.24	0.12, 0.99, 0.76	0.08, 0.33, 0.60	0.49, 0.41, 0.86

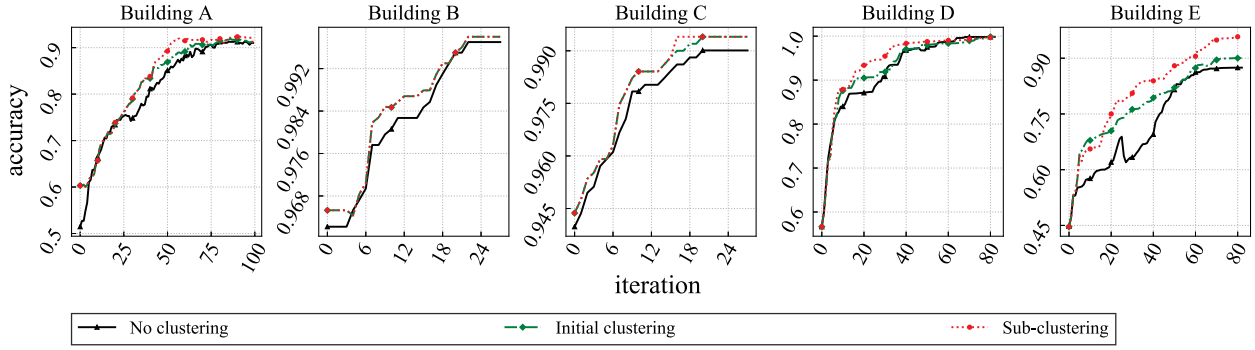

Figure 6: Comparison of classifier accuracy with different clustering strategies in our selective sampling solution.

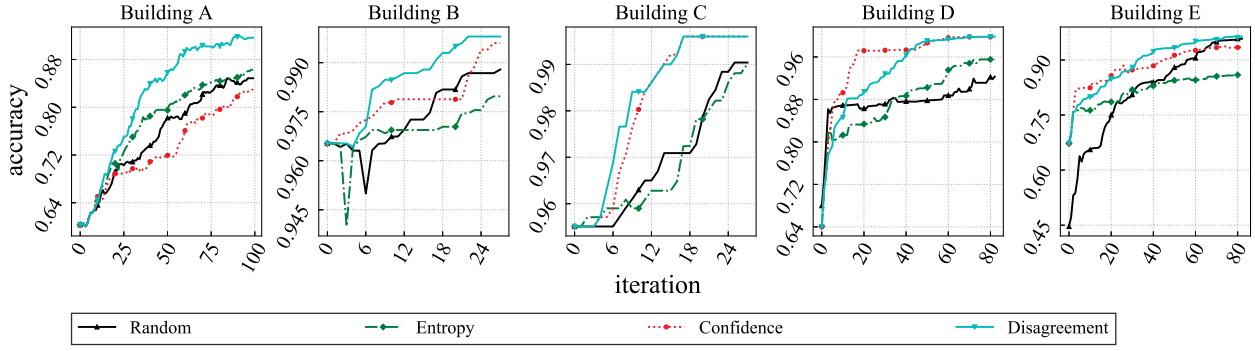

Figure 7: Type classification accuracy of our selective sampling solution under different query selection strategies.

Figure 7 illustrates the classifier accuracy under different query selection strategies. All the selection strategies can guide the classifier to reach accuracy over 80% quickly. Suffering from lack of guidance, random sampling's performance is unstable. Cluster entropy based selection combines the current classifier's prediction and possible effect on the propagated labels. But it does not utilize the confidence of aggregated labels, which indicates the contribution of an instance to label aggregation. As a result, its performance is limited in this respect. Exactly on the opposite, confidence-based selection strategy only considers the uncertainty in weak label aggregation, but ignores the classifier's prediction and the representativeness of instances. It leads the selection to fully depend on the noisy labels.

We also observe that the performance of query selection strategies is often highly related to the properties of the dataset. Some buildings (e.g., building B, C, and D) have multiple types dominant in size and the name features of intra-cluster instances are very similar, while the inter-cluster distances are relatively farther. In these buildings, the name feature-based clusters can well approximate the underlying type distribution. Consequently, the drawback of confidence-based selection strategy can be easily overcome after several queries cover each cluster. The cluster entropy based selection strategy loses its advantage, because the clusters are already of high purity. In other cases (e.g., building A) where the instance name features are harder to distinguish, the performance of cluster-entropy selection strategy is more efficient. Our disagreement-based strategy improves its robustness in different buildings by jointly considering the representativeness of instances as well as the informativeness suggested by both the classifier and

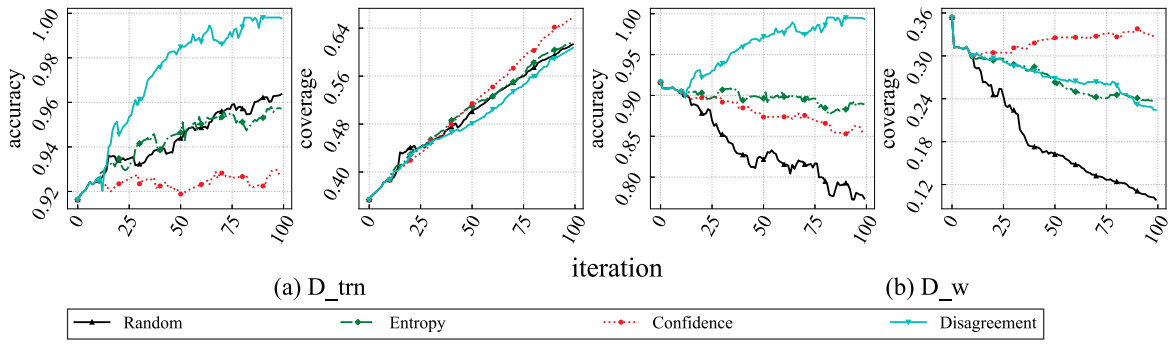

Figure 8: Accuracy and coverage of the labels in D_{trn} and D_w with different selection strategies on building A.

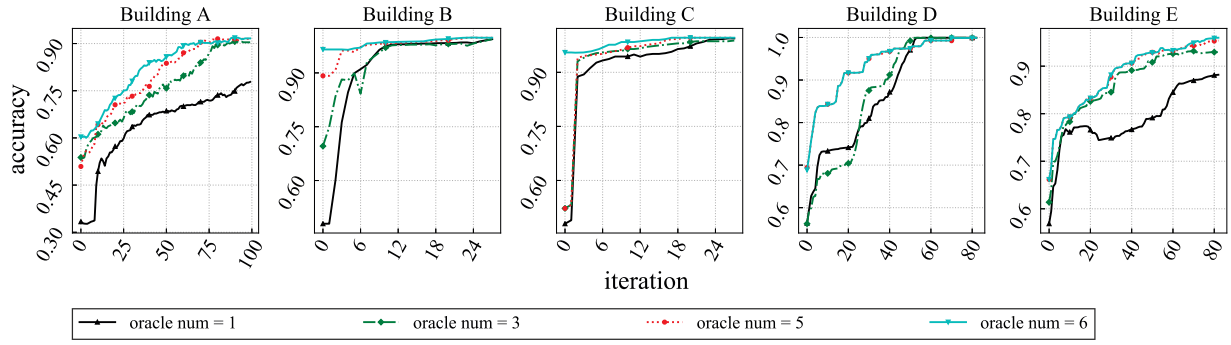

Figure 9: Type classification accuracy of our selective sampling solution under different numbers of weak oracles.

the label aggregation component. In Figure 7, in most of the buildings, we observe substantial improvements for our disagreement-based selection strategy, which outperforms the other baselines by $3\% - 15\%$.

In order to further investigate the impact of the selection strategies on classifier training and label aggregation, we compare the accuracy and coverage on the actively created training set D_{trn} and the aggregated labels D_w during selective sampling. Due to space limit, in Figure 8 we only present the results on building A when $C = 0.9$. Figure 8(a) shows the accuracy and coverage of labels on D_{trn}, and Figure 8(b) shows the results on the "augmented" instance set with aggregated weak labels D_w. With improved accuracy from the learnt classifier, accuracy of the aggregated labels in D_{trn} becomes more and more important: Existing study [34] shows that even a small number of erroneous labels may significantly degrade the performance of the classifier. From Figure 8, compared to other strategies, although the coverage for D_{trn} by our disagreement-based strategy is slightly lower than the others, it can always achieve much higher accuracy on both D_{trn} and D_w. Furthermore, if we zoom into D_w, the disagreement-based and confidence-based selection strategies can actually provide more reliable aggregated labels to complement the classifier training. The main reason is that these two strategies take advantage of uncertainty in weak label aggregation, while the update of D_w with the selection strategy driven by uncertainty of the classifier (e.g. entropy-based selection) is more likely to overlap with the update in classifier prediction and label propagation.

4.5 Number of Weak Oracles

To evaluate the robustness of our selective sampling solution, we vary the number of weak oracles, which makes it harder to infer high-quality labels. Figure 9 presents the learnt classifier's accuracy under different numbers of weak oracles. We remove the weak oracles one by one according to their ground-truth accuracy, i.e., the best one first. In this way, it becomes more and more challenging for the framework to provide useful aggregated labels. In the extreme case where only one weak oracle is preserved, there is no need to vote. The label aggregation component can only use the queried ground-truth labels to evaluate the weak oracle in each cluster, which leads to the worst initial performance of selective sampling in all buildings. Fortunately, with several more oracles joining in, the performance quickly improves and approaches the result when all the weak oracles are available.

5 CONCLUSION

In this paper, we address the problem of building sensor type classification over disparate forms of sensor names. We build a selective sampling framework upon a clustering-based label aggregation method to exploit the abundant free yet noisy labels from multiple information sources. The estimate of weak oracles' reliability is continuously updated as ground-truth labels are actively selected. The query selection strategy simultaneously enhances both the classifier and the component for noisy label aggregation, in order to obtain better type labels. The proposed framework is evaluated on a large collection of real-world building data with over 11,000 sensors from five office buildings with different metadata naming

conventions. The experimental results show that our framework can significantly reduce the amount of manual labeling by synergizing the classifier for predicting sensor types and the mechanism for aggregating noisy labels from other information sources.

Our work particularly focuses on sensor type classification; as more useful information can be identified from sensor names, such as location and equipment id, it would be meaningful to extend our solution to a richer scope of metadata mapping. When considering more context, the importance of a sensor/actuator to certain applications could be included as part of the label from human and serve as another factor of informativeness for deciding which instance to query. As it evolves into a structured prediction problem, label aggregation becomes more challenging, e.g., a weak oracle might be good at recognizing different segments of point names. In addition, although we assume weak oracles are free for labeling, in practice, they might incur different costs based on their level of confidence or according to the difficulty of annotation tasks. And in practice, there might not exist any strong oracle which always provides perfect answers. It is thus important to extend our selective sampling framework to such more general settings, where we not only need to decide which instance to query, but also which oracle to query.

ACKNOWLEDGMENTS

We thank our shepherd and the reviewers for helpful comments. This work was supported by NSF 1940291, 1718216, and Department of Energy DE-EE0008227.

REFERENCES

[1] [n. d.]. Project Haystack. http://project-haystack.org/.
[2] Yuvraj Agarwal, Bharathan Balaji, Seemanta Dutta, Rajesh K Gupta, and Thomas Weng. 2011. Duty-cycling buildings aggressively: The next frontier in HVAC control. In *IPSN*. IEEE, 246–257.
[3] Bharathan Balaji, Arka Bhattacharya, Gabriel Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, Jason Koh, Joern Ploennigs, Yuvraj Agarwal, et al. 2016. Brick: Towards a unified metadata schema for buildings. In *BuildSys*.
[4] Bharathan Balaji, Chetan Verma, Balakrishnan Narayanaswamy, and Yuvraj Agarwal. 2015. Zodiac: Organizing large deployment of sensors to create reusable applications for buildings. In *Proceedings of the 2nd BuildSys*. ACM, 13–22.
[5] Bharathan Balaji, Jian Xu, Anthony Nwokafor, Rajesh Gupta, and Yuvraj Agarwal. 2013. Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings. In *SenSys*. ACM, 17.
[6] Vladimir Bazjanac and DB Crawley. 1999. Industry foundation classes and interoperable commercial software in support of design of energy-efficient buildings. In *Building SimulationâĂŹ99*, Vol. 2. 661–667.
[7] Arka A Bhattacharya, Dezhi Hong, David Culler, Jorge Ortiz, Kamin Whitehouse, and Eugene Wu. 2015. Automated metadata construction to support portable building applications. In *Proceedings of the 2nd BuildSys*. ACM, 3–12.
[8] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4 (1996), 129–145.
[9] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.
[10] U. DOE. 2013. Better buildings challenge. http://www4.eere.energy.gov/challenge/sites/default/files/uploaded-files/may-recognition-fs-052013.pdf
[11] Pinar Donmez and Jaime G Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *CIKM*. ACM, 619–628.
[12] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM KDD*. ACM, 259–268.
[13] Varick L Erickson, Stefan Achleitner, and Alberto E Cerpa. 2013. POEM: Power-efficient occupancy-based energy management system. In *IPSN*. ACM, 203–216.
[14] Meng Fang, Jie Yin, and Dacheng Tao. 2014. Active learning for crowdsourcing using knowledge transfer. In *Twenty-Eighth AAAI*. 1809–1815.
[15] Thomas S Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics* (1973), 209–230.
[16] Romain Fontugne, Jorge Ortiz, Nicolas Tremblay, Pierre Borgnat, Patrick Flandrin, Kensuke Fukuda, David Culler, and Hiroshi Esaki. 2013. Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. In *IPSN*. IEEE, 129–140.
[17] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine learning* 28, 2-3 (1997), 133–168.
[18] Jingkun Gao, Joern Ploennigs, and Mario Berges. 2015. A data-driven meta-data inference framework for building automation systems. In *Proceedings of the 2nd ACM BuildSys*. ACM, 23–32.
[19] Dezhi Hong, Quanquan Gu, and Kamin Whitehouse. 2017. High-dimensional time series clustering via cross-predictability. In *Artificial Intelligence and Statistics*. 642–651.
[20] Dezhi Hong, Jorge Ortiz, Kamin Whitehouse, and David Culler. 2013. Towards automatic spatial verification of sensor placement in buildings. In *BuildSys*. ACM.
[21] Dezhi Hong, Hongning Wang, Jorge Ortiz, and Kamin Whitehouse. 2015. The building adapter: Towards quickly applying building analytics at scale. In *Proceedings of the 2nd ACM BuildSys*. ACM, 123–132.
[22] Dezhi Hong, Hongning Wang, and Kamin Whitehouse. 2015. Clustering-based active learning on sensor type classification in buildings. In *CIKM*. ACM, 363–372.
[23] Merthan Koc, Burcu Akinci, and Mario Bergés. 2014. Comparison of linear correlation and a statistical dependency measure for inferring spatial relation of temperature sensors in buildings. In *BuildSys*. ACM, 152–155.
[24] Jason Koh, Bharathan Balaji, Vahideh Akhlaghi, Yuvraj Agarwal, and Rajesh Gupta. 2016. Quiver: Using control perturbations to increase the observability of sensor data in smart buildings. *arXiv preprint arXiv:1601.07260* (2016).
[25] Jason Koh, Bharathan Balaji, Dhiman Sengupta, Julian McAuley, Rajesh Gupta, and Yuvraj Agarwal. 2018. Scrabble: transferrable semi-automated semantic metadata normalization using intermediate representation. In *BuildSys*.
[26] Jason Koh, Dezhi Hong, Rajesh Gupta, Kamin Whitehouse, Hongning Wang, and Yuvraj Agarwal. 2018. Plaster: An integration, benchmark, and development framework for metadata normalization methods. In *BuildSys*. ACM, 1–10.
[27] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
[28] Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. 2015. A bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SDM*. SIAM, 181–189.
[29] Christina S Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. 2004. Mismatch string kernels for discriminative protein classification. *Bioinformatics* 20, 4 (2004), 467–476.
[30] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J Franklin. 2016. Crowdsourced data management: A survey. *IEEE TKDE* 28, 9 (2016), 2296–2319.
[31] Hongwei Li and Bin Yu. 2014. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086* (2014).
[32] Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd WWW*. ACM, 165–176.
[33] Shuheng Li, Dezhi Hong, and Hongning Wang. 2020. Relation Inference among Sensor Time Series in Smart Buildings with Metric Learning. (2020).
[34] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review* 33, 4 (2010), 275–306.
[35] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. 2015. Combining crowd and expert labels using decision theoretic active learning. In *Third AAAI conference on human computation and crowdsourcing*.
[36] U.S Department of Energy. 2019. Better Buildings Initiative Progress Report.
[37] Jorge Ortiz, Catherine Crawford, and Franck Le. 2019. DeviceMien: network device behavior modeling for identifying unknown IoT devices. In *IoTDI*. 106–117.
[38] Marco Pritoni, Arka A Bhattacharya, David Culler, and Mark Modera. 2015. Short paper: A method for discovering functional relationships between air handling units and variable-air-volume boxes from sensor data. In *BuildSys*. ACM, 133–136.
[39] Anika Schumann, Joern Ploennigs, and Bernard Gorman. 2014. Towards automating the deployment of energy saving approaches in buildings. In *BuildSys*.
[40] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244–258.
[41] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence* 214 (2014), 89–111.
[42] Weimin Wang, Michael R Brambley, Woohyun Kim, Sriram Somasundaram, and Andrew J Stevens. 2018. Automated point mapping for building control systems: Recent advances and future research needs. *Automation in Construction* 85 (2018).
[43] Peter Welinder and Pietro Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 25–32.
[44] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. 2011. Active learning from crowds.. In *ICML*, Vol. 11. 1161–1168.
[45] Jing Zhang, Victor S Sheng, Jian Wu, and Xindong Wu. 2015. Multi-class ground truth inference in crowdsourcing with clustering. *IEEE TKDE* 28, 4 (2015).
[46] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.