Improving Experience Sampling with Multi-view User-driven Annotation Prediction

Jonathan Liono*, Flora D. Salim*, Niels van Berkel†, Vassilis Kostakos† and A. K. Qin‡
*Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia
†The University of Melbourne, Melbourne, Australia
‡Department of Computer Science and Software Engineering, School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia
{jonathan.liono,flora.salim}@rmit.edu.au;
nielsv@student.unimelb.edu.au; vassilis.kostakos@unimelb.edu.au; kqin@swin.edu.au

***Abstract*—A fundamental challenge in real-time labelling of activity data is user burden. The Experience Sampling Method (ESM) is widely used to obtain such labels for sensor data. However, in an in-situ deployment, it is not feasible to expect users to precisely label the start and end time of each event or activity. For this reason, time-point based experience sampling (without an actual start and end time) is prevalent. We present a framework that applies multi-instance and semi-supervised learning techniques to perform to predict user annotations from multiple mobile sensor data streams. Our proposed framework estimates users' annotations in ESM-based studies progressively, via an interactive pipeline of co-training and active learning. We evaluate our work using data collected from an in-the-wild data collection.**

***Index Terms*—Annotation prediction, Experience Sampling Method, User annotation, Human activities, User-driven data collection, Experience improvement, Semi-supervised learning, Multi-view learning, Multi-instance learning, Mobile sensing**

I. Introduction

The Experience Sampling Method (ESM) [1] provides opportunities to record ground-truth data through self-reports (i.e. annotations) from the participants in a data collection campaign. Originally, ESM was widely used in the domain of psychological research, for example [2]. However, it has offered significant benefits for ubiquitous computing research in recent years, for example, emotion recognition [3], mobile user intimacy and smartphone usage [4], [5], human activity recognition [6], lifelogging [7]–[10] and mobile sensor data collection [11]–[14].

ESM can be configured in various ways, such as having either regular or intermittent sampling. Inherently, human annotations acquired from ESM in a pervasive sensing environment can be associated with their past or recent activities, events, social encounters and spatiotemporal contexts (e.g. proximity of locations and surrounding point-of-interest (POI) categories in a certain time segment). These annotations can be requested based on specific events or changes of sensor signals. Asking users to annotate their activities, events and contexts while these are ongoing can be challenging because of users' subjective mental states and cognitive workload during the annotation processes. Moreover, identifying such changes or events in the data streams is also a significant challenge because of the reliability of these perceived human annotations in the wild. A less subjective way to identify specific activities or events can be performed with a restricted experiment setting. In this case, these events can be distinguished based on sudden changes in multidimensional sensor channels or streams, such as fall detection [15] and human activity recognition [16], [17].

In a typical application of in-the-wild data collection, ESM must be performed in a low-burden manner to produce a higher rate of compliance [18]. To perform a specific task in daily life, retrospective memory [19] is an essential aspect of remembering previous events or human activities, which can also affect the process of annotation in a real-world scenario. Ideally, an annotation should be attained interactively through a ubiquitous instrument (e.g. surveys through mobile apps), given the possibility of undefined time boundaries for such activities and the contextual information recorded. For example, daily annotations can be performed by users as they perform their activities.

Our proposed framework applies multi-instance learning (MIL) to the features extracted from the multivariate sensor data, which correspond to the recent time duration of a given user's annotation. Additionally, a semi-supervised learning component corresponds to the usage of both co-training and active learning to predict and improve the annotation classifier progressively. In this case, the aim of annotation prediction is for an ESM system to be confident to obtain the next annotation interactively through accurate inference of user context. Consequently, the direct implication of our contribution is targeted towards process optimisation in ESM-based data collection — in particular, by reducing the burden of annotations (e.g. minimising choice overload in a survey form).

Our pioneering work shows its effectiveness in reducing the burden during an ESM study by predicting user annotations just before ESM-based surveys are triggered. Further, its capability in progressive learning is based on active feedback from its corresponding user and a variety of sensor data streams from mobile devices. The outcome of our work considers the following aspects in mobile sensor data collection (especially the in-the-wild data collection and sensing applications that rely on ESM-based annotations):

- Our framework can **predict user annotations** during

978-1-5386-9148-9/19/$31.00 ©2019 IEEE

an ESM study, and it enables the model to **adapt** progressively based on a mutual agreement between co-trained models from heterogeneous data sources (mobile sensors). In other words, a semi-supervised learning approach is applied to the small amount of labelled data during bootstrapping, which aims to predict the annotation accurately before an annotation (e.g. ESM-based survey) is requested from the mobile user. Consequently, the model can evolve progressively (through a model re-training mechanism) based on the inclusion of newly unlabelled data in the training pool.

- As a result of semi-supervised learning, our work is resilient to **missing sensor data**. For example, the light sensor in a smartphone might not always be available during a human activity performed just before the user is requested to participate in an ESM-based survey. **Multi-view** (i.e. co-training) and **active** learning approaches are applied to feature subsets of the unlabelled sequences streamed from available sensors at the time of annotation prediction.
- The **design considerations** are important, to improve the interaction and engagement of prevalent ESM-based surveys for user-driven mobile data collection in-the-wild. Hence, our initial work aims to provoke the ubiquitous computing research to increase the reliability and quality of annotations by providing context-aware human-computer interaction in intelligent applications.

II. Related Work

Experience Sampling Method. The ESM is a prevalent approach used in many domains [6], [20]–[23] to recall recent or past activities of a user. Its reliability and validity have been empirically studied in [1], which provides convincing results for the labels (activities) that are obtained through a systematic random sampling of daily life. Experience-Sampling Forms (ESF) can be easily embedded in mobile phone applications. As Csikszentmihalyi and Larson detailed in [1], ESF is typically designed for a short (in-situ) survey or self-report questionnaire that should take no more than two minutes to complete. Many studies in recent years have focused on reducing the cognitive workload of the ESM by leveraging the unique characteristics of mobile users' behaviours or activities—for example, ESM that is driven by micro-usage of mobile applications [24] and break-points between a user's activities [25]. According to [3], the experience sampling could be triggered for the mobile users from the signal, event or time (at regular intervals). Moreover, the users in [6] self-annotated the start and end time for before and after their activities. However, these types of data collection typically require the users to be actively engaged in defining the start time and end time of their activities. When the data collection is performed through participatory or opportunistic sensing [26] in the wild (such as daily commuting journeys), users may forget to define the end time of the activities due to their environmental contexts and the constant distractions within their vicinity. In several cases, experience sampling can be performed to ask about the recent or current activity of a user without strictly defining the start and end time of activity. Hence, it is inherently challenging to extract relevant data related to each experience sampling label recorded at a particular timestamp (i.e. point-based experience sampling) and build suitable models to predict the annotations ahead of time.

In this paper, the challenge of annotation prediction is inherently different to a forecasting problem. Annotation prediction refers to the classification of a label just before a user is presented with information that may be relevant to the final prediction output (e.g. ESM-based surveys where the questions can be relevant to recent user activities). In contrast, a forecasting problem is targeted towards the future occurrence of the annotations. Minor and Cook [27] proposed an activity forecasting method to predict the expected time until a target activity occurs using a regression tree classifier. In fact, such a method could also be leveraged to infer when is the best time to prompt the user for an ESM-based survey.

Multi-instance and Multi-view Learnings. MIL can be used to tackle problems in behavioural studies where the boundary of target labels is unclear because of subjective experience during the user's annotations at those moments. Typically, the research problems in this space are formulated so that data can be continuously streamed, which can then be organised into bags for inference purposes.

In a real-world scenario of mobile sensor data collection, the availability of reliable training data is often seen as a critical issue for building a better predictive model. In this case, building a classifier based on small subsets of data might not be enough for accurate prediction of ESM annotations because they might also be influenced by the mobile user's activities and environmental contexts. In [28], semi-supervised learning was used to solve the multi-instance problem by treating instances in the positive bags as unlabelled data. A common semi-supervised method that has been used in real-world applications is co-training [29], which allows the training of two distinct classifiers from multi-view perspectives by labelling unlabelled instances for each other. For instance, this concept has then been adapted to the application of activity recognition in [30]. In ubiquitous environments, sensor data can be collected through streaming from multiple sources. Hence, a multi-view perspective is needed for the inference of subjective human behaviours. In [31], multi-task multi-kernel learning (MTMKL) exploits the kernel functions that are represented from different views or modalities for affective computing studies. Due to its single task objective, MTMKL does not suit the purpose of annotation prediction for ESM-based surveys. Co-training was also applied in multi-transfer [32] for cross-domain knowledge transfer. Since annotation prediction in a typical ESM process is targeted to one domain, such a transfer learning technique may not be feasible in our case.

Active Learning. Inherently, a model can be improved progressively by reliable annotations (ground-truth) during the data collection process. This improvement can be achieved by the application of active learning, to determine the most informative points based on direct feedback from a user.

In [33], active learning was applied to the annotation process in a crowdsourcing scenario in which multiple annotators were required to provide their own activity labels. However, this solution could be over-generalised since they are generally used for determining informative sensing data on a specific community of individuals. In our case, the daily activities are more tailored to each person for personal intelligent mobile sensing (i.e. first-person based activity recognition). Moreover, the true label complexity in the authors' proposed framework was heavily dependent on the number of clusters derived from unlabelled data instances. In a typical ESM-based survey, this complexity can be simplified since the true label is obtained based on the mental state of the user at a given time. To the best of our knowledge, we are the first to investigate and propose a continuous learning framework for predicting annotations in ESM, using multi-view multi-instance learning.

III. Methodology

We address the following main research question: *Given an ESM-based annotation acquired from time-point based experience sampling, can a smartphone predict the annotation just before an ESM-based survey is presented to the user?*

A. Problem Formulation

We first formulate the problem we are addressing, in terms of human activities and contextual information captured in an ESM study. Hence, we define the following notations:

Let $\boldsymbol{S} = \{\boldsymbol{S_1}, \boldsymbol{S_2}, ..., \boldsymbol{S_n}\}$ be the set of sensors available during the collection of data on a mobile user, where i is the index of i-th sensor, $\boldsymbol{1 \leq i \leq n}$ and $\boldsymbol{n}$ is the total number of sensors. Let sensor $\boldsymbol{S_i}$ be the source of time series data containing sequences of real-valued numbers. It should be noted that the time series data streamed from $\boldsymbol{S_i}$ could be composed of multiple time series (e.g. an accelerometer sensor that produces the reading of acceleration in x, y and z axes, and its magnitude).

Let the discrete label $\boldsymbol{a}$ be a unique member of a label set $\boldsymbol{A} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, ..., \boldsymbol{a_d}\}$, where $\boldsymbol{d}$ is the number of unique experience sampling labels $\boldsymbol{a}$ in $\boldsymbol{A}$.

Let $\boldsymbol{S_{ia}} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, ..., \boldsymbol{s_m}\}$ be the particular time series streamed by sensor $\boldsymbol{S_i}$ in which a point-based experience sampling label $\boldsymbol{a}$ exists after the last instance (i.e. $\boldsymbol{j = m}$), where j is the index of j-th instance in the observed time series $\boldsymbol{S_{ia}}$, $\boldsymbol{1 \leq j \leq m}$ and $\boldsymbol{m}$ is the length of $\boldsymbol{S_{ia}}$ within a certain time-interval boundary $\boldsymbol{t_\Delta \leq t_\delta}$ before the occurrence of $\boldsymbol{a}$ and $\boldsymbol{t_\delta}$ is a constant for a maximum time range of observed time series for $\boldsymbol{a}$.

Consider this scenario. The time series of magnitude for the accelerometer sensor contains two annotations (see Figure 1). Let $\boldsymbol{t_\delta}$ be a constant of 30 minutes that results in the observed time series with the duration of 30 minutes before the annotation 'Bus Riding' (i.e. $\boldsymbol{t_\Delta = t_\delta}$). However, the duration of 'Light Rail Riding' is less than 30 minutes (i.e. $\boldsymbol{t_\Delta = t_\delta - z}$) since the time portion $\boldsymbol{z}$ of $\boldsymbol{t_\delta}$ belongs to 'Bus Riding'.

In a scenario of ESM-based surveys that are triggered at particular time points, the experience sampling labels are given by the users. Hence, we formulate the problem in which labelled data are scarce while not all sensors are available within the duration of t_δ before the ESM-based survey is triggered. Let us consider the following application scenario in which the mobile app is constantly recording sensor data in the background. If the annotation can be predicted correctly before the app notifies the mobile user, an interactive survey form can be constructed based on such intelligent inference. Hence, a simple binary choice can be presented instead of having potential overloaded options that may disengage or demotivate the user to contribute high-quality annotations.

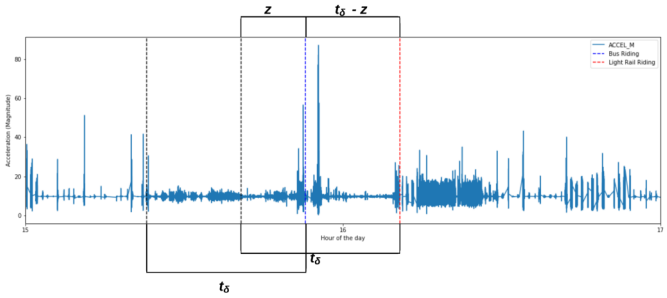

Fig. 1. Point-based experience sampling label problem for user annotations. Blue ('Bus Riding') and red ('Light Rail Riding' dashed lines are the ESM data points (i.e. point-based experience sampling labels).

Therefore, the problem of annotation prediction is formulated as follows: given an unlabelled time series set for all sensors $\boldsymbol{S_u} = \{\boldsymbol{S_{1u}}, \boldsymbol{S_{2u}}, ..., \boldsymbol{S_{nu}}\}$ within the constrained time interval $\boldsymbol{t_\Delta}$, predict the annotation that the mobile user will choose during an ESM process, where $\boldsymbol{i}$ corresponds to $\boldsymbol{i}$-th sensor $\boldsymbol{S_i}$ of $\boldsymbol{S_{iu}}$ and $\boldsymbol{1 \leq i \leq n}$.

Let a_u be the label to be predicted for the recent time range $\boldsymbol{t_\Delta}$ containing $\boldsymbol{S_u}$. Hence, the objective of annotation prediction is to accurately classify a_u from A (i.e. $\boldsymbol{a_u}$ in $\boldsymbol{A}$).

B. Implementation

In a typical ESM scenario, a robust and progressive model is needed to predict the annotation just before the user is asked. Therefore, we design a framework based on the assumption that only a small amount of data are available for those annotations. In other words, there exists the initial subset of data corresponding to each member a of A. Here we present a semi-supervised framework (CoAct-nnotate) to predict a user's experience sampling labels at the time they are about to be requested. Thus, our framework aims to predict users' ESM annotations and continuously learn to improve the model over time.

An overview of CoAct-nnotate's architecture is presented in Figure 2. This framework consists of multi-instance and semi-supervised modules. Instances from the mobile sensors are organised into bags where the representative features of each bag need to be extracted in the multi-instance module. A classifier is then trained for each data source (i.e. each mobile sensor). These initially trained classifiers are based on a small subset of data. For example, training of a classifier is based on the first occurrence (instances in the first bag) of a

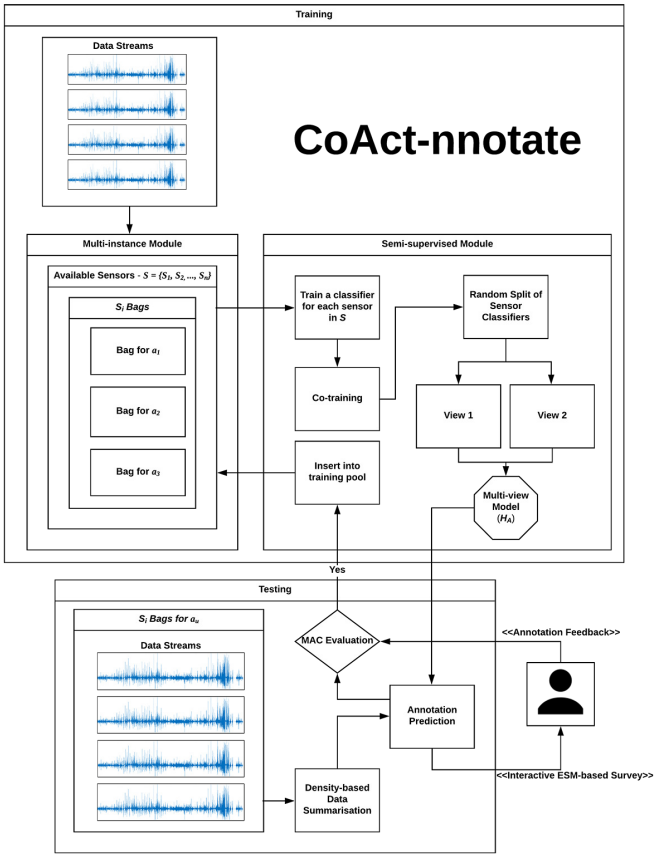

Fig. 2. **CoAct-nnotate**: user-driven annotation prediction framework for mobile experience sampling labels.

particular annotation. Next, the semi-supervised module aims to improve the overall performance of annotation prediction based on the inputs of predicted annotations in multi-view perspectives (from co-trained classifiers).

C. Multi-instance Learning for Experience Sampling Labels

To address the loosely-coupled nature of experience sampling labels on the recent streaming of sensor data, MIL is applied, whereby the boundary of an annotation is weakly assumed on a sequence of training instances. In a typical task of MIL, the ultimate aim is to predict a class label from a bag of instances, which contains at least one positive instance for the true label. As shown in Figure 3, the process of MIL is generalised to allocate all instances from each sensor into a bag first, which is labelled as a. For learning and prediction purposes, feature bags for a are prepared through feature construction and extraction. In our work, feature construction refers to the process of creating new information that can be derived from instances within the dimension of all mobile sensors S, for instance, the magnitude of acceleration that can be computed from all three axes of x, y and z from the accelerometer of a wearable device.

Consequently, feature extraction corresponds to the derivation of new information through a mapping function. This process is typically performed in a time interval manner (e.g. extracting features from temporal and frequency domains within a given time window). Thus, the final product of the MIL component in our proposed framework is the set of *feature bags*, which will be used for learning and prediction purposes. A feature bag refers to a representation of a multi-instance set. Each bag contains instances of extracted features (from a particular sensor). Each set of feature bags (for all sensors) is associated with an annotation.

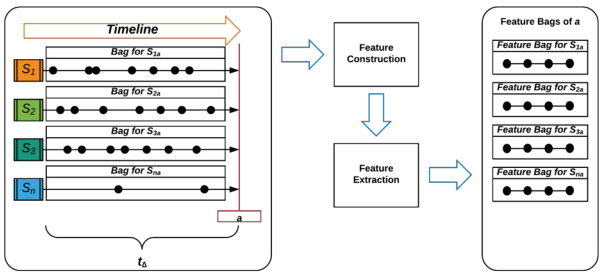

Fig. 3. Workflow of multi-instance learning of instance streaming from multiple mobile sensors.

A sensor feature bag for label a is represented as S_{ia} throughout this paper. The purpose of this process is to derive the sets of representative data from sensors with respect to all possible annotations A. In the CoAct-nnotate framework, we propose one classifier should be trained on each set of sensor feature bags of A. In other words, there would be at least one classifier trained for each mobile sensor. This approach is preferred due to the real-world scenario where there would be the possibility of no data (instance) to be streamed from a particular sensor S_i within a recent time duration t_Δ.

In this paper, MIL problem entails the aim to predict a bag of unlabelled data containing the final product of feature construction and extraction processes (refer to Figure 3). Hence, we define the MIL prediction problem as follows:

Let us define unlabelled feature bags S_u, where $\boldsymbol{Su} = \{\boldsymbol{S_{1u}}, \boldsymbol{S_{2u}}, ..., \boldsymbol{S_{nu}}\}$, i is the index of i-th feature bag for i-th sensor, $\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n}$ and $\boldsymbol{n}$ is the total number of feature bags. A feature bag S_{iu} can contain no feature instances (i.e. $count(S_{iu}) \geq 0$).

Feature instances in a feature bag is defined as a set $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_l}\}$, k is the index of k-th feature instance in a bag, $\mathbf{1} \leq \boldsymbol{k} \leq \boldsymbol{l}$ and $\boldsymbol{l}$ is the total number of instances in the feature bag.

A classifier H_i is used to predict the annotation/class label of S_{iu}, where $\boldsymbol{H_A} = \{\boldsymbol{H_1}, \boldsymbol{H_2}, ..., \boldsymbol{H_n}\}$, i is the index of i-th classifier for i-th sensor, $\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n}$ and $\boldsymbol{n}$ is the total number of sensor classifiers.

In a real-world setting, a sensor may be unavailable or turned off by users. For instance, a user may turn off the Bluetooth and Wi-Fi sensors or location services to preserve her smartphone's battery. Therefore, the condition of $count(S_{iu}) \geq 0$ holds a conclusive inference when an experience sampling label a may have no entry of feature instances computed within the recent t_Δ.

As a sensor may stream no data for a, the MIL component in CoAct-nnotate trains a classifier for each sensor on the feature instances (contained in feature bags for all experience sampling labels A). Consequently, each feature instance in the unlabelled sensor feature bag S_{iu} can be predicted for its annotation by the posterior probability $Pr(y|X)$ of trained classifier H_i. Ultimately, annotation prediction can be performed on the unlabelled bag S_{iu} by gaining consensus of annotation for all its feature instances. The simplest form of the consensus is the majority voting mechanism, which is used in our CoAct-nnotate implementation in this paper. In other words, the bag labels can be defined as $y_{iu} = \max_l(y_{iul})$, where y_{iul} are the instance labels inferred from S_{iu} using H_i and y_{iu} is the product of annotation prediction inferred from maximum count function over all y_{iul}. Our CoAct-nnotate framework is not restricted to this maximum inference function for the annotation prediction.

D. Co-training of Sensor Classifiers

In everyday settings, the availability of sensors and annotations is one of the primary roadblocks to enable intelligent sensing and ESM applications. By nature, the signals that are streamed from the sensors embedded in a smart device (i.e. smartphone) can characterise the traits of human activities and their contextual information, which can be analysed and differentiated in a multifaceted perspective (i.e. multi-view annotation prediction from heterogeneous sensor streams).

To build a multi-view model for annotation prediction, the co-training approach is applied in our framework by randomly allocating sensor classifiers to two distinct views (refer to Figure 2). In theory, co-training (also known as co-regularisation) [34] is a multi-view consensus learning approach that leverages two feature representations (i.e. 'views') to minimise the misclassification rate by maintaining the consistency of classification decisions from two independent classifiers. Hence, this semi-supervised learning approach is adapted to our problem to predict annotations reliably from the two distinct views of heterogeneous sensor classifiers.

In this case, the splitting of sensor classifiers should be performed evenly into two subsets (corresponding to first and second views). Hence, these two sets of classifiers would be used as a joint-model to predict an annotation for unlabelled bags of sensor data. The multi-view model evolves by including the sample of unlabelled data in the training pool (to rebuild the classifiers) upon mutual agreement between the two views. The main objective of co-training, in this case, is to improve the performance of classifiers by mutual agreement of predicted annotations from two views. Inherently, the consensus of annotation prediction in a view should be achieved via an intrinsic mechanism to select the predicted annotation amongst all instances in each sensor feature bag. Hence, the simplest form that can be used is majority voting from predicted class labels from all sensor bags (i.e. $y_{uv} = \max_i(y_{iu})$).

Given the view V, y_{iu} corresponds to the inferred annotation of the unlabelled sensor feature bag S_{iu} contained in V and $y_u v$ is the product of annotation prediction from maximum count function over all y_{iu} in V. V is a general representation of view for either first view V_{first} or second view V_{second} in the co-training process of CoAct-nnotate's semi-supervised module (as shown in Figure 2).

For an unlabelled bag S_u, the prediction of annotation can be performed with a three-step process: summarisation of instances in unlabelled bags, prediction of annotation and improvement of the overall multi-view model based on the evaluation of mutual agreement of classification decisions.

E. Data Summarisation of Feature Instances

Since the number of instances contained in the sensor feature bags can be unpredictable (given the natural settings of mobile data collection), it is important to derive the representative instances that can be used for prediction (which can also be included in the training pool for the progressive improvement of the proposed multi-view model). In this case, data summarisation is leveraged to derive representative instances by clustering the instances of features for one sensor bag based on density measures.

For the unlabelled time series of a sensor S_{iu}, data summarisation is performed before annotation prediction. We employ a density based data summarisation based on cluster change of sequential instances of the sensor data. Previously in [35], density based data summarisation has been studied to maintain reliable inter-rater agreement between machine learning models while inducing a high space saving ratio. In this case, such space saving factor provides beneficial inputs for the co-training mechanism of training sensor classifiers to allow progressive learning over time and according to the mobile user's behaviour in the wild and dynamic environments.

As a result of the data summarisation process in the proposed CoAct-nnotate framework, representative features can be ultimately obtained in a compact form. This compact form is then used for multi-view annotation prediction. In several cases, the direct benefit can be directed towards the model that may require more time for prediction, such as nearest neighbours based classifiers.

The product of this data summarisation process is not only beneficial for multi-view annotation prediction but also to improve the overall performance of a multi-view model with less data to be included in the proceeding training phase (after the process of annotation prediction and active feedback obtained from the user). Without the data summarisation process, the time taken for multi-view model re-training would be exponential when the system is deployed and used progressively.

F. Multi-view Annotation Prediction

Once the sets of summarised bags ($summarisedBags_{first}$ and $summarisedBags_{second}$) are acquired through the process in Section III-E, the subsequent objective of CoAct-nnotate is to predict the annotation accurately. The prediction can be achieved in a multi-view approach, utilising the concept of co-training by allowing sensor classifiers that were previously trained to predict the annotation for a given set of summarised bags, corresponding to a particular view. Let us denote y_{first} as the predicted annotation for $summarisedBags_{first}$ and y_{second} as the predicted annotation for $summarisedBags_{second}$. The corresponding sensor classifiers in a view will predict the annotation according to a consensus mechanism in the bags. Consequently, the concept of co-training is applied to improve the overall prediction model. This enables the views to benefit each other by being able to continuously learn or improve the sensor classifiers based on the mutual agreement of predicted annotations from each view. The posterior probability of predicted annotation (i.e. $Pr(y|X)$) from a view can be acquired by the inference of all predicted annotations of sensor feature bags in the corresponding view.

Irrespective of whether a mutual agreement ($y_{first} == y_{second}$) is reached or not, the posterior probability ($Pr(y_{agreed}|X)$) of the multi-view annotation predictor can be inferred from the highest posterior probability of the two views, either $Pr(y_{first}|X_{first})$ or $Pr(y_{second}|X_{second})$. Moreover, if there is no mutual agreement between the two views, the predicted annotation can be inferred from the view that has the highest posterior probability. Conversely, a random selection process would be used if the posteriors are equivalent (i.e. $Pr(y_{first}|X_{first}) == Pr(y_{second}|X_{second})$). At the end of the prediction procedure for the summarised bags, a special evaluation module is included to determine whether the summarised bags need to be thrown to the training pool where each sensor classifier could be re-trained after performing the annotation prediction. The output of this evaluation is denoted as $macEvaluated$. We call this process **M**utually **A**greed **C**onfidence (MAC) evaluation, whereby its binary value is based on the condition of mutual agreement in the multi-view prediction, and either the posterior probability ($Pr(y_{agreed}|X)$) is below a given threshold β or there is a disagreement between predicted annotation and true annotation. Hence, the purpose of this MAC evaluation module is to determine the needs to improve the sensor classifiers if the confidence level of multi-view annotation prediction is insufficient. Conclusively, the binary output of MAC evaluation can be expressed with the following equation:

$$macEvaluated = M_A \cdot ceil\left(\frac{ceil(\beta - Pr(y_{agreed}|X)) + D_A}{2}\right) \quad (1)$$

where M_A is the binary value indicating a mutual agreement occurrence (i.e. $M_A \in \{0, 1\}$), β is the parameter threshold for confidence evaluation of the posterior $Pr(y_{agreed}|X)$ on the predicted annotation, and D_A is the binary value indicating a disagreement between the predicted annotation and true annotation (i.e. $D_A \in \{0, 1\}$). In this case, the true annotation refers to the actual label provided by the user through an active feedback mechanism (i.e. an answer to the ESM-based survey). Consequently, this true annotation is also used for the following active learning component to improve the classifiers in CoAct-nnotate (explained in Section III-G below).

G. Improvement of Sensor Classifiers

In this section, the process of improving sensor classifiers is elaborated in detail, given the intrinsic output (i.e. $macEvaluated$) produced in the previous process (annotation prediction in Section III-F). In a real-world scenario, we take the input from the mobile user as the consideration to improve the performance of sensor classifiers for the multi-view annotation prediction. The acquisition of such input is based on the data collected by the ESM protocol, which inherently conducts a query of annotation feedback from the mobile user in an interactive manner. Hence, the process of improvement for sensor classifiers is based on the binary condition of $macEvaluated$ with an additional input $userAnnotation$ acquired from the mobile user's feedback. Within the improvement process in the co-training module of CoAct-nnotate, active learning is applied whereby the true annotation is obtained through the ESM process and is used as the expected annotation to label S_u, for which the contained bags need to be included in the training pool. Inherently, the usage of semi-supervised learning in CoAct-nnotate (consisting of co-training and active learning) is applicable for both generative and discriminative base classifiers of the respective sensor feature bags. When $macEvaluated$ returns zero, there would be no improvement process undertaken by CoAct-nnotate. In other words, the feature bags (with the user's annotation) will not be included in the training pool.

To resolve the potential issue of data imbalance in the summarised unlabelled bags (i.e. $summarisedBags$), upsampling (or oversampling) is applied to each sensor feature instance in the corresponding bag, thereby increasing the number of possibly important data points (i.e. feature instances) within a summarised sensor bag. The simplest form of upsampling is the duplication of a feature instance. In this case, k-number of duplication is applied to a given summarised feature instance, where k is obtained from a Poisson distribution with a rate parameter δ (i.e. Poisson(δ)). The method of upsampling is not restricted to instance replication because other forms, such as generative approaches of sampling (also known as generative oversampling [36], [37]), can be performed on a given feature instance (extracted from a sequence of feature instances in a summarised sensor bag). Ultimately, these upsampled bags (labelled with $userAnnotation$) are then added to $TrainingPool$ to re-train all sensor classifiers.

IV. Experimental Evaluation

A. Dataset

We use the CrowdSignals dataset [11] for our analysis. This dataset contains rich sensor data from smartphones and wearables in the wild (annotated by participants). In our experiment, the prediction of annotations is based on multiple sensors in Android smartphones. For the construction of instances in a bag, the temporal value of t_δ is set to 30 minutes. Thus, each sensor bag in the MIL phase contains at least the data points within the duration of t_Δ. For the standard approach of preparation to train the classifiers, we use a window size of one-minute time interval with 50% overlapping windows of temporal segmentation.

The CrowdSignals dataset consists of daily logs for more than 30 Android smartphone users. In our analysis, the datasets of nine participants are sampled for the experiment, and timestamped ESM labels are extracted from their data. Using these labels, we simulate a scenario in which the users are asked to respond to the ESM questions, at the time of these timestamped labels. Although only smartphone sensor data are used within the scope of our experiment, it should be noted that other data sources (e.g. smartwatches, wearable sensors) could be used to enrich the contextual inference to enable better annotation prediction.

Given the rich amount of data collected in the CrowdSignals campaign, we leverage the following sensor data: *Accelerometer, Gyroscope, Magnetic field, Rotational vectors, Battery, Light, Screen status, Step counter* and *Pressure*.

The following ESM labels (annotations) are derived from the end timestamps of time-interval labels that the participants recorded: *Riding bus, Riding train, Riding light rail, Riding ferry, Riding in a car, Riding a bicycle, Riding an elevator, Riding an escalator, Riding a Scooter, Walking, Walking on stairs, Drinking water* and *Playing video game*.

B. Experimental Setup

During the initial training process of each sensor classifier, only one sensor bag is used per ESM label provided by a mobile user. Our model is trained with a limited amount of sample data for all labels (i.e. one bag per class label), which then need to perform annotation prediction progressively throughout the simulated data collection in a day-to-day manner. In other words, the objective of the experiment is to perform annotation prediction accurately based on the streaming of multidimensional sensor data during an ESM study, given the influence of in-situ contexts of the mobile user. Consequently, this experiment compares the performance of annotation prediction by general approaches with our proposed semi-supervised approach.

In our work, the density-based bag summarisation component employs the same strategy as [38], [39] and [35] by setting the parameters $Eps = 0.3$ and $min_{pts} = ln(n)$ for the given DBSCAN algorithm (for density-based clustering), where n is the number of feature instances in an unlabelled sensor feature bag S_{iu}. In the co-training process, a random split operation is performed proportionally on the set of sensor classifiers to produce two different views V_{first} (View 1) and V_{second} (View 2). In this case, the number of distinct sensor classifiers in a view is at least $(n/2)$. At the end of the annotation prediction process, the binary value of MAC evaluation is calculated under the condition of a mutual agreement between the views of sensor classifiers where $y_{first} == y_{second}$, and its agreeable posterior (i.e. $Pr(y_{agreed}|X)$) is being under a certain threshold $\beta = 0.9$. Therefore, a MAC evaluation is considered valid when it satisfies the output of Equation 1 where $macEvaluated == 1$. Before the summarised sensor bags are added to $TrainingPool$ (given a valid MAC evaluation) for the sensor classifiers to be re-trained, the upsampling operation is performed on the summarised sensor bag by using the k-number of the instance replication strategy, where k is withdrawn from the $Poisson(\delta)$ distribution with $\delta = 5$. To simulate the active learning component of the semi-supervised module in CoAct-nnotate, we leverage the actual annotation at the end time of the time interval based on the actual user labels in the CrowdSignals dataset. For the time duration of recent sensor data on the given annotation a, 30 minutes of past mobile sensor data (i.e. $t_\delta = 30$ minutes) are used to construct a bag (containing a sequence of raw sensor data) for the respective sensor channel.

Since annotation prediction is crucial for mobile data collection in the wild, we simulate an experiment in which the end time of self-annotation (user-driven labelling) is the time point of ESM annotation. All participants involved in CrowdSignals data collection are mobile users who own Android smartphones. Different phone models are noticeable within the dataset since the capability of smartphones to sense their context and environments varies. Due to the diversity of sensors in different smartphone models, the performance of annotation prediction can be greatly influenced by the limited composition of sensor classifiers contained within a view.

As the base classifier of the mobile sensors, we leverage the following algorithms in our evaluation (using scikit-learn [40]):

- Naive Bayes (**NB**)
- Support Vector Classifiers (**SVC**)
- Multilayer Perceptron (**MLP**) with 0.00001 as the L2 penalty (regularisation parameter), L-BFGS [41] as the solver for weight optimisation and structure of two hidden layers (consisting of five neurons for the first layer and two neurons for the second layer)
- Random Forests (**RF**) with 100 trees
- Decision Tree (**DT**)
- k Nearest Neighbour (k-NN) with $k = 1$ (**1NN**)

For the baseline of annotation prediction, we leverage the general approaches that can be used for annotation prediction as follows:

- Multivariate time-window based annotation prediction (denoted as **MAP**). In the MAP approach, only one classifier is trained for all sensor feature dimensions and instances in $TrainingPool$.

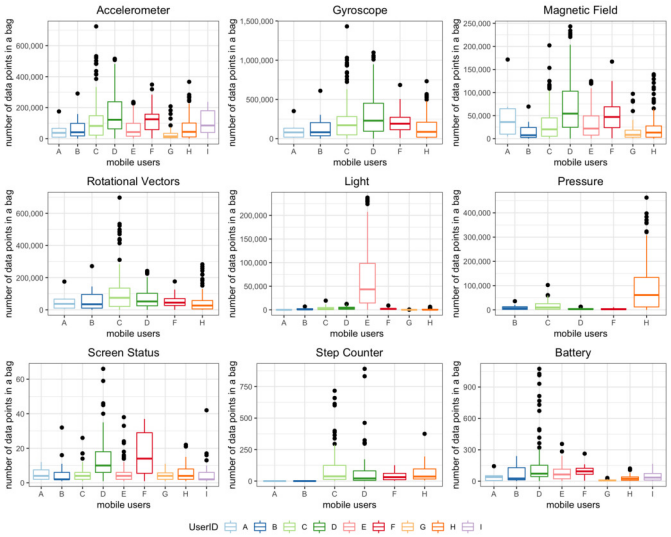

Fig. 4. The granularity of data points in sensor bags.

- Non-multivariate time-window based annotation prediction (denoted as **1C1S**). In 1C1S approach, one classifier is trained per sensor.
- 1C1S with co-training (denoted as **Co-1C1S**). In the Co-1C1S approach, the concept of co-training is applied to perform multi-view annotation prediction. The basic operation of view split is similar to CoAct-nnotate, except for the process of sensor classifiers improvement. For the improvement process, the predicted annotation (i.e. y_{agreed}) is used to label S_u, which will be included in $TrainingPool$ only if there is a mutual agreement (i.e. $M_A == 1$) between two views of sensor classifiers.

For both the MAP and 1C1S approaches, the training of classifiers is based on the bags of all first occurrences of each a in A. In other words, only one bag is used for a class label during the training phase, which results in no progressive learning over time. In contrast, both Co-1C1S and CoAct-nnotate employ the concept of progressive learning by a co-training mechanism. The only difference between Co-1C1S and CoAct-nnotate is in the criteria for sensor classifier improvement and cost-efficient performance of bag summarisation for S_u in CoAct-nnotate. In the feature extraction process of all annotation prediction approaches (MAP, 1C1S, Co-1C1S and CoAct-nnotate), time-interval based temporal segmentation is used for a given bag whereby the size of the time window is set to 60 seconds (1 minute) with 50% overlapping parameters. In terms of general evaluation performance of annotation prediction, the **correctness** metric is used to measure the accuracy of an annotation predictor. Consequently, the correctness metric can be measured by calculating the fraction of the total count of correct predictions over the number of annotation prediction, as expressed in the following equation:

$$Correctness = \frac{\sum_{u=1}^{v} annotation^{u}_{correct}}{v} \quad (2)$$

where v is the total number of annotation predictions and $annotation^{u}_{correct}$ is the binary value whether the u-th annotation prediction is correct or not. To evaluate the performance of the systems empirically, the experiment is performed with 10 iterations per base classifier on each approach.

C. Results

As shown in Figure 4, we leverage nine sensor channels (mentioned in Section IV-A) as the source of data streams, and use those to predict the ESM labels in the dataset.

In our dataset, there are several instances of incomplete sensor channels that are due to the smartphone hardware. For instance, user E's dataset lacks gyroscope, rotational vectors, step counting and air pressure. Although air pressure data are available for users B, C and D, the step counter sensor channel is missing for user B. Similarly, battery information is missing for user C within t_δ (30 minutes) before all occurrence of ESM annotations. Due to the variability of sensors that may be missing and their inconsistent sampling in a given S_u, this increases the difficulty of ESM label prediction. Despite the inconsistent number of data points (with many noticeable outliers) for heterogeneous sensor channels, shown in Figure 4, the time lengths of data points are varied with fewer outliers, as shown in Figure 5. In this case, the time length t_{length} can be computed by $t_{length} = t_{max} - t_{min}$, where t_{max} is the maximum timestamp and t_{min} is the minimum timestamp of data points in a given sensor bag S_{ia}.

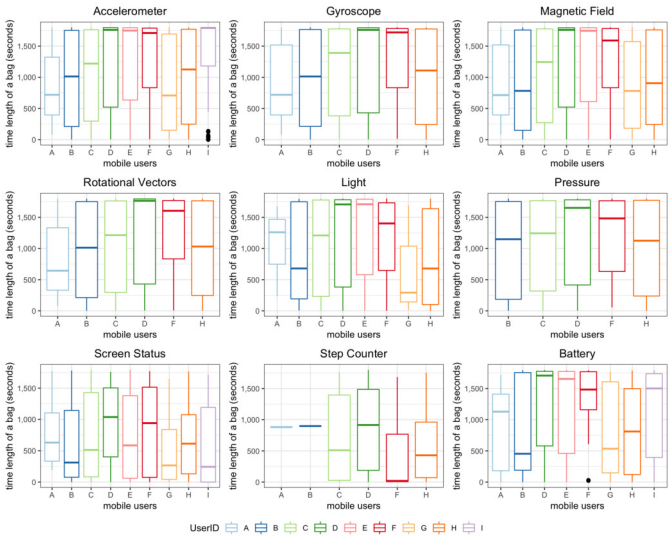

Fig. 5. The time length of data points in sensor bags (seconds).

Although multiple metrics can be used for evaluating the performance of classifiers, we leverage the correctness metric as the dominant measurement for the performance of annotation prediction. As shown in Table I, we believe that one classifier should be trained for each sensor (refer to the 1C1S experiment result). By observing the average correctness values of base classifiers from our iterative experiment, the maximum performance gain of 28.9% is noticeable by training one

TABLE I
CORRECTNESS OF ANNOTATION PREDICTION (NORMALISED FROM 0 TO 1)

User ID	Number of Classes	MAP						1C1S						Co-1C1S						CoAct-nnotate					
		NB	SVC	MLP	RF	DT	1NN	NB	SVC	MLP	RF	DT	1NN	NB	SVC	MLP	RF	DT	1NN	NB	SVC	MLP	RF	DT	1NN
A	3	0.125	0.125	0.125	0.125	0.15	0.375	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.138	0.138	0.188	0.188	0.138	0.125	**0.613**	**0.4**	**0.7**	**0.713**	**0.863**
B	9	0.14	0	0	0.067	0.033	0.087	0.173	0.207	0.18	0.213	0.18	0.167	0.181	0.125	0.131	0.106	0.144	0.144	**0.281**	**0.413**	**0.363**	**0.563**	**0.706**	**0.744**
C	10	0.232	0.246	0.008	0.128	0.129	0.169	**0.304**	0.289	0.297	0.293	0.299	0.288	0.283	0.215	0.016	0.22	0.269	0.275	0.295	**0.515**	**0.437**	**0.727**	**0.785**	**0.818**
D	10	0.175	0.206	0.206	0.121	0.156	0.079	0.281	0.262	**0.275**	0.271	0.279	0.265	**0.357**	0.268	0.17	0.284	0.29	0.281	0.343	**0.386**	0.268	**0.679**	**0.754**	**0.792**
E	9	0.088	0.256	0.26	0.089	0.073	0.129	**0.229**	0.236	0.23	0.235	0.228	0.245	0.198	0.277	**0.262**	0.194	0.175	0.21	0.19	**0.296**	0.22	**0.463**	**0.553**	**0.579**
F	4	0.18	0.2	0.2	0.26	0.26	0.35	0.15	0.15	0.15	0.15	0.15	0.15	**0.285**	0.305	0.345	0.4	0.33	0.185	0.235	**0.585**	**0.4**	**0.83**	**0.89**	**0.9**
G	11	0.033	**0.597**	0	0.10	0.153	0.10	0.147	0.133	0.15	0.157	0.15	0.137	0.11	0.206	0.097	0.09	0.119	0.129	**0.258**	0.561	**0.729**	**0.858**	**0.877**	**0.884**
H	10	0.093	0.139	0.139	0.22	0.141	0.247	0.3	0.314	0.312	0.307	0.307	0.306	0.26	0.222	0.148	0.235	0.24	0.286	**0.37**	**0.441**	**0.421**	**0.951**	**0.891**	**0.915**
I	7	0.148	0.093	0.093	0.109	0.161	0.351	0.335	0.317	0.322	0.367	0.337	0.361	0.131	0.109	0.085	0.093	0.093	0.111	**0.528**	**0.748**	**0.716**	**0.781**	**0.77**	**0.763**

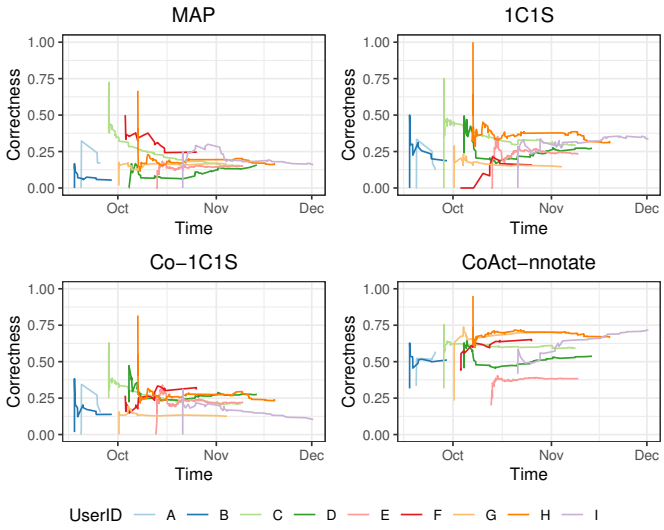

Fig. 6. The average progression of correctness over the time for sampled users.

classifier per sensor (i.e. 1C1S) over a simplistic multivariate setup (i.e. MAP). However, the repeated measure of ANOVA test found a statistically significant mean difference between the correctness of MAP and 1C1S, $F(1, 1078) = 152.2$, $p < .001$. The results of the two-sample t-test (assuming unequal variance) also found a statistically significant evidence of a difference of mean correctness between MAP and 1C1S, $t(df = 1004.7) = 12.34$, $p < .001$, 95% CI for the difference in means [0.06, 0.08].

It should be noted that both MAP and 1C1S do not use progressive learning. In this case, the models are constructed based on the first set of sensor feature bags for each label. Hence, the overall performance is insufficient. Even by including the co-training process for progressive learning (refer to Co-1C1S), the difference in correctness measurements is not substantial in comparison with non-progressive learning. This argument is evident from the results of two-sample t-test (assuming unequal variance) between progressive learning (i.e. Co-1C1S) and non-progressive learning (i.e. MAP and 1C1S), which found no statistically significant evidence of a difference for the mean correctness values, $t(df = 1145.3) = 0.715$, $p = 0.475$, 95% CI for the difference in means [-0.01, 0.01].

Ultimately, our proposed CoAct-nnotate pipeline can significantly improve annotation prediction (increasing average correctness by 35.94%) over all baselines. From the results of the two-sample t-test assuming unequal variance, there is statistically significant evidence of a difference of mean correctness between CoAct-nnotate and all baselines, $t(df = 588.2) = 33.302$, $p < .001$, 95% CI for the difference in means [0.37, 0.42]. In other words, co-training alone (refer to the result of Co-1C1S) is not enough to enhance the predictive performance over time in daily annotation tasks. It is evident that by combining both co-training and active learning (i.e. CoAct-nnotate), the outcome becomes progressively accurate (as shown in Figure 6).

As shown in Figure 6, the average correctness values are aggregated per user over time (for the iterative experiment on all base classifiers), spanning from late August to the end of November in 2016. In fact, this is aligned with the duration of the data collection campaign of the CrowdSignals dataset in which each user participated for four to six weeks of automatic logging of their smartphone sensor data in daily life.

From the visualisation of average correctness over the time, we can conclude that our proposed CoAct-notate clearly outperforms all the baseline approaches in annotation prediction in most the cases. For the co-training approach without active feedback from the users (Co-1C1S), the average performance degrades at an alarming pace in comparison with MAP and 1C1S. Unfortunately, the weakness of original co-training is known to result in degrading performance over time if the sampling bias shifts towards the unlabelled bags with mutual agreement and misclassification of class labels (i.e. incorrect annotation predictions). Therefore, this weakness is tackled in our proposed framework by integrating active learning (feedback from the users) to reduce the bias shifting towards the misclassification of class labels.

For over 50% of the time length of annotation prediction, our CoAct-nnotate visually demonstrates steady improvement of average correctness, which is also supported by the trend depicted in Figure 7. We plot the smooth line of the linear model (using a second degree polynomial term) on all correctness values of classifiers in the iterative experiment on all users within the normalised scale of time. Thus, we see a stable increase of the performance of CoAct-nnotate by the early convergence starting from 40% of the time duration of annotation prediction. Considering there are 13 annotations in

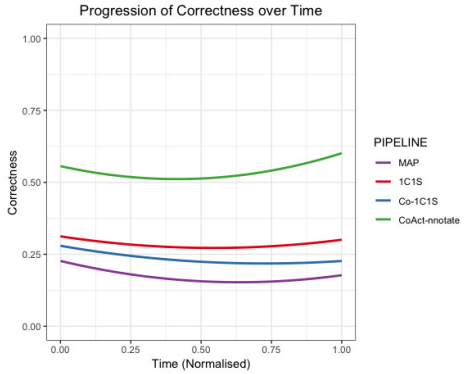

Fig. 7. The progression of correctness over the time for MAP, 1C1S, Co-1C1S and CoAct-nnotate.

total that can be predicted for users, the baseline accuracy can be set to 7.7% (1/13) for an application of annotation prediction. Therefore, it can be concluded that our proposed CoAct-nnotate pipeline can guess the correct user annotation 50% of the time, which is significantly above this baseline.

It should be noted that our experiment is limited to evaluation in which we assume the test bags to have active feedback from users. In a real scenario of ESM studies, users might ignore such survey notifications and not provide any feedback on the underlying predicted annotations. Further, the correctness measure presented in this paper is based on the notion of single-annotation prediction. If the ESM survey question is presented with multiple choices, then *top-k* predicted annotations can be displayed based on their ranked posterior probabilities. However, an option of 'other' should be displayed in an interactive annotation process of a real-world application to provide an alternative for the user that reveals more choices or inputs an answer via free text input. Our study aims to reduce such a choice overload issue during the ESM annotation process. An immediate challenge for future work is to measure the real-time performance of annotation prediction and evaluate it based on actual experience (in terms of user burden). Given such challenges, future research is required to improve the techniques used in ESM studies, leading to fewer interruptions and burdens for participants.

Ideally, the model training should be performed in a powerful instance (e.g. in the cloud) because mobile devices are restricted in terms of their computational resources. Therefore, the time taken to perform training on mobile devices is not evaluated in our current study. The summarisation technique that we applied in the experiment aims to derive a more compact representation of the given feature bags. Our previous results [35] show that the applied summarisation technique tends to maintain a relatively stable and reliable inter-rater agreement between machine learning models. Training the model on smart devices should be considered as another significant challenge that will lead to more intelligent mobile sensing applications (e.g. for assistive technologies). Nevertheless, the main contribution of this paper is to improve the model of annotation prediction over time by using both concepts of co-training and active learning.

V. Conclusion

This paper presents a framework to reduce user burden in ESM studies. Specifically, our work shows how semi-supervised learning can be used to predict the ESM labels that could be relevant to users at the time of questioning. We demonstrate the ability to predict the annotations before they are acquired from the users through an active feedback mechanism. Through the application of both co-training and active learning in our proposed multi-view models, the overall accuracy of annotation prediction systems is increased by 35.94% in comparison with conventional approaches. Therefore, researchers can customise the scheduling of ESM questionnaires to collect labels from all required instances. This can help overcome situations in which less frequent instances are not captured due to the limited sampling rate of ESM studies.

CoAct-nnotate is designed as a system for generic prediction of ESM labelling. Although the target application in this paper is for activity recognition, this approach can also be used for other types of applications, such as mood or emotional changes (assuming different sets of sensors are deployed, e.g. wearables for emotion prediction). Moreover, we envision that the future intelligent digital assistants (e.g. Amazon Alexa, Google Assistant and Microsoft Cortana) would be able to infer and support daily user activities and tasks [42], [43] through ubiquitous sensing. In this case, our proposed framework can be used to improve such virtual assistants to be more aware of the contexts of a mobile user and adapt accordingly based on active feedback.

Evaluation of actual user burden could be considered in future work of an intelligent ESM annotation process. In this study, we assume a scenario in which the user provides an annotation at a given time for an experiment performed on an existing dataset. Moreover, the selection of appropriate features and learning parameters can have direct effects on the accuracy of an annotation prediction. In our study, we chose the parameter values heuristically. Therefore, an efficient and intelligent selection of features and learning parameters (which also evolves over time based on user contexts) would require further study in this era of ubiquitous computing research.

VI. Acknowledgments

Jonathan Liono is supported by the Australian Government Research Training Program Scholarship from the RMIT University. This research was partially supported by Microsoft Research.

References

[1] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," in *Flow and the foundations of positive psychology*. Springer, 2014, pp. 35–54.

[2] M. Fuller-Tyszkiewicz, M. McCabe, H. Skouteris, B. Richardson, K. Nihill, B. Watson, and D. Solomon, "Does body satisfaction influence self-esteem in adolescents' daily lives? an experience sampling study," *Journal of adolescence*, vol. 45, pp. 11–19, 2015.

[3] S. Ghosh, V. Chauhan, N. Ganguly, B. Mitra, and P. De, "Impact of experience sampling methods on tap pattern based emotion recognition," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 713–722.

[4] M. Gustarini, M. P. Scipioni, M. Fanourakis, and K. Wac, "Differences in smartphone usage: Validating, evaluating, and predicting mobile user intimacy," *Pervasive and Mobile Computing*, vol. 33, pp. 50–72, 2016.

[5] N. van Berkel, C. Luo, T. Anagnostopoulos, D. Ferreira, J. Goncalves, S. Hosio, and V. Kostakos, "A systematic assessment of smartphone usage gaps," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: ACM, 2016, pp. 4711–4721. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858348

[6] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," *Pervasive computing*, pp. 1–17, 2004.

[7] C. Gurrin, A. F. Smeaton, A. R. Doherty *et al.*, "Lifelogging: Personal big data," *Foundations and Trends® in Information Retrieval*, vol. 8, no. 1, pp. 1–125, 2014.

[8] I. Li, A. Dey, and J. Forlizzi, "A stage-based model of personal informatics systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 557–566.

[9] R. Gouveia and E. Karapanos, "Footprint tracker: supporting diary studies with lifelogging," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2921–2930.

[10] U. Atz, "Evaluating experience sampling of stress in a single-subject research design," *Personal and ubiquitous computing*, vol. 17, no. 4, pp. 639–652, 2013.

[11] E. Welbourne and E. M. Tapia, "Crowdsignals: A call to crowdfund the community's largest mobile dataset," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 873–877.

[12] Y. Vaizman, N. Weibel, and G. Lanckriet, "Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, p. 168, 2018.

[13] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 554.

[14] N. v. Berkel, D. Ferreira, and V. Kostakos, "The experience sampling method on mobile devices," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 93, 2017.

[15] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, "Wearable sensors for reliable fall detection," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*. IEEE, 2006, pp. 3551–3554.

[16] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008.

[17] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A survey on activity detection and classification using wearable sensors," *IEEE Sensors Journal*, vol. 17, no. 2, pp. 386–403, 2017.

[18] J. Gershuny, "Costs and benefits of time sampling methodologies," *Social Indicators Research*, vol. 67, no. 1-2, pp. 247–252, 2004.

[19] R. Brewer, M. Morris, and S. Lindley, "How to remember what to remember: Exploring possibilities for digital reminder systems," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 38, 2017.

[20] D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone, "A survey method for characterizing daily life experience: The day reconstruction method," *Science*, vol. 306, no. 5702, pp. 1776–1780, 2004.

[21] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones," in *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 2007, pp. 57–70.

[22] A. B. Krueger and D. A. Schkade, "The reliability of subjective well-being measures," *Journal of public economics*, vol. 92, no. 8, pp. 1833–1845, 2008.

[23] J. Froehlich, T. Dillahunt, P. Klasnja, J. Mankoff, S. Consolvo, B. Harrison, and J. A. Landay, "Ubigreen: investigating a mobile tool for tracking and supporting green transportation habits," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 1043–1052.

[24] D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, and A. K. Dey, "Contextual experience sampling of mobile application micro-usage," in *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM, 2014, pp. 91–100.

[25] M. Obuchi, W. Sasaki, T. Okoshi, J. Nakazawa, and H. Tokuda, "Investigating interruptibility at activity breakpoints using smartphone activity recognition api," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1602–1607.

[26] N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, and A. T. Campbell, "Urban sensing systems: opportunistic or participatory?" in *Proceedings of the 9th workshop on Mobile computing systems and applications*. ACM, 2008, pp. 11–16.

[27] B. Minor and D. J. Cook, "Forecasting occurrences of activities," *Pervasive and mobile computing*, vol. 38, pp. 77–91, 2017.

[28] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1167–1174.

[29] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[30] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *Embedded and Real-Time Computing Systems and Applications, 2007. RTCSA 2007. 13th IEEE International Conference on*. IEEE, 2007, pp. 469–475.

[31] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multi-task, multi-kernel learning for estimating individual wellbeing," in *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, vol. 898, 2015.

[32] B. Tan, E. Zhong, E. W. Xiang, and Q. Yang, "Multi-transfer: Transfer learning with multiple views and multiple sources," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 7, no. 4, pp. 282–293, 2014.

[33] H. S. Hossain, M. A. A. H. Khan, and N. Roy, "Active learning enabled activity recognition," *Pervasive and Mobile Computing*, vol. 38, pp. 312–330, 2017.

[34] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Bayesian co-training," *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2649–2680, 2011.

[35] J. Liono, P. P. Jayaraman, A. Qin, T. Nguyen, and F. D. Salim, "Qdas: Quality driven data summarisation for effective storage management in internet of things," *Journal of Parallel and Distributed Computing*, 2018.

[36] A. Liu, J. Ghosh, and C. E. Martin, "Generative oversampling for mining imbalanced datasets." in *DMIN*, 2007, pp. 66–72.

[37] B. Das, N. C. Krishnan, and D. J. Cook, "Racog and wracog: Two probabilistic oversampling techniques," *IEEE transactions on knowledge and data engineering*, vol. 27, no. 1, pp. 222–234, 2015.

[38] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[39] W. Shao, F. D. Salim, A. Song, and A. Bouguettaya, "Clustering big spatiotemporal-interval data," *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 190–203, 2016.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] G. Andrew and J. Gao, "Scalable training of l1-regularized log-linear models," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 33–40.

[42] J. Liono, J. R. Trippas, D. Spina, M. S. Rahaman, Y. Ren, F. D. Salim, M. Sanderson, F. Scholer, and R. W. White, "Building a Benchmark for Task Progress in Digital Assistants," in *Proceedings of WSDM'19 Task Intelligence Workshop (TI@WSDM19)*, 2019.

[43] J. R. Trippas, D. Spina, F. Scholer, A. H. Awadallah, P. Bailey, P. N. Bennett, R. W. White, J. Liono, Y. Ren, F. D. Salim, M. S. Rahaman, and M. Sanderson, "Learning About Work Tasks to Inform Intelligent Assistant Design," in *Proceedings of the 2019 Conference on Human Information Interaction & Retrieval*, ser. CHIIR '19, 2019.