# Situation Inference and Context Recognition for Intelligent Mobile Sensing Applications

A thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy

Jonathan Liono
Bachelor of Information Technology and Systems (Hons) ⟨⟨ BITS Hons ⟩⟩
[ Monash University ]

School of Science
College of Science, Engineering and Health
RMIT University
Melbourne, Australia
September, 2019

# Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Jonathan Liono
School of Science
RMIT University
$5^{th}$ September, 2019

# Acknowledgement

I would like to spend this special section to express my utmost gratitude for all precious people that have given me invaluable support in my life, especially towards the completion of this thesis. My research would have been impossible without the constant aid, undying support and complete trust of my academic supervisors, A/Prof. Flora Dilys Salim and A/Prof. Kai Qin, during the last five years. I appreciate their enthusiasm for my research and their assistances in writing this thesis. Without their constructive criticisms and innovative inspirations, I would not be able to grow and push myself beyond my comfort zone to be a better scientist. Specifically, I thank Flora for always being available and patient from start to the end of my PhD candidature, while providing invaluable inputs in research techniques and constant encouragement of never-giving-up during my research. At one particular moment, I was about to quit the candidature in the middle of my PhD journey due to my mental breakdown from an unfortunate event that happened to one of my family members. Flora was the only person who believed in me, which made me stronger and firm enough to not give up my research halfway. This chain of events made me truly learn and understand more about resilience and perseverance.

I take this opportunity to show my special appreciation to the people that have been providing me with a more fulfilling progression towards the end of my PhD journey at RMIT University. First mentions would be for my colleagues who are the initial members of the Context Recognition and Urban Intelligence (CRUISE) research group. I would like to share my greatest thanks to both Dr. Mohammed Saiedur Rahaman and Dr. Irvan Bastian Arief Ang as my loyal companions throughout my PhD candidature. We truly spent our pain and joy together in various crucial moments of all our PhD candidatures, keeping our perseverance and staying up night by night over deadlines. Another mention is dedicated to Dr. Wei Shao, for providing everyday entertainment that has made my PhD life more colourful, with his mind-wandering topics which can be eye-opening. Other initial CRUISE group members are Dr. Amin Sadri, Dr. Hui Song and Rumi Shakila Khan, and with them, I shared constant encouragement and built precious friendships during the course of my doctorate. I am profoundly grateful to Johanne R. Trippas for supporting and helping me in the last two years of my PhD research.

Special gratitude I would also like to convey to my previous supervisors Prof. Shonali Krishnaswamy, Dr. Nicholas Nicoloudis and Prof. Mohamed Medhat Gaber for their support in the completion of my Honours degree of Bachelor of Information Technology and Systems. Their legacy and pioneering innovation on mobile data mining have certainly helped to inspire and shape my skills in research, prior to pursuing this PhD degree.

I would like to express my special thanks and dedication to Stephanie Christianty, for providing constant mental support in my life and being my precious companion, who is also at this moment progressing towards the completion of her PhD candidature.

Moreover, my extended gratitude is given towards all my friends that directly or indirectly supported me. Thanks to Jingyi Xu for sharing all the memes, which also have given me precious mental support during my PhD journey. A special dedication is given to my friends: Sharif Yahya, Daniel Kim, Suryanata Wongtomo, Christopher Rianto, Victor Augusteo and the rest of "EX-K Crew" members, who have encouraged me with this question: "when are you going to finish your PhD?". With full respect, I also appreciate Jimmy Liong's daily visions and inspirations. Heartfelt thanks go to RMIT alumni and friends: Irwan Fario Subastian, Jason Liyanage, Danial Tham, Samuel Elsmore and Barti Murugesan, who had provided invaluable support, even before I started my PhD candidature.

I would also like to thank my scholarship providers, RMIT University and the Government of Australia for providing financial support in the form of an Australian Government Research Training Program (RTP) scholarship. A special acknowledgement is also mentioned for Microsoft Research, who has given the top-up scholarship, opportunities and support for parts of my research.

Last but not the least, I would like to thank my family: parents, brothers and sisters, for understanding and supporting me spiritually throughout my PhD and life journeys.

# Statistical Summary

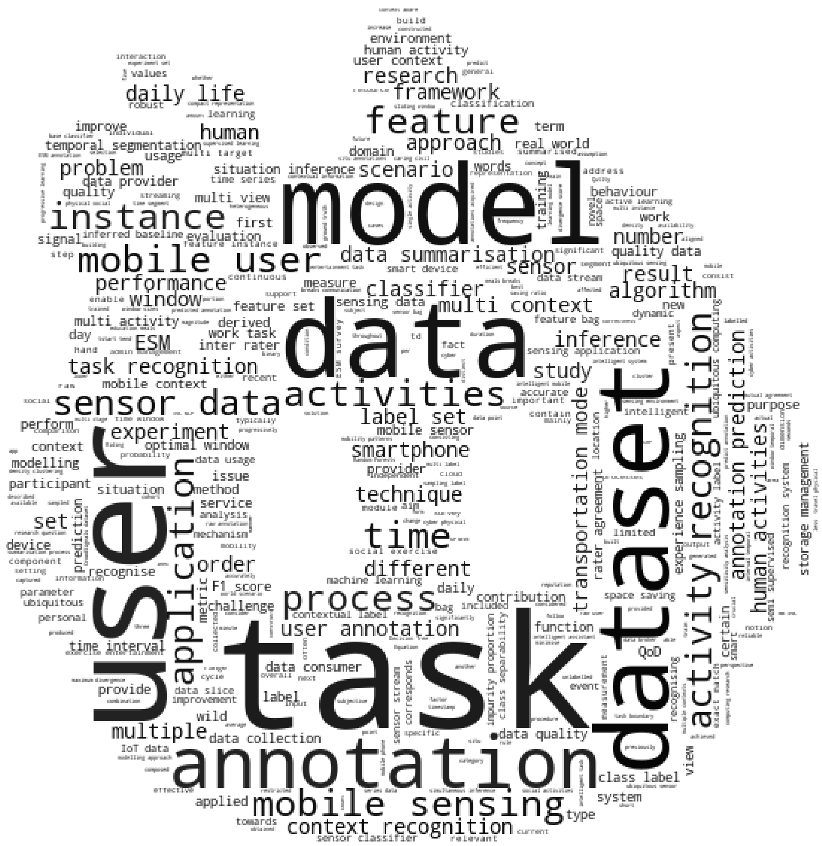
Total number of pages: 197 pages

Total number of word count: 59,395 words

Total number of figures : 62 figures

Total number of tables: 13 tables

Total number of references: 225 scientific literatures



# Credits

Portions of the material in this thesis have previously appeared in the following publications:

- Jonathan Liono, A. K. Qin, and Flora D. Salim. "Optimal time window for temporal segmentation of sensor streams in multi-activity recognition". *In the Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services(MobiQuitous 2016)*, pages 10-19. Hiroshima, Japan. November 28 - December 01, 2016. DOI:10.1145/2994374.2994388 [Liono et al., 2016] **(CORE Rank A conference)**
- Jonathan Liono, Zahraa S. Abdallah, A. K. Qin, and Flora D. Salim. "Inferring Transportation Mode and Human Activity from Mobile Sensing in Daily Life". *In the Proceedings of the 15th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2018)*, pages 342-351. New York, NY, USA. November 05 - 07, 2018. DOI:10.1145/3286978.3287006 [Liono et al., 2018a] **(CORE Rank A conference)**
- Jonathan Liono, Flora D. Salim, Niels van Berkel, Vassilis Kostakos, and A. K. Qin "Improving Experience Sampling with Multi-view User-driven Annotation Prediction". *In 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Kyoto, Japan. March 10 - 15, 2019.
DOI:10.1109/PERCOM.2019.8767394 [Liono et al., 2019a] **(CORE Rank A* conference)**
- Jonathan Liono, Johanne R. Trippas, Damiano Spina, Mohammad S. Rahaman, Yongli Ren, Flora D. Salim, Mark Sanderson, Falk Scholer and Ryen W. White. "Building a Benchmark for Task Progress in Digital Assistants". *In the Proceedings of WSDM'19 Task Intelligence Workshop (TI@WSDM19)*, Melbourne. 2019. [Liono et al., 2019b] **(Workshop at a CORE Rank A* conference)**

- Jonathan Liono, Prem Prakash Jayaraman, A. K. Qin, Thuong Nguyen, and Flora D. Salim. "QDaS: Quality driven data summarisation for effective storage management in Internet of Things". *Journal of Parallel and Distributed Computing*. 2018. DOI:10.1016/j.jpdc.2018.03.013 [Liono et al., 2018b] **(Impact Factor:1.815, SJR: Q1)**
- Jonathan Liono, Mohammad Saiedur Rahaman, Flora D. Salim, Yongli Ren, Damiano Spina, Falk Scholer, Johanne Trippas, Mark Sanderson, Paul N. Bennett and Ryen White. "Intelligent Task Recognition: Towards Enabling Productivity Assistance in Daily Life". *In 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. **Under review**. **(CORE Rank A* conference)**

My research is primarily supported by the Australian Government Research Training Program Scholarship from RMIT University. In this thesis, the research in Chapters 4 and 5 is partially supported by Microsoft Research through Cortana Intelligence Institute (a co-funded initiative between Microsoft Research, Cortana Research and RMIT University).

# Contents

# List of Figures

# List of Tables

# Abstract

The usage of smart devices is an integral element in our daily life. With the richness of data streaming from sensors embedded in these smart devices, the applications of ubiquitous computing are limitless for future intelligent systems. Situation inference is a non-trivial issue in the domain of ubiquitous computing research due to the challenges of mobile sensing in unrestricted environments. There are various advantages to having robust and intelligent situation inference from data streamed by mobile sensors. For instance, we would be able to gain a deeper understanding of human behaviours in certain situations via a mobile sensing paradigm. It can then be used to recommend resources or actions for enhanced cognitive augmentation, such as improved productivity and better human decision making.

Sensor data can be streamed continuously from heterogeneous sources with different frequencies in a pervasive sensing environment (e.g., smart home). It is difficult and time-consuming to build a model that is capable of recognising multiple activities. These activities can be performed simultaneously with different granularities. We investigate the separability aspect of multiple activities in time-series data and develop OPTWIN as a technique to determine the optimal time window size to be used in a segmentation process. As a result, this novel technique reduces need for sensitivity analysis, which is an inherently time consuming task. To achieve an effective outcome, OPTWIN leverages multi-objective optimisation by minimising the impurity (the number of overlapped windows of human activity labels on one label space over time series data) while maximising class separability.

The next issue is to effectively model and recognise multiple activities based on the user's contexts. Hence, an intelligent system should address the problem of multi-activity and context recognition prior to the situation inference process in mobile sensing applications. The performance of simultaneous recognition of human activities and contexts can be easily affected by the choices of modelling approaches to build an intelligent model. We investigate the associations of these activities and contexts at multiple levels of mobile sensing perspectives to reveal the dependency property in multi-context recog-

nition problem. We design a Mobile Context Recognition System, which incorporates a Context-based Activity Recognition (CBAR) modelling approach to produce effective outcome from both multi-stage and multi-target inference processes to recognise human activities and their contexts simultaneously. Upon our empirical evaluation on real-world datasets, the CBAR modelling approach has significantly improved the overall accuracy of simultaneous inference on transportation mode and human activity of mobile users.

The accuracy of activity and context recognition can also be influenced progressively by how reliable user annotations are. Essentially, reliable user annotation is required for activity and context recognition. These annotations are usually acquired during data capture in the world. We research the needs of reducing user burden effectively during mobile sensor data collection, through experience sampling of these annotations in-the-wild. To this end, we design CoAct-nnotate — a technique that aims to improve the sampling of human activities and contexts by providing accurate annotation prediction and facilitates interactive user feedback acquisition for ubiquitous sensing. CoAct-nnotate incorporates a novel multi-view multi-instance learning mechanism to perform more accurate annotation prediction. It also includes a progressive learning process (i.e., model retraining based on co-training and active learning) to improve its predictive performance over time.

Moving beyond context recognition of mobile users, human activities can be related to essential tasks that the users perform in daily life. Conversely, the boundaries between the types of tasks are inherently difficult to establish, as they can be defined differently from the individuals' perspectives. Consequently, we investigate the implication of contextual signals for user tasks in mobile sensing applications. To define the boundary of tasks and hence recognise them, we incorporate such situation inference process (i.e., task recognition) into the proposed Intelligent Task Recognition (ITR) framework to learn users' Cyber-Physical-Social activities from their mobile sensing data. By recognising the engaged tasks accurately at a given time via mobile sensing, an intelligent system can then offer proactive supports to its user to progress and complete their tasks.

Finally, for robust and effective learning of mobile sensing data from heterogeneous sources (e.g., Internet-of-Things in a mobile crowdsensing scenario), we investigate the utility of sensor data in provisioning their storage and design QDaS — an application agnostic framework for quality-driven data summarisation. This allows an effective data summarisation by performing density-based clustering on multivariate time series data from a selected source (i.e., data provider). Thus, the source selection process is determined by the measure of data quality. Nevertheless, this framework allows intelligent systems to retain comparable predictive results by its effective learning on the compact representations of mobile sensing data, while having a higher space saving ratio.

This thesis contains novel contributions in terms of the techniques that can be employed for mobile situation inference and context recognition, especially in the domain of ubiquitous computing and intelligent assistive technologies. This research implements and extends the capabilities of machine learning techniques to solve real-world problems on multi-context recognition, mobile data summarisation and situation inference from mobile sensing. We firmly believe that the contributions in this research will help the future study to move forward in building more intelligent systems and applications.

# Chapter 1

# Introduction

Situation awareness (SA) has become a growing research aim for many interdisciplinary studies in the past decades to support human decision making and understanding of their environments [Feng et al., 2009, Foresti et al., 2015, Panteli and Kirschen, 2015, Preden et al., 2015, Yin et al., 2015, Adams et al., 2017, De Maio et al., 2017, Stanton et al., 2017]. Although off-the-shelf smart devices are embedded with cheaper and greater computational capacity, inferring the situation of mobile users is still considered to be a tremendous challenge in ubiquitous and pervasive computing research. It is believed that introducing situation awareness to ubiquitous computing via the mobile sensing paradigm is crucial, in order to provide more meaningful assistance for users in an intelligent manner and proactively, if possible. The benefits of such proactive intelligence could result in the reduced cognitive workload of an individual (i.e., a user of smart devices), in a particular situation at that point in time and space.

In the early work on situational awareness, most studies focused on command and control research for tactical decision-making purposes, such as the context-aware intelligent assistant approach (weather assistant) to improve the situation awareness of an aircraft pilot [Spirkovska and Lodha, 2004]. Given the prevalent usage of ubiquitous devices in daily human life, there are limitless opportunities for future intelligent systems to offer proactive situation-aware assistance to mobile users. In-the-wild mobile sensing data can be utilised to recognise human activities, contexts and even their behaviours to provide more significant insights for situation inference for its users. Understanding human behaviours in certain situations will then allow intelligent systems to recommend resources or actions for enhanced augmentation (e.g., improved productivity and better decision making) through mobile sensing within the smart environments.

In the realm of pervasive computing [Ye et al., 2012], a situation is defined as "an abstraction of the events occurring in the real world derived from context and hypotheses about how observed context relates to factors of interest to designers and applications". Domain knowledge and the expected behaviour of the observed phenomena and its contexts (derived from spatial and temporal dimensions), are required to construct the notion of a particular situation. In this case, the underlying situation inference in context-aware computing relies heavily on the availability of sensor data and effective behaviour modelling of mobile users in smart sensing environments. Hence, user-driven and data-driven modelling strategies are crucial in helping the intelligent systems make sense of the relevant contexts from the mobile sensor data. It should be noted that these data can be streamed from various sources (e.g., multiple devices, physical, virtual and logical sensors) in an irregular manner. Nevertheless, reliable multi-context recognition and situation inference should provide a greater avenue for a user to have informed decision-making anytime and anywhere, through intelligent mobile sensing in pervasive environments.

The general overview of our vision is presented in Figure 1.1. The processes of context recognition and human task recognition play essential roles for intelligent systems and/or applications to offer proactive assistance (e.g., intelligent notifications). All of these processes are part of the mobile situation inference umbrella framework. In order to enable an affective situation inference, the collective composition of relevant contexts needs to be defined and refined. This can be achieved through systematic steps, starting from the processing of multivariate time series data of ubiquitous sensors, recognising co-occurring contexts to modelling, and recognising the tasks of mobile users. Nevertheless, the output of context recognition, task recognition, or proactive assistance, can be used to enrich the insights in sensing applications for a mobile user. For future intelligent systems, they should be not only effective but also affective [Picard, 2003]. We believe that the long term vision of affective computing is to intelligently understand more about its mobile user's situation and consequently provides responses that are more affective.

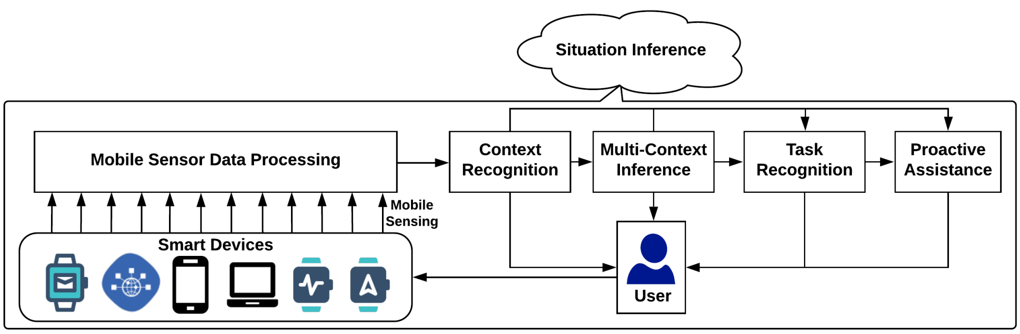



Figure 1.1: Context recognition and situation inference for intelligent mobile sensing applications.

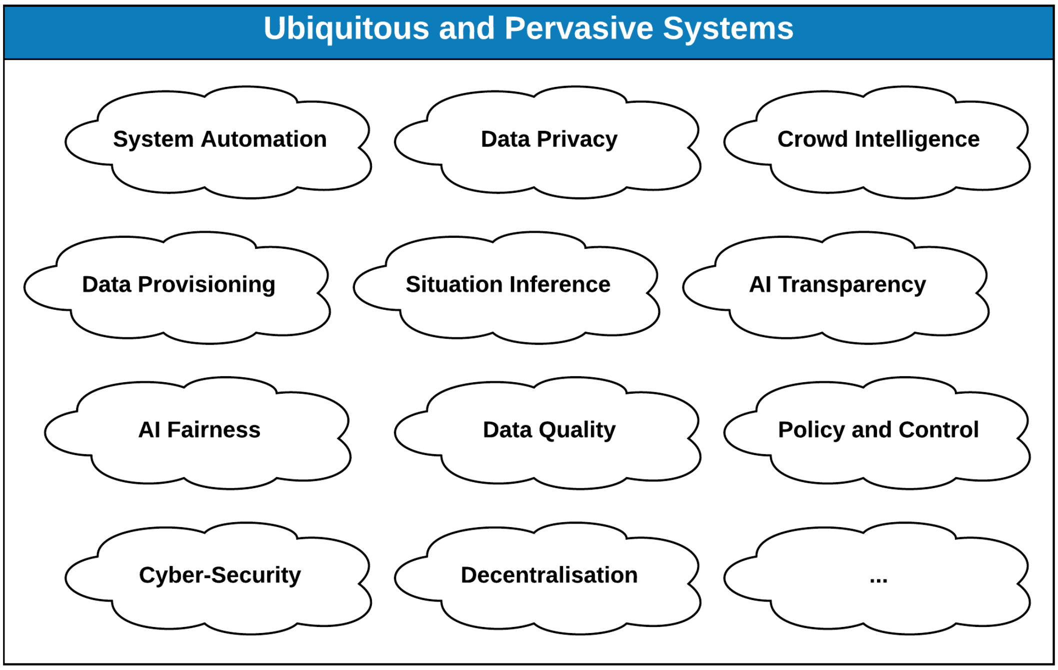



Figure 1.2: Emerging trends and situation inference in ubiquitous and pervasive systems.

Figure 1.2 shows the bigger picture related to ubiquitous and pervasive systems, beyond just situation inference for intelligent mobile sensing applications. Many of these emerging topics should be addressed to support effective and affective situation awareness in the next decades. Within the scope of our novel research and contributions, we focus on specific domains in situation inference (including context recognition), data quality and provisioning.

In this thesis, we focus on the research gaps of context recognition and situation inference from the perspective of mobile sensing. Therefore, the fundamental challenges of mobile situation inference are addressed. These consist of 1) temporal segmentation and 2) recognition of human activities and contexts over multi-dimensional spaces, 3) mobile-based task recognition, and 4) scaling up mobile sensing applications for big data and efficient machine learning processes.

## 1.1 Background and Motivations

In the domain of ubiquitous computing research, human activities and their relevant contexts can be characterised by the intrinsic patterns derived from mobile sensor data (sourced by various smart devices). According to Cook and Das [2007] (as identified in [Ye et al., 2012]), ubiquitous sensing focuses on capturing data based on:

- **Environment aspect**, where sensing data are associated with the environment surrounding the device, such as temperature, humidity, barometric pressure, light, electricity and gas usages.
- **Device aspect**, in which sensing data can be influenced by the state of a device. The information that is associated with state transitions and the availability of its sensors sets immediate challenges for ubiquitous computing to provide intelligent assistance for the users.
- **User aspect**, where sensing data can be used to characterise the user and relevant contexts, such as motion, human activities and mobility.
- **Interaction aspect**, where sensing data can be affected by direct or indirect interaction with its user.

The concept of automaticity should be embraced in the applications of intelligent mobile sensing to offer proactive assistant services. This can be achieved by raising the situation awareness of a mobile user through the processes of situation inference and multi-context recognition. It is also aligned with the inherent notion of context-aware computing for mobile sensing and its future directions [Yürür et al., 2016], where the user state is closely associated with a situation composed of the relationships of all high-level contexts (inference of device, user, physical and temporal contexts) and low-level contexts (sensed from physical, virtual and logical sensors).

Within the scope of this dissertation, we focus on the steps that lead to situation awareness from the individual/first-person perspective based on multivariate mobile sensor data. Many real-world problems can be tackled through the applications of ubiquitous sensing by recognising human activities alone. For instance, individuals may be monitored via real-time applications of activity recognition in the smart home environment [Chen et al., 2012b] and for elderly care [Torres et al., 2013]. Moreover, the derived patterns can be used for knowledge discovery on a user, such as early diagnosis of psychiatric diseases [Tacconi et al., 2008]. Geib et al. [2015] highlighted the needs for machines to be able to recognise, understand and predict human actions in order to allow context-aware human-computer interaction (HCI) in ever more complex situations. In the last few decades, the study of situation inference and multi-context recognition has not been properly explored for intelligent mobile sensing applications, despite the aforementioned research efforts by Geib et al. [2015] on activity, plan and intent recognition in ubiquitous computing. Our research encompasses the real-world problems where sensor data (including their relevant contexts) can be streamed continuously in an irregular manner over time. Consequently, the evolvability [Chang et al., 2009] of user contexts should be considered for a ubiquitous system to adapt to dynamic and everchanging user requirements (e.g., behaviour change). Let us consider a future scenario where intelligent applications are necessary to inform the users for better decision-making and proactively assist them in achieving the goals of their activities in daily life. The challenges and impacts of our research are simply breathtaking, with the overall motivation of unleashing the full potential of mobile sensing applications to provide situation-aware mobile assistance in-the-wild.

## 1.2 Research Challenges

In this thesis, we aim to maximise the capability of smart devices to support their users in daily life through intelligent applications that are enabled through mobile sensing. With the prevalent use of these ubiquitous devices (e.g., smartphones) in daily life, it is becoming increasingly important that they should be more aware of what, where and when the users would like to be assisted in certain situations. The time series characteristics of sensor data can be easily influenced by the noise within the mobile user's vicinity (e.g., environmental factors). This remains a continuous challenge for ubiquitous computing to tackle, due to dynamic changes in human activities, and the contexts and properties associated with user's sensing environments. As a consequence, there is an inherent risk often associated with the performance degradation of a predictive model deployed in an assistive system, especially with in-the-wild sensing scenarios.

In summary, the core challenges in this research include, but are not limited to, the following:

- Fusion of activities and contextual information relevant to the mobile user, to infer a certain situation.
- Providing intelligent recognition of human activities and contexts in multiple dimensions, given the magnitude of sensor data streaming from multiple channels.
- Effective model building and learning from sparse sensing data and the user's subjective annotations.
- Optimisating and minimising learning time of annotated multivariate mobile sensor data.
- Scaling mobile sensing applications for Big Data and efficient machine learning processes.

## 1.3 Research Questions

In order to progress the research in situation inference (especially for intelligent mobile sensing), the following research questions are defined and addressed:

- ***RQ-1.*** *How to find an optimal window size in processing multivariate sensor data for multi-context recognition?*
  In this research, the aim of situation inference is to understand the relationships of all contextual elements of a user that can be derived from mobile sensing. Foremost, such intelligent inference requires systematic cleaning, processing and feature extraction on time series data (logged from the streaming of mobile sensors).

  In order to accelerate the learning process of context recognition, we first address the windowing problem in processing multivariate time series data, given the multi-dimensional label spaces of contexts associated with mobile users. Finding optimal window size for time-interval based temporal segmentation is a non-trivial task, and it can be time-consuming to build an effective model for multi-context recognition. For this research question, we use the scenario of multi-activity recognition of users (with on-body sensors) in smart home environments.

- ***RQ-2.*** *How to perform multi-context recognition from multidimensional sensor data and user annotations?*
  Moving beyond the temporal segmentation problem in **RQ-1**, we discovered that independent classification of contextual labels (derived from multi-dimensional label space) typically results in inaccurate multi-context inference. In this research question, we address the modelling approach for simultaneous recognition of multiple contexts including human activities, which can be derived from human annotations and mobile sensing in uncontrolled environments. Inherently, the accuracy of multi-context recognition is easily influenced by how reliable human annotations are acquired from in-the-wild sensing. Consequently, we also address the problem of annotation prediction that aims to improve the interactivity of experience sampling of human annotations and reduce the burden during mobile sensor data collection. For example, the user can be presented with the most probable selection of answers relevant to the user's contexts, to minimise "choice overload" issues in a survey form.

- ***RQ-3.*** *How to recognise tasks of a mobile user from continuous contextual signals?*
  Once the collection of sensor data and annotations can be achieved effectively from mobile users in-the-wild (through experience sampling), the contextual data are then utilised to recognise their

daily tasks. In this case, we use the scenario where intelligent assistant should be proactive in accurately recognising which type of task a user is currently engaged in.

- ***RQ-4.*** *How to evaluate the quality of crowdsourced data from mobile sensing environments?* Given the proliferation of data collection on a larger scale, we address the problem of deriving a compact representation of the mobile sensor data of individuals based on the notion of data quality. We expand the application scenario where cloud resources are limited in terms of the storage of raw sensor data (streamed from the Internet of Things). By having a compact representation of data from mobile sensing empowers the learning process to be more efficient and effective in producing intelligent predictive models for multi-context recognition and situation inference.

## 1.4 Research Contributions

To address the aforementioned research questions, the contributions of this thesis are as follows:

1. **Optimal time-windowing technique for multi-context recognition in smart sensing environments**
   To facililate the model building for multi-context recognition, we investigated the effect of window size on the accuracy of multi-activity recognition in smart home environments. As a result, we developed OPTWIN as a technique that can determine optimal window size to be used in the segmentation process of continuous time series data from ubiquitous sensors. As human activities can overlap during a segmentation and recognition process, the measure of impurity (the number of overlapped windows of human activity labels on one sequence dimension) is minimised while maximising their class separability, to improve the overall performance of human activity recognition. Consequently, this technique is also expanded to adapt with a multi-activity scenario, where a user can carry out multiple activities at once (i.e., different high-level and low-level human activities over multi-dimensional label spaces).

2. **Simultaneous multi-context recognition for intelligent mobile sensing**
   In relation to the first contribution, we further investigate the association of multiple contexts and human activities from multi-dimensional label space. In fact, independent recognition of contexts and human activities from respective label sets produces lower accuracy in comparison with the assumption of label dependency for simultaneous recognition. This research is validated on the crowdsourced mobile sensing data in urban cities, based on user annotations of daily home activities and outdoor-commuting contexts. In this case, a problem of simultaneous transportation

mode inference and human activity recognition is addressed. Consequently, a novel modelling approach coined as **C**ontext-**b**ased **A**ctivity **R**ecognition (**CBAR**) is proposed to be used in a robust mobile context recognition system by combining the concepts of multi-stage and multi-target inferences. Substantially, this modelling approach outperforms the models that are built on the traditional approach, where there is no assumption of dependency between multiple contextual label sets for the given human activities. Nonetheless, such accurate multi-context recognition can be beneficial for mobile situation inference (e.g., identifying crowdedness in public transportation situations from intelligent mobile sensing).

3. **Multi-view user annotation prediction for interactive mobile sensing**
   The accuracy of multi-context recognition can be influenced by the reliability of user annotations. The acquisition of user annotations relative to recent activities, including their contexts, is typically enabled via the Experience Sampling Method (ESM) in psychological research and affective computing. Mobile sensing provides an avenue to lead the study further for building intelligent systems. To improve the process of the ESM-based survey in mobile sensing settings, we present CoAct-nnotate, a novel semi-supervised learning technique tailored for annotation prediction based on the time series data that are sourced from multiple sensors. This pipeline leverages both multi-view and multi-instance learning techniques based on feature instances stored inside sensor bags. Moreover, CoAct-nnotate can also be used for improving the model progressively over time from active learning based on user feedback. Nevertheless, the ultimate aim of accurate annotation prediction in this contribution is to minimise the user burden by reducing the number of choices in ESM-based surveys in mobile sensor data collection.

4. **Intelligent recognition of human tasks in daily life from mobile sensing**
   Once the annotations for human tasks have been acquired from mobile sensing in-the-wild, recognising them in daily life is challenging due to the subjectivity of user annotations and also the unclear boundaries of task annotations (acquired in-situ). In this case, we present the problem of intelligent task recognition in a daily mobile sensing scenario. The main contribution of our research is related to our novel **I**ntelligent **T**ask **R**ecognition (**ITR**) framework for daily mobile sensing. The ITR framework is comprised of 1) a presence-based task boundary construction mechanism (on in-situ annotations), and 2) a learning module based on **C**yber-**P**hysical-**S**ocial (**CPS**) contextual data and modelling. From the results of our empirical evaluation, it can be concluded that an intelligent system needs to fully utilise all cyber, physical and social activities in order to provide a greater predictive result and insights for situation inference.

5. **Quality-driven data summarisation for effective and scalable mobile sensing applications** To expand the mobile sensing data collection and experiments on a larger scale (e.g., crowdsensing), an effective technique is needed to retain a compact representation of informative elements in sensor data. In this case, we introduce the scenario of cloud instances having limited capacity to store raw sensing data from Internet-of-Things (IoT). We develop QDaS, a novel domain agnostic framework for effective storage and management of IoT data in the cloud. Based on the proposed data quality estimation technique (using the notion of the utility value of sensor data), QDaS leverages its unsupervised module to perform density-based data summarisation on continuous multivariate time series sensor data. Our experimental results show the effectiveness of this data summarisation technique, by having higher space saving ratio while maintaining reliable inter-rater agreement between machine learning models. In fact, this smart data summarisation technique has been proven to be effective in a semi-supervised learning module of CoAct-nnotate framework for multi-view user annotation prediction (refer to Chapter 4).

## 1.5 Thesis Organisation

This chapter mainly discusses the challenges and motivation behind situation inference and context recognition from mobile sensing. The rest of this dissertation is structured as follows. In the next chapter, we provide the necessary background on different aspects of mobile sensing and intelligent applications of ubiquitous computing in daily life. The main contributions of this dissertation are included in Chapters 2 to 6 as shown in Figure 1.3. In Chapter 2, a technique for discovering the optimal time window size in temporal segmentation process is presented, which scales up to multi-context recognition in mobile sensing. In Chapter 3, we address the multi-context recognition problem for multivariate time series data of mobile sensors, by considering the factor of context dependency of human activities. In Chapter 4, we present a novel semi-supervised technique for predicting human annotations to improve the experience of data collection for intelligent mobile sensing applications. In Chapter 5, we address the situation inference problem by recognising human tasks in a daily mobile sensing scenario, utilising the CPS activities. To scale up situation inference and context recognition efficiently from individual mobile users (e.g., in a mobile crowdsensing scenario), we present an application agnostic framework for quality-driven data summarisation (Chapter 6) to derive the compact representation of mobile sensing data. Finally, Chapter 7 concludes this thesis with a summary of our contributions, key findings, and discussion of future works.

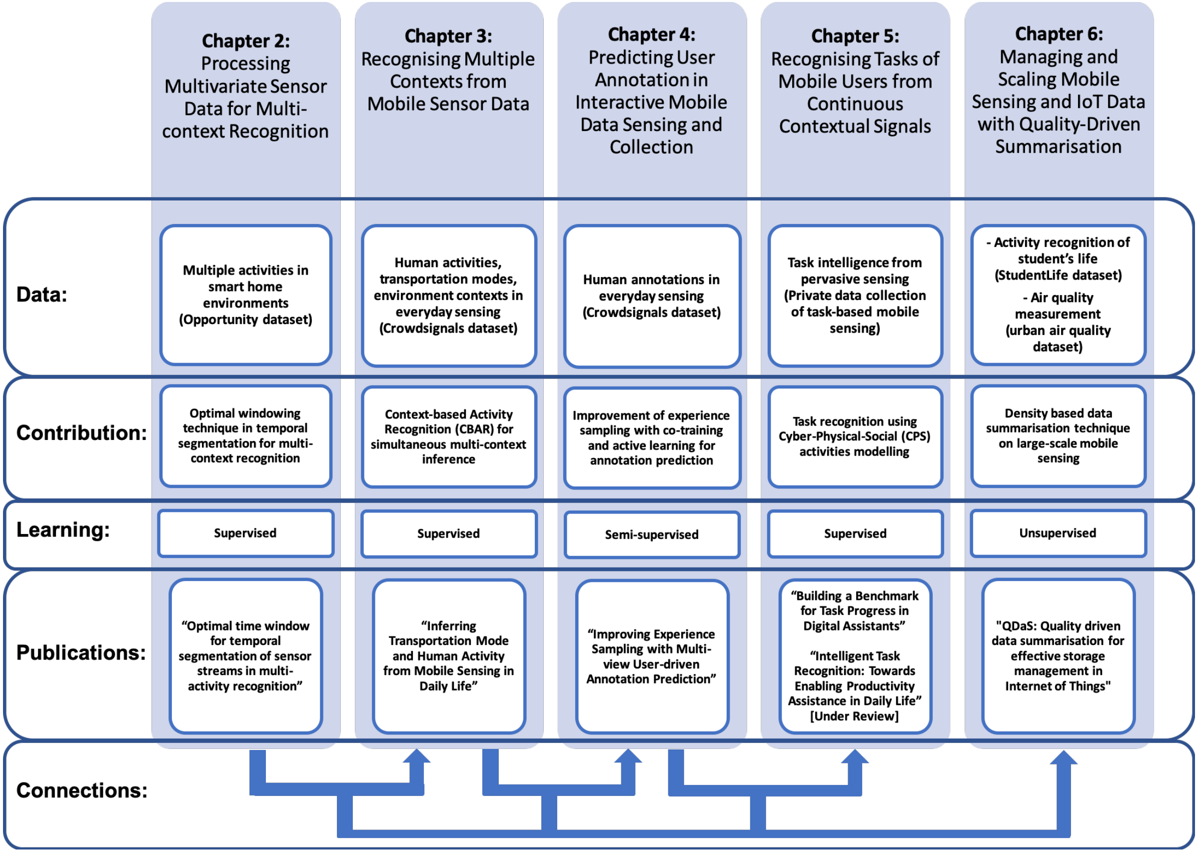



Figure 1.3: Overview of thesis structure and organisation.

**Notes for connections:**

- **Chapter 2 to Chapter 3:** The problem of multi-activity recognition in Chapter 2 is expanded to simultaneous multi-context (activities and related contexts) recognition.
- **Chapter 3 to Chapter 4:** Multiple contextual labels and human activities in Chapter 3 are the results of the decomposition process of raw human annotations. We expanded the problem of acquiring these raw annotations (Chapter 4) to improve the interactivity of experience sampling.
- **Chapter 4 to Chapter 5:** Beyond context recognition and prediction, Chapter 5 addresses the issue of situation inference in the form of human task recognition for intelligent assistants through mobile sensing.
- **Chapters 2, 3, 4 and 5 to Chapter 6:** The topic of this research is constituted within the realm of mobile sensing from first-person based situation inference and context recognition. Chapter 6 extends all of the preceding chapters by scaling up mobile sensing for large-scale experiments.

Based on the thesis organisation, the following breakdown is presented with respect to their paper publications in conferences and journals (including exclusive acknowledgement on several articles which were partially supported by Microsoft Research):

- **Chapter 2: Processing multivariate sensor data for multi-context recognition**
  Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from our paper published as:
  - Jonathan Liono, A. K. Qin, and Flora D. Salim. "Optimal time window for temporal segmentation of sensor streams in multi-activity recognition". *In the Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services(MobiQuitous 2016)*, pages 10-19. Hiroshima, Japan. November 28 - December 01, 2016. DOI:10.1145/2994374.2994388 [Liono et al., 2016] **(CORE Rank A conference)**
- **Chapter 3: Recognising multiple contexts from mobile sensor data**
  Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from our paper published as:
  - Jonathan Liono, Zahraa S. Abdallah, A. K. Qin, and Flora D. Salim. "Inferring Transportation Mode and Human Activity from Mobile Sensing in Daily Life". *In the Proceedings of the 15th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2018)*, pages 342-351. New York, NY, USA. November 05 - 07, 2018. DOI:10.1145/3286978.3287006 [Liono et al., 2018a] **(CORE Rank A conference)**
- **Chapter 4: Predicting user annotation in interactive mobile data sensing and collection**
  Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from our paper published as:
  - Jonathan Liono, Flora D. Salim, Niels van Berkel, Vassilis Kostakos, and A. K. Qin "Improving Experience Sampling with Multi-view User-driven Annotation Prediction". *In 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Kyoto, Japan. March 10 - 15, 2019. DOI:10.1109/PERCOM.2019.8767394 [Liono et al., 2019a] **(CORE Rank A* conference)**
    **[The research in this paper is partially supported by Microsoft Research]**

- **Chapter 5: Recognising Tasks of Mobile Users from Continuous Contextual Signals**
  Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from our paper published as:

  – Jonathan Liono, Johanne R. Trippas, Damiano Spina, Mohammad S. Rahaman, Yongli Ren, Flora D. Salim, Mark Sanderson, Falk Scholer and Ryen W. White. "Building a Benchmark for Task Progress in Digital Assistants". *In the Proceedings of WSDM'19 Task Intelligence Workshop (TI@WSDM19)*, Melbourne. 2019. [Liono et al., 2019b] **(Workshop at a CORE Rank A* conference)**
    **[The research in this paper is partially supported by Microsoft Research]**

  – Jonathan Liono, Mohammad Saiedur Rahaman, Flora D. Salim, Yongli Ren, Damiano Spina, Falk Scholer, Johanne Trippas, Mark Sanderson, Paul N. Bennett and Ryen White. "Intelligent Task Recognition: Towards Enabling Productivity Assistance in Daily Life". *In 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)* **(Under review)**
    **[The research in this paper is partially supported by Microsoft Research]**

- **Chapter 6: Managing and Scaling Mobile Sensing and IoT Data with Quality-Driven Summarisation**
  Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from our paper published as:

  – Jonathan Liono, Prem Prakash Jayaraman, A. K. Qin, Thuong Nguyen, and Flora D. Salim. "QDaS: Quality driven data summarisation for effective storage management in Internet of Things". *Journal of Parallel and Distributed Computing*. 2018.
    DOI:10.1016/j.jpdc.2018.03.013 [Liono et al., 2018b] **(Impact Factor:1.815, SJR: Q1)**

# Chapter 2

# Processing Multivariate Sensor Data for Multi-context Recognition

## 2.1 Introduction

As discussed in Chapter 1, intelligent sensing can be enabled by accurate recognition of user contexts, including multiple human activities in a smart environment. This chapter discusses our research contributions in optimising the overall temporal segmentation process of multiple activity recognition, by finding the optimal window size for feature extraction purposes.

The advances in ubiquitous computing in recent years have driven the research to automatically observe, monitor and recognise contextual information for various benefits, which are enabled by direct applications of machine learning algorithms. As an example, resident monitoring for smart home applications can facilitate tracking and recognition of daily activities for Alzheimer's patients and elderly care. Several initiatives such as [Kidd et al., 1999, Cook et al., 2003, Intille et al., 2006, Kröse et al., 2008, Cook et al., 2009] have brought various researchers to achieve significant contributions for activity recognition in smart environments.

Human annotated data may not be aligned consistently in terms of the boundary between activities. In other words, the impurity of the associated activity labels within a temporal segment may affect the performance quality of classification tasks for activity recognition. It should be noted that throughout the course of this chapter, labels and annotations are used interchangeably in the context of activity recognition corresponding to class labels.

On the other hand, it is crucial to consider a more dynamic and scalable use case for activity recognition. In a real-world situation, multiple activities that can be observed from the streaming of ubiquitous sensor data. However, this is not well explored in the current literature. Existing works mainly focus on single activity recognition from one label set. In common practice, the techniques associated to temporal segmentation involve optimised processes for one-dimensional space of an activity label set. It is a standard mechanism before recognition process (i.e., classification of activities). However, these techniques may need to be adjusted for multi-activity recognition scenario.

For many years, major works have been performed in order to understand the human behaviour and mobility patterns in various interdisciplinary studies. Previous work by Gonzalez et al. [2008] had shown the individual mobility patterns from trajectories of 100,000 mobile phone users in order to signify that each individual is characterised by time-independent travel distance and the probability of return to frequently visited locations. As the results, they could obtain the likelihood of a mobile phone user in any location. Given the opportunities by ubiquitous computing capabilities, the application scenarios of activity recognition in unconstrained environment could be limitless.

To illustrate the importance such recognition in a smart environment, possible scenarios that can be applied in the medical sector include automatic recognition of user (staff) activities in smart hospital [Sánchez et al., 2008] and emotion recognition for measuring mental fitness [Tacconi et al., 2008, Rachuri et al., 2010]. Therefore, the inferences of multi-activity recognition from ubiquitous sensors can be leveraged to facilitate human mobility and contextual modelling in these dynamic smart environments.

These directions suggest the tremendous needs for innovations and efficient methods to handle streaming of ubiquitous, yet irregular multivariate sensor data. Hence, it is a significant challenge to improve the performance of multi-activity recognition. This is due to the fact that human activity in a smart environment is bound to be dynamic. In addition, dynamic changing of activities in data streams will incur a mixture of these activity labels during temporal segmentation process. Therefore, methodologies in minimising the purity in temporal segments are needed without sacrificing much on the performances of activity recognition. In this chapter, a window refers to a temporal segment in sliding window model. Moreover, scalability of these methodologies is required to fit into multi-activity recognition scenario.

The contributions of this chapter are entailed on the preprocessing of ubiquitous sensor streams for activity recognition tasks. In particular, our proposed technique provides these following key contributions:

1. The multi-objective technique of finding optimal window size for time-interval based temporal segmentation in streaming fashion. It is derived based on gaining the balance between minimising label impurity in a segment and maximising factor for class separability (divergence) of ubiquitous sensor features towards class labels.

2. Robust recommendation of optimal time-interval window sizes for temporal segmentation in multi-activity recognition.

## 2.2 Related Works

The study of human activity recognition is a well-known area in many research communities, including computer vision [Turaga et al., 2008, Aggarwal and Ryoo, 2011]. As the proliferation of mobile and ubiquitous devices become prominent, activity recognition from streaming sensor data becomes an emerging research area. It is difficult to be neglected due to its apparent realisation in near future by enabling Internet of Things (IoT) technology. The inherent challenges of dealing with data streaming from these ubiquitous sensor devices (such as wearables, mobile devices and on-body sensors) are related to noisy sensor reading due to hardware limitation and environmental influences. Su et al. [2014] recently presented a general overview of techniques and challenges in performing activity recognition from the smartphone sensors. In the study of activity recognition, it is often treated as a classification problem since there are annotations associated with certain states (e.g., human locomotion activities) for training and testing phases. However, it is common practice to use one second window size for temporal segmentation in many experiments [Pirttikangas et al., 2006, Zhu and Sheng, 2009, Wang et al., 2009, Gordon et al., 2014]. Torres et al. [2013] verified in their recent study that the fixed time window (FTW) achieved high performance of real-time activity recognition. In this chapter, we refer their FTW as time-interval based segmentation. They also have performed comprehensive performance evaluation of other segmentation techniques such as activity windowing, dynamic windowing and mutual information windowing.

In addition, Guo et al. [2012] have proposed an adaptive approach for online segmentation through PCA feature selection and model selection. In their approach, the window size can expand based on feature selection and model selection criteria in order to incorporate the next frame. Unfortunately, the mixture of labels is not included as an essential factor during their segmentation process. It is important to note that human activities are dynamic and subject to continuous change in data streams. Therefore, we consider that maintaining high label purity in a temporal segment is crucial for activity recognition application. However, it is insufficient to simply consider label purity alone. Another objective that can be considered in maintaining label purity of temporal segments is related to the importance of features for activity labels. In many image classification experiments such as [Guo et al., 2008], class separability is crucial to select important features in order to improve and accelerate classification tasks.

Therefore, our study in this chapter is aimed towards finding the optimal time window size to mitigate from simple selection of window size from common heuristics (e.g., one second windowing). In the past literature, dynamic programming [Bellman, 1961] and $k-$segmentation [Himberg et al., 2001] can be used to perform time series segmentation for the purpose of context recognition. However, the drawback of these techniques is that it requires offline data processing for such context recognition. In real world application, streaming of sensor data can be irregular, especially when a sensor device is unavailable under certain circumstances. For example, GPS sensor of smartphone is unreliable when the user is inside a building, or in an underground tunnel. Such temporal data (especially event-based sensor data) can be challenging to process in a streaming fashion. On the other hand, Krishnan and Cook [2014] recently proposed a sliding window approach to perform activity recognition in a streaming fashion, which can be adapted for event-based sensor streaming. Their method incorporates time decay and mutual information based weighting of sensor events within a window. They claimed that different activities can be characterised with different lengths of sensor events. Consequently, their experimental result is validated and tested on the real-world smart home dataset, which mostly consists of event-based sensing data. However, the scope of their study is restricted towards one dimension of activity set.

Most of the research problems in past studies [Chen et al., 2012a, Riboni and Bettini, 2011, Hemminki et al., 2013, Feng and Timmermans, 2013, Kim et al., 2014] addressed the challenge to perform a context recognition on one label set. In other words, the primary objective is to classify a class label correctly from a context label set. For example, a system needs to recognise user walking from home to work, given an activity label set that is composed of “walking”, “standing” and “sitting”. In many real-world scenarios, multiple activities are required to recognise to indicate a user’s intention or action. To provide meaningful contextual information, multi-label classification can be leveraged to facilitate such needs. The problem of multi-label classification is practical in real-world scenarios as a subject can be associated with multiple annotations at a time. For example, a person is “sitting” in a cafe while “drinking coffee”. On another occasion, she can be “running” while “listening to music” from a smartphone. In the medical application, multi-label classification can be used for diagnosing diseases, such as diabetes and prostate cancer [Tsoumakas and Katakis, 2006]. Furthermore, it can also be used for classifying several characteristics during real-time ECG (Electrocardiography) analysis of data streaming from on-body sensors. Moreover, the multi-label problem is not strictly limited to smart environment scenario. Consequently, it is also applicable for modelling the mobility of a user in a dynamic urban environment. For example, Read et al. [2016] proposed a multi-label oriented technique to preprocess sensing data to produce labelled data that are reliable for human mobility modelling. Due to various benefits that can be attained for multi-label classification, the motivations of our studies are influenced to pursue multi-activity recognition from continuous streaming of sensor data. A very simplistic approach for solving multi-label classification is to construct a new label set that is composed of the possible combination of labels from predefined label sets.

Throughout this chapter, we define the steps to find a reasonable parameter (window size) for time-interval based temporal segmentation. They are designed to be robust for multi-activity recognition.

## 2.3 Problem Definition

### 2.3.1 Impure Windows in Temporal Segments

In order to identify certain activity label on data streams from ubiquitous sensors, it is a common practice that the data points are annotated in a time-interval manner. In several scenarios, these sensor data could arrive in irregular manner at different point of time. Temporal segmentation is necessary in order to define the boundary of feature extraction process for recognition phase.

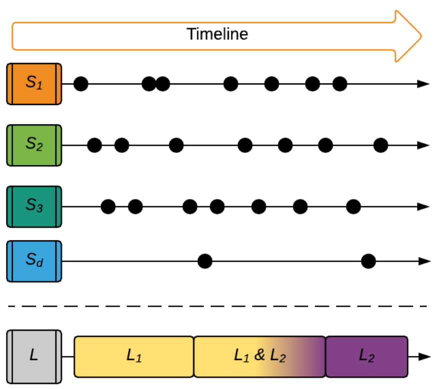



Figure 2.1: Non-overlapping temporal segmentation of sensor data streams with a given label set $L$. An impure segment is composed of more than one label.

Let $S = \{S_1, S_2, S_3, S_4, ..., S_d\}$ be the $d$ number of sensors and $S_i$ $(1 \leq i \leq d)$ is a sensor identifer within the range of $d$ sensor streams. Each arrival of the sensor reading instance $I_{ji}$ can be associated to a unique timestamp $t_j$ that is continuous in a sensor stream $S_i$.

Let us consider a temporal segmentation through sliding window technique where a data stream is processed in a continuous manner. It is apparent that temporal segmentation would produce several windows that are composed of mixtures of multiple annotations from a label set $L = \{L_1, L_2, ..., L_n\}$ (Figure 2.1). Inherently, the most dominant label would be selected as the final label for the corresponding window.

Assuming the process is performed via a time-interval based temporal segmentation approach with non-overlapping sliding window, the problem is formulated as to how to find a reasonable time window given the impurity of segments in ubiquitous sensor streams. The complexity of the problem increases when these sensor streams need to be synthesised due to the variation of data arrival from heterogeneous sensors. Synthesis in this chapter refers to the synchronisation process to align the time segments of heterogeneous sensor streams in a format that is suitable for features extraction.

It is assumed that a complex streaming scenario involves different timestamp for the arrival of data from each sensor. A sensor may be inactive for a certain period of time depending on particular predefined rules. For example, the GPS sensor of a smartphone user is activated only when the person is close to specific locations. When the battery level of the smartphone is low, the sampling frequency of sensor data may be reduced or scheduled in a short time-span in a controlled interval. Hence, a common approach, such as frame-based temporal segmentation would be inappropriate due to these irregular behaviours of sensor data streaming. Hence, the problem definition defined in this chapter is constrained to time-interval based temporal segmentation for streaming ubiquitous sensor data.

A common practice is to assume a predefined fixed window size for the segmentation process. For activity recognition on motion sensors (such as the accelerometer), the one-second window size is commonly assumed in many experiment settings [Pirttikangas et al., 2006, Zhu and Sheng, 2009, Wang et al., 2009, Gordon et al., 2014], including applications in the medical field such as monitoring for Parkinson's disease rehabilitation [Cancela et al., 2015]. Another alternative method is to perform sensitivity analysis on various time-interval window size based on quality metrics of recognition results (e.g., accuracy of classifier validation results). In most cases, sensitivity analysis requires a significant amount of time to evaluate. In the recognition phase (i.e., classification tasks), another sensitivity analysis may need to be performed on each classifier for parameter tuning of machine learning algorithms.

### 2.3.2 Multi-label Problem of Temporal Segments

As described previously, real world applications may require multiple annotations to be associated with an instance. This is commonly addressed as multi-label problem. Assume that label vector $L$ can be constructed with multiple label sets $L = \{L_1, L_2, ..., L_k\}$, which are derived from human annotations. Each label set $L_k$ is composed of multiple labels $L_k = \{L_{k.1}, L_{k.2}, ..., L_{k.n}\}$. Therefore, it is clear that $I_{ji}$ can be associated with annotations from each $L_k$ label set as depicted in Figure 2.2.

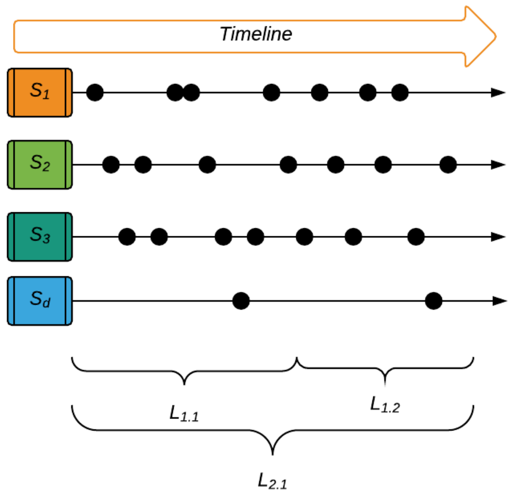



Figure 2.2: Sensor data streams (annotated with multiple labels).

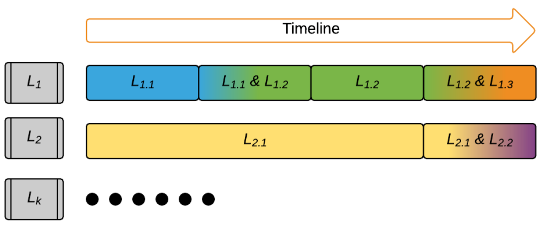



Figure 2.3: Impure segments in multi-label scenario.

Previously the problem of impure segments in temporal segmentation is explained. However, this problem is not limited to annotations from one label set. Hence, the impure segments can be scaled to several dimensions, given the annotations from heterogeneous label sets (as illustrated in Figure 2.3). Each label set may be associated with others or be independent according to certain contexts. Thus, finding a reasonable window size for temporal segmentation presents as a crucial problem in multidimensional scales.

Intuitively, it is feasible to find a balance for recommending optimal time-interval window size by minimising the impurity proportion of windows in temporal segmentation. Furthermore, the magnitude of complexity increases when sensor data streams are associated with annotations from multiple label sets. This is due to the assumption that each label set may have different optimal window size. For example, detecting a human activity from {standing, walking, sitting} may have smaller window size in comparison to recognising higher level activity from {cooking, exercising, relaxing on couch, eating} set. In addition, window size may be significantly different in the scenario where multiple activities must be predicted from given $k$ label sets. In other words, the dynamic combination of items from heterogeneous label sets requires a balanced window size. Therefore, our contribution aims to improve and to accelerate part of processes in multi-activity recognition.

## 2.4 Methodology

Finding a reasonable window size for temporal segmentation requires a significant amount of time and effort from a typical sensitivity analysis of output quality of classifier results, especially in multi-activity scenarios. Hence, we present a data-driven approach of time-interval window size recommendation from annotated multivariate data streams. The objective of our algorithm is to achieve the balance between minimising impurity of segments in data streams and maximising factor for class separability based on given $k$ number of label sets. Given the proposed multi-objective method of data-driven approach in finding optimal window size, we formalise the following metrics:

1. Impurity proportion of segments
2. Class separability

### 2.4.1 Impurity Proportion

In real world scenario, a window from temporal segmentation may contain a mixture of multiple annotations. Therefore, it is impractical to assume that a segment can contain only one annotation. We consider the portion of annotation mixture is important in the case of transition points between activities. In this chapter, we define impure segments for the windows that contain more than one annotation from temporal segmentation.

The first objective of our method requires minimising the impurity proportion. Impurity proportion can be calculated via:

$$pimpure = \frac{m_{impure}}{m} \tag{2.1}$$

where $m_{impure}$ is the count of impure segments over total $m$ segments in data streams.

### 2.4.2 Class Separability

Class separability for data streams refers to the notion of representation for distinct separation between features in data streams with respect to each class label. Typically, the class separability score is calculated for each feature in a given data distribution. In this case, feature values of sensor streams are temporarily stored, which then be processed in order to identify which features are dominant towards a class label. In our method, Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] is used to calculate the class separability score. This is a common technique to measure the difference between two probability distributions $P$ and $Q$.

It is assumed that the streaming of sensor data is constrained by numeric features. The overall values of each feature can be represented in histogram (frequency distribution) format. In other words, there would be one histogram per feature for each class. Hence, KL divergence can be calculated for each feature after the histogram representation has been attained from discretization using $b$ equally-spaced bins. This histogram representation would fit into different scenarios where the numeric features are discrete or continuous. In other problem domain, class separability is used for feature selection [Benoit et al., 2015, Guo et al., 2008, Liu et al., 2010, Oh et al., 1999, Zhang et al., 2013, Zhou et al., 2014].

For each feature $f$, we calculate maximum score of the class separability (i.e., maximum divergence of $f$) as:

$$mdivergence_f = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} KL_f(i, j), \tag{2.2}$$

where $n$ is number of class labels in a label set $L_k$ and $KL_f(i, j)$ is the KL distance between two distributions (histograms) corresponding to class labels $i$ and $j$:

$$KL_f(i, j) = \sum_{m=1}^{b} Pr_f(m \mid i) \; log \left( \frac{Pr_f(m \mid i)}{Pr_f(m \mid j)} \right) \tag{2.3}$$

where $b$ is the number of bins in the histograms. Both probabilities $Pr_f(m \mid i)$ and $Pr_f(m \mid j)$ should never be 0; instead we define $0.0000001 \leq Pr_f(m \mid l) \leq 1$ heuristically where $l$ is the class label index. In any case, a larger value of $mdivergence_f$ relates to greater separability of $f$ over all labels in $L_k$. As KL divergence holds asymmetric property, it is practically useful to derive which top-$k$ features provide significant contributions for a classifier by ranking class separability scores. However, this can be replaced with a symmetric solution such as Jensen-Shannon (JS) divergence [Lin, 1991] when a strict metric is required. Despite the distinction between symmetrical and asymmetrical divergence distances, it is prominent that the average of these pairwise measures is typically used for the extension to multi-class scenario [Guo et al., 2008]. In our case, the mean is derived from maximum divergences for all features. Maximum divergence was also used as scoring in [Abeel et al., 2009] for feature selection.

### 2.4.3 Multi-objective Function

Based on the above two critical metrics, multi-objective function is leveraged that aims to gain balance between minimising the impurity proportion in data streams and maximising the divergence in terms of class separability of activity recognition. The balance is found by looking for optimal window size that has impurity proportion measurement close to the intersection of metrics within the defined window size search range. The details of multi-objective operation would be elaborated in the following algorithm section.

### 2.4.4 Algorithm

Our proposed method is formalised as Optimal Window (OPTWIN) for time-interval based temporal segmentation. In order to recommend time-interval window size for segmentation, it requires finding the balance between the above impure proportion and class separability measure.

As described in Algorithm 1, the recommendation of window size include the initial parameters of:

1. Three main inputs for the search range of window size:
    a) Smallest window size $ws_{start}$.
    b) Biggest window size as the upper boundary $ws_{end}$.
    c) Time-interval step $ws_{step}$ for window size.
2. An additional input for data streams from ubiquitous sensors $D_{streams}$.

**Algorithm 1** OPTWIN based window size recommendation

**procedure** $recommendWindowSize(ws_{start}, ws_{end}, ws_{step}, D_{streams})$
    **global** $ws_{current} \leftarrow ws_{start}$
    **while** $ws_{current} \leq ws_{end}$ **do** ▷ Label purity and divergence derivation
        **for each** $L_k \in L$ **do**
            $D_{segments} \leftarrow temporalSegmentation(ws_{current}, L_k)$
            $pimpure_{L_k}[ws_{current}] \leftarrow impurityProportion(D_{segments})$
            $multiActivities \leftarrow isMultiActivities(L_k)$
            **if** $multiActivities$ **then**
                $discore_{L_k}[ws_{current}] = \frac{1}{i} \sum_{i=1}^{n} discore_{L_i}[ws_{current}]$
            **else**
                **for each** $f$ **in** $D_{segments}$ **do**
                    $maxdivergences_{L_k}[f] \leftarrow maxKL(values_f, L_k)$
                **end for**
                $discore_{L_k}[ws_{current}] = \frac{1}{f} \sum_{i=1}^{f} maxdivergences_{L_k}$
            **end if**
        **end for**
        $ws_{current} \leftarrow ws_{current} + ws_{step}$
    **end while**

    **for each** $L_k \in L$ **do** ▷ Optimal window size recommendation
        $ws_{current} \leftarrow ws_{start}$
        **while** $ws_{current} \leq (ws_{end}$ **do**
            $dideviation_{current} \leftarrow nrmlDvt(discore_{L_k}[ws_{current}], discore_{L_k})$
            $metrics_{current} \leftarrow (pimpure_{L_k}[ws_{current}],\ dideviation_{current})$
            $ws_{next} \leftarrow ws_{current} + ws_{step}$
            $dideviation_{next} \leftarrow nrmlDvt(discore_{L_k}[ws_{next}], discore_{L_k})$
            $metrics_{next} \leftarrow (pimpure_{L_k}[ws_{next}],\ dideviation_{next})$
            $ws_{optimal} \leftarrow intersect(metrics_{current},\ metrics_{next})$
            **if** $ws_{optimal}$ **then**
                $ws_{rec}[L_k] \leftarrow ws_{optimal}$
                **break;**
            **end if**
            $ws_{current} \leftarrow ws_{current} + ws_{step}$
        **end while**
    **end for**

    **return** $ws_{rec}$
**end procedure**

This algorithm requires scanning operation for impurity proportion and class separability measure of every window size (with incremental of $ws_{step}$) within the boundary of $ws_{start}$ and $ws_{end}$. The computation of impurity proportion $pimpure_{L_k}$ and divergence score from all maximum divergence of features $discore_{L_k}$ are stored for each label set $L_k$. In this case, the average function of all maximum divergences of features is used to compute $discore_{L_k}$.

The process of window size recommendation is composed of two stages. Firstly, the necessary metrics (e.g., impurity proportion and divergence score) are computed and stored for each $L_k$ from a given $L$ label sets in data streams $D_{streams}$. This stage is referred as label purity and divergence derivation in Algorithm 1. Within iteration of every possible window size, the derivation begins with temporal segmentation procedure where time-interval segmentation is performed on the given $D_{streams}$. In temporal segmentation procedure, feature extraction is performed to derive the summary of data points in each time window. It should be noted that the results of feature extraction would be later used in calculating the class separability score. Afterwards, the impurity proportion of each window size $pimpure_{L_k}[ws_{current}]$ is calculated through Equation 2.1. The divergence score $discore_{L_k}[ws_{current}]$ is then calculated subsequently through the mean maximum divergence from $maxdivergences_{L_k}[f]$ of all $f$ features. The corresponding $maxdivergences_{L_k}[f]$ is computed through Equation 2.2. All of these metrics are stored temporarily for optimal window size recommendation stage.

Furthermore, for any $L_k$ that is defined as multi-activity (i.e., $L_k$ is composed by combination of several other $L_k$ in $L$ label sets), mean maximum divergence can be computed from the following:

$$discore_{L_k}[ws_{current}] = \frac{1}{i}\sum_{i=1}^{n} discore_{L_i}[ws_{current}] \tag{2.4}$$

where $L_k$ is identified as multi-activity label set and $L_i$ is single-activity label set that is part of $n$ label sets being associated with composition of corresponding $L_k$. For example, let us consider a multi-activity scenario where $L_k$ is composed of combination between locomotion activity $L_1$ label set and gesture activity $L_2$ label set. Both of label sets $L_1$ and $L_2$ are considered as single-activity label sets. Therefore, the result $discore_{L_k}[ws_{current}]$ is derived from averaging divergence score $discore_{L_i}[ws_{current}]$ from $L_1$ and $L_2$ corresponding to the same window size.

The second stage after metrics derivation refers to optimal window size recommendation in Algorithm 1. It involves finding optimal window size for each label set $L_k$ with given metrics computed from the first stage. For each label set $L_k$, the optimal window size is derived from metrics intersection between current window size and next window size in an iterative operation. In order to decide the window size to be optimal, the metrics for impurity proportion and normalised squared deviation of divergence score are used.

Given the divergence score $ds_i$ and mean of divergence scores ($\overline{ds}$) from all possible window sizes, normalised squared deviation $ndd_i$ can be computed via:

$$ndd_i = \frac{dd_i - \min(dd)}{\max(dd) - \min(dd)} \tag{2.5}$$

where squared deviation $dd_i$ of $ds_i$ is defined as:

$$dd_i = (ds_i - \overline{ds})^2 \tag{2.6}$$

As a result, divergence deviation of available window sizes would be scaled from 0 to 1 and can be used to find the intersection with impurity proportion. For temporal segmentation, larger time-interval window size corresponds to the following assumptions:

1. Greater impurity proportion is included.

2. Increasing mean of maximum divergences from all features.

These assumptions would be validated via the experiments in next section. In Algorithm 1, the intersection function would return the optimal window size $ws_{optimal}$. The $ws_{optimal}$ is non-empty in a condition of $range_{pimpure} \cap range_{dideviation}$ where

$$range_{pimpure} = (pimpure_{L_k}[ws_{current}],\ pimpure_{L_k}[ws_{next}]) \tag{2.7}$$

and

$$range_{dideviation} = (dideviation_{current},\ dideviation_{next}) \tag{2.8}$$

Given the ranges $range_{pimpure}$ and $range_{dideviation}$, the optimal window size is defined based on the first intersection occurrence. Hence, the candidate of optimal window size (either $ws_{current}$ or $ws_{next}$) for $ws_{optimal}$ depends on closest absolute distance of impurity proportion (either $pimpure_{L_k}[ws_{current}]$ or $pimpure_{L_k}[ws_{next}]$) to average value (centroid) of $S$ value set:

$$S = \left\{ \begin{array}{l} pimpure_{L_k}[ws_{current}], \\ pimpure_{L_k}[ws_{next}], \\ dideviation_{current}, \\ dideviation_{next} \end{array} \right\}$$

The final output of algorithm results in optimal window sizes $ws_{rec}$ corresponding to all $L_k$ in $L$ label sets.

## 2.5 Experiments and Evaluation

In this section, we present the experimental settings and evaluation. The method is validated with the benchmark OPPORTUNITY Activity Recognition Dataset [Sagha et al., 2011] from UCI repository. It is a dataset that was collected from wearable, object and ambient sensors for activity recognition. This rich dataset contains 242 features in total. For the purpose of this study, we only use 101 features that are related to body sensors and wearables. The remaining unused features are associated with sensors that are attached to objects such as knife, spoon, plate and fridge.

There are several characteristics of this dataset that can be associated to our problem definition:

1. Dynamic sensor data from ubiquitous sensors in an **irregular manner**. In several occasions, feature values in a sensor stream can be empty due to unavailability of specific ubiquitous device. For example, unavailability of accelerometer sensor in certain time period would result in empty reading for three axes of accelerometer values. Therefore, the data contains inherent sparsity problem, which may result in less accurate and inconsistent performance of a classifier model.

2. The dataset contains **multi-activity label sets**, which is suitable for our evaluation in terms of scalability of the proposed method.

### 2.5.1 Data Preparation

The OPPORTUNITY Activity Recognition Dataset contains several activity sets including 5 high level activities, 4 locomotion activities, 17 gesture activities, low-level actions relating 13 actions to 23 objects. There are 4 users associated with the sensor data, 6 recordings for each user. In our experiments, we randomly selected 2 recordings for each user. Time-interval based temporal segmentation is performed for each recording in a streaming fashion. All temporal window instances from selected recordings of users are combined, producing a dataset that would be used for training and testing phases.

Consequently, our method would be demonstrated and validated with this dataset for multi-activity recognition scenario. In this instance, we use the following label sets:

1. ***High level activity (HLA)*** label set:

$$L_1 = \left\{ \begin{array}{l} \text{Relaxing,} \\ \text{Coffee time,} \\ \text{Early morning,} \\ \text{Cleanup,} \\ \text{Sandwich time,} \\ \text{None} \end{array} \right\}$$

2. ***Locomotion activity (LA)*** label set:

$$L_2 = \left\{ \text{Stand, Walk, Sit, Lie, None} \right\}$$

3. ***Gesture activity (GA)*** label set:

$$L_3 = \left\{ \begin{array}{l} \text{Open Door 1, Close Door 1,} \\ \text{Open Door 2, Close Door 2,} \\ \text{Open Fridge, Close Fridge,} \\ \text{Open Dishwasher, Close Dishwasher,} \\ \text{Open Drawer 1, Close Drawer 1,} \\ \text{Open Drawer 2, Close Drawer 2,} \\ \text{Open Drawer 3, Close Drawer 3,} \\ \text{Clean Table, Drink from Cup, Toggle Switch,} \\ \text{None} \end{array} \right\}$$

For the instance without class label $L_k$ above, it is automatically assigned to "None". Furthermore, feature extraction is subsequently performed on each window that is produced by temporal segmentation process. The generated features include: *mean*, *median*, *minimum*, *maximum*, *standard deviation*, *IQR (interquartile range)*, *median* and *RMS (root-mean-square)*.

As a result, there are 707 features that can be used for training and testing a classifier. For recognition of activities, these following classifiers are used:

1. Naive Bayes (**NB**) classifier
2. Decision Tree (**J48**) classifier
3. Random Forests (**RF**) classifier

Throughout our experiments, the activity recognition models are built using a well-known data mining software: Weka 3.8 [Hall et al., 2009] with default parameters corresponding to each classifier. For multi-activity recognition, a new label set is constructed from the combination of three label sets. In other words, the new label set of multi-activity recognition $L_4$ contains labels in a given structure $L_1$:$L_2$:$L_3$ (denoted as **HLA:LA:GA**), e.g., "Coffee time:Sit:Drink from Cup".

The boundary to search for optimal window size is defined as the following:

- $ws_{start} = 600$ milliseconds (0.6 seconds).
- $ws_{end} = 5000$ milliseconds (5 seconds).
- $ws_{step} = 100$ milliseconds (0.1 seconds).

### 2.5.2 Observation of OPTWIN Metrics

In order to validate our assumptions on increasing window size for temporal segmentation towards OPTWIN metrics, impurity proportion and mean of maximum divergences (from all features) are observed in this section.

#### 2.5.2.1 Impurity Proportion

As shown in Figure 2.4, the impurity proportion increases as window size used for temporal segmentation is larger. The decrease of purity is especially prominent for **LA**, **GA** and **HLA:LA:GA** label sets. However, it is less noticeable for **HLA** as its class labels have larger activity duration compared to other label sets. It is within the expectation that the impurity proportion for **HLA:LA:GA** label set would be significantly dominant. Essentially, this can be caused by the dynamic combination of classes from available label sets.

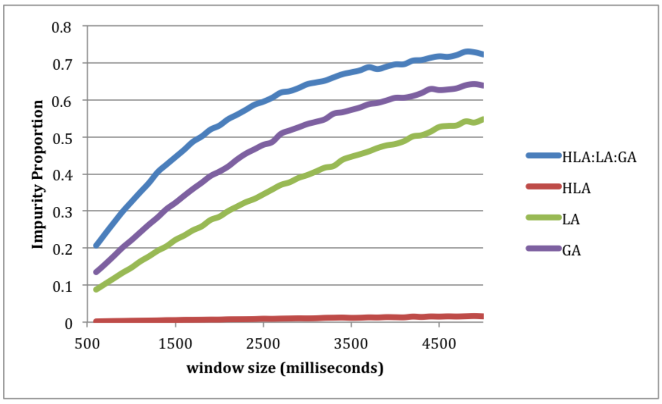



Figure 2.4: Impurity proportion.

Furthermore, we noticed that the locomotive state changes frequently in OPPORTUNITY dataset (as described in Figure 2.5). These activities include instances without a class label, which are reassigned as "none". As a result, the frequent changes of locomotive state lead to impurity level exceeding 50% at 4.5 seconds (window size) for locomotion activities.

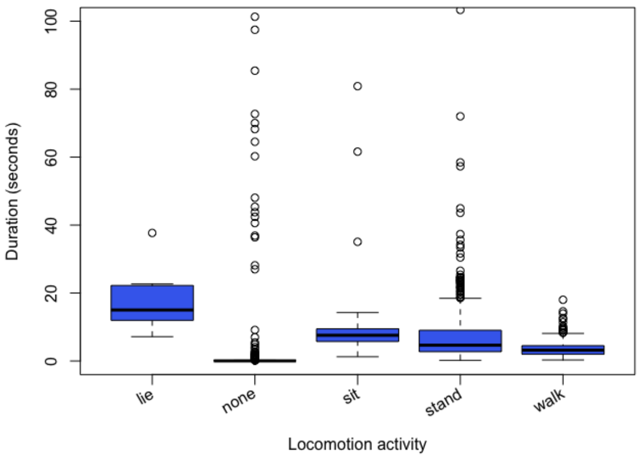



Figure 2.5: Box plot of locomotion activity duration (zoomed). Maximum locomotion activity duration is 224.994 seconds for "none" class label.

#### 2.5.2.2 Mean Maximum Divergence

Before calculating maximum divergence for each feature, the feature values are converted to a histogram as explained in the methodology section. However, the sparsity problem is inherent in the OPPORTUNITY dataset since a sensor can be unavailable for a period of time in a stream. To solve this problem in relation to KL divergence computation, the empty feature values are allocated to the middle of defined $b$ bins. In this case, 100 bins ($b = 100$) are used for the normalisation to a histogram. For the label set that is used for multi-activity recognition, $discore_{L_4}$ (mean maximum divergence of **HLA:LA:GA** label set) is computed by averaging the mean maximum divergences of associated label sets. In other words, $discore_{L_4} = \frac{1}{3}(discore_{L_1} + discore_{L_2} + discore_{L_3})$.

Similarly, the mean average divergence metrics appear to be increasing corresponding to larger time window size (up to 10 seconds) as shown in Figure 2.6. Moreover, the mean maximum divergence score is expected to degrade after reaching the peak at a certain time window size. This phenomenon is clearly shown in the sudden drop of $discore_{L_4}$ and $discore_{L_3}$ at 33 seconds window size. In many applications, this could be viewed as a convex optimisation problem. However, our observation revealed that larger window size would eventually reduce the number of unique class labels (especially for $L_3$ and $L_4$). This observation indicates that the optimal divergence score would not be convincing since the number of unique class labels will significantly decrease if activities are shifted rapidly. It should be noted that the range of divergence score to be used in our analysis is defined between 600 milliseconds and 5000 milliseconds for finding optimal window size.

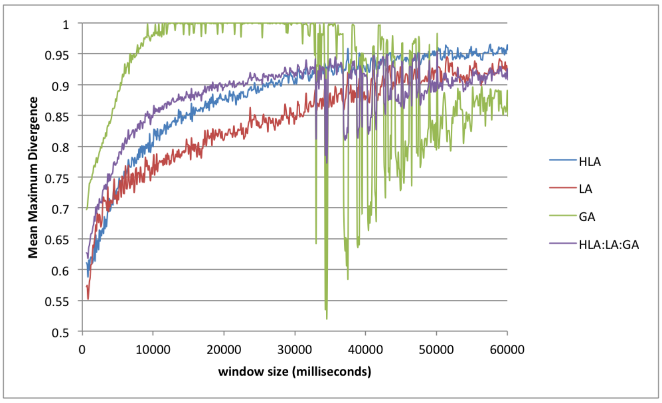



Figure 2.6: Mean of maximum divergences.

### 2.5.3 Optimal Window Sizes

The OPTWIN algorithm finds the optimal window sizes for label sets based on the above metrics within a constrained search range. The optimal window size is set from the closest window size's impurity proportion to the intersection of impurity proportion and normalised deviation of divergence score. As an example, the optimal window size for **HLA:LA:GA** label set is found to be 1.2 seconds as depicted in Figure 2.7.

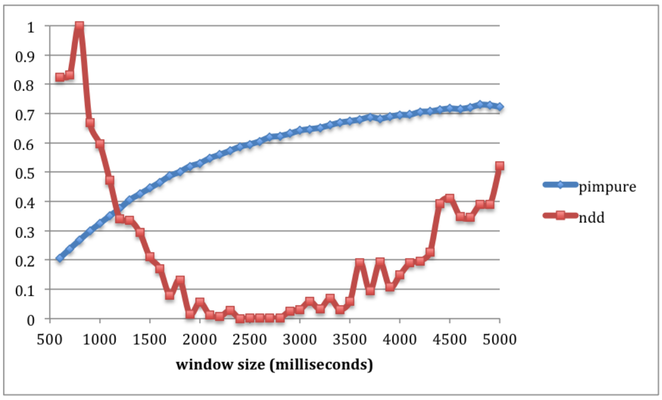



Figure 2.7: Intersection between impurity proportion (pimpure) and normalised squared deviation of divergence score (ndd).

Performing sensitivity analysis of every label set would be a time consuming task in overall activity recognition operation, excluding parameter tuning for each classifier. The proposed method specifically operates as multi-objective function that intends to gain the balance between minimising the impurity proportion of time segments and maximising class separability measure. Moreover, the algorithm is scalable to tackle multi-activity recognition problem. For our experiments, 1-second window size is used as the baseline comparison. It was mentioned previously that many experiments leveraged 1-second as a heuristic way to select the window size, especially for activity recognition. The performances of classifiers are tested with F1 score (harmonic mean of precision and recall) via 10-folds CV (Cross Validation) on optimal and baseline window sizes.

The following $ws_{optimal}$ are returned from the window size recommendation:

1. **HLA**: 2.4 seconds - single-activity recognition
2. **LA**: 1.5 seconds - single-activity recognition
3. **GA**: 1.4 seconds - single-activity recognition
4. **HLA:LA:GA**: 1.2 seconds - multi-activity recognition

### 2.5.4 Sensitivity Analysis

In this section, the result of sensitivity analysis for multi-activity recognition is shown by the performance of classifiers on $L_4$. Thus, this brief analysis will verify that general classifier performance varies for each window size. The performances of classifiers are tested with both F1 score and accuracy via 10-folds CV. As shown in Figure 2.8 and Figure 2.9, fluctuation of quality performances on the classifiers are prominent for both F1 score and accuracy as the window size increases, especially on RF (best classifier). Thus, this is essentially aligned with the motivations and challenges in this research to find optimal window size for multi-activity recognition. It should be clear that multi-activity recognition in this chapter aims to detect multiple activities (one activity from each label set) for a given time window instance.

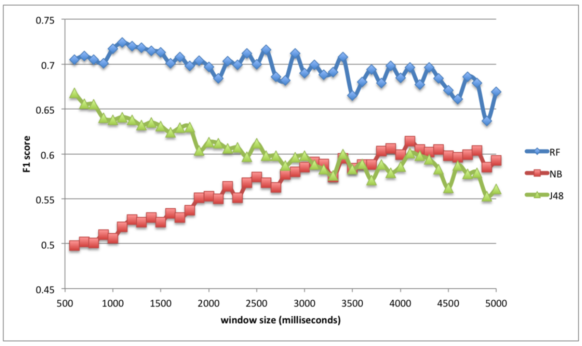



Figure 2.8: Sensitivity analysis - F1 score (OPPORTUNITY).

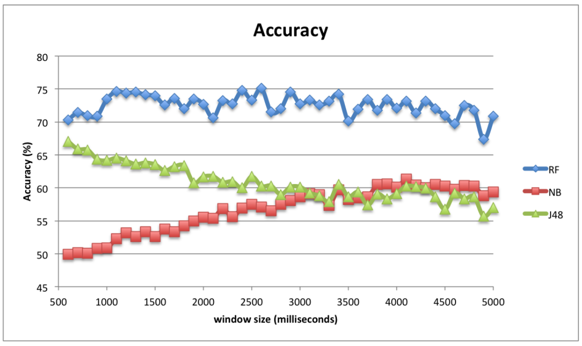



Figure 2.9: Sensitivity analysis - Accuracy (OPPORTUNITY).

### 2.5.5 Discussion for Optimal Window Size

From all performance comparisons of F1 scores shown in Figure 2.10 and Figure 2.11, NB classifier gains significant improvement for F1 scores for the defined label sets. This suggests that finding optimal window size would generally improve the performance of generative model based classifier such as NB in comparison to the baseline (1-second window size). It is important to note that in several applications of activity recognition, generative models can outperform discriminative approaches, especially when there are errors associated with activity labels (e.g., streaming data from crowdsensing [Ganti et al., 2011]). However, the performances of F1 scores appear to be slightly worse for the optimal window size of single-activity recognition of **HLA** and **LA**. The slight degradation of these classifier performances can be found as the trade-offs to finding balanced window size for recognising multiple activities. However, an improvement of single activity recognition for the atomic activity (**GA**) is gained. In short, finding optimal window sizes through OPTWIN algorithm corresponds to better and balanced performance for multi-activity recognition (e.g., **HLA:LA:GA**) in comparison to single activity recognition from a single label set. Furthermore, the recommended window size can be used for real-time activity recognition according to user application context (either single-activity recognition or multi-activity recognition).

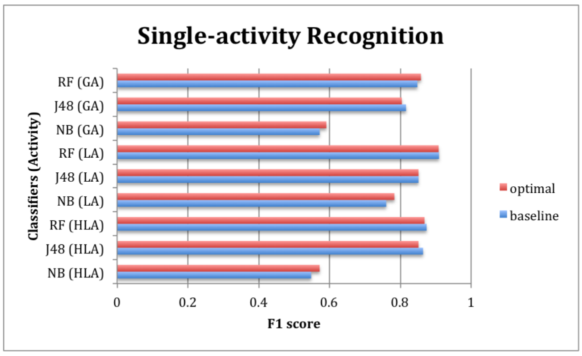



Figure 2.10: F1 score comparison between optimal and baseline window size for single-activity recognition (HLA, LA and GA).

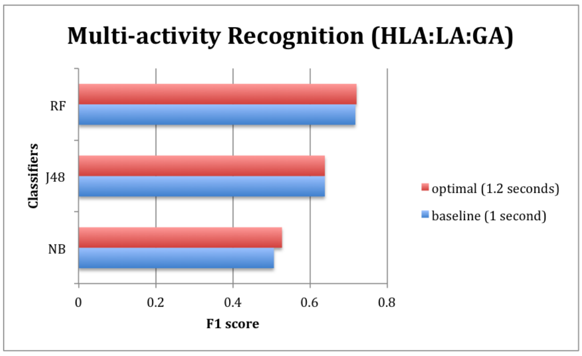



Figure 2.11: F1 score comparison between optimal and baseline window size for multi-activity recognition (HLA:LA:GA).

## 2.6 Conclusion

In this chapter, we have proposed a specialised technique of finding optimal window sizes for time-interval based temporal segmentation. This technique is composed of multi-objective function that aims for the balance between minimising the impurity segments during temporal segmentation and maximising class separability measure. The scalability of proposed technique is demonstrated by the capabilities to produce optimal window size for each label set, including one for the combined label set (for multi-activity recognition). From the experimental results, the optimal window size for multi-activity recognition is found to be improving the quality of classifier performances from the baseline window size. In addition, the improvement appeared to be dominant for generative model based recognition techniques (such as Naive Bayes classifier). Nevertheless, OPTWIN can be used as an alternative method for automatic selection of window size instead of simple one second heuristic or a very time consuming sensitivity analysis.

The contribution of this chapter is mainly targeted towards time-interval temporal segmentation for multi-context recognition purposes. Moreover, another challenge is associated with the study of dependency between activity and context label sets. In this chapter, the multi-label classification is achieved by constructing a new label set. This new label set consists of the combination class labels that we can derive from heterogeneous label sets. Unfortunately, this approach is not robust in terms of inferring the accuracy for each activity label. An ideal solution is needed for the mobile sensing scenario where the human activity and context label sets can be dependent or independent from each other. Therefore, this particular issue of simultaneous multi-context recognition from mobile sensor data is addressed in the next chapter.

# Chapter 3

# RECOGNISING MULTIPLE CONTEXTS FROM MOBILE SENSOR DATA

## 3.1 Introduction

As discussed previously in Chapter 2, the problem of multi-activity recognition in constrained environments (e.g., smart homes) can be expanded further to simultaneous context recognition in-the-wild, via intelligent mobile sensing. In this chapter, we prove that such dependency assumption between contextual labels from multi-dimensional label spaces is extremely important, especially in cases where smart devices (e.g., smartphones) are used to recognise both human activities and transportation modes through daily mobile sensing.

Rapid development and population growth in urban areas have gained tremendous attention from researchers to study human dynamics via mobile sensing technology. In order to close the gaps towards a smarter city, it is also important for the local authorities to discover the relevant knowledge from the observation of human mobility patterns within various transportation contexts. The human movements within the cities are becoming more dynamic and complex due to the high availability of the public transportation services and dynamic environmental factors, especially with the involvement of surrounding spatial and temporal elements that could affect the accuracy in mobile sensing applications.

In daily commuting routines, smartphone users travel between urban areas via certain transportation modes. The dynamic mobility of the humans in each journey involves activities across multiple transportation modes. Hence, the mode of transportation associated with human activities always changes over the time, depending on user's spatial and temporal contexts. For example, in the morning a mobile user travels from home to the office using a bus, train and light rail while he could be standing or sitting while inside those transportation mediums. Furthermore, during the transition between these transportation mediums, the user might be required to walk for a certain distance in order to get onboard on different transportation mode (e.g., train or light rail). At a certain time slot, the user could be relaxing while sitting on a couch at home, which means having no transportation mode for the given contexts of this mobile user. It should be noted that a mobile user in this chapter is defined as a user that frequently moves (i.e., commute) and utilises mobile devices (not limited to smartphones) in daily life.

In major metropolitan areas, smartphones are heavily used and carried everywhere by individuals in order to perform their daily tasks. These smartphones are embedded with various built-in mobile sensors that are capable of perceiving the user contexts, such as transportation modes, human activities and types of environment the mobile users are contained in, from the signals captured within the vicinity of smart devices. In this case, the smart devices are not only limited to first-person perspective (smartphone users), but also to the ubiquitous devices that could be deployed throughout the city. In this chapter, the scope of our study is focused on a non-trivial issue to perform simultaneous inference of human activities, transportation modes and their related contexts from mobile users' perspective for the purpose of intelligent applications. For instance, let us imagine the future intelligent assistant that can automatically understand human commuting behaviour and recommend the best route based on user contexts (including the past behaviour from user's mobility patterns).

In this chapter, transportation mode, human activity and their related contexts (e.g., environment type) can be inferred from the data streamed from the ubiquitous sensors embedded in a smartphone. Such independent assumption of modelling human activities and transportation mode may result in a wrong conceptual interpretation of an action, which is unfavourable for applications such as intelligent assistants. For example, it may be less practical for a human to "stand" in a car while the actual activity in such transportation could be "driving" or "sitting". Furthermore, we focus on the simultaneous inference of mobile user contexts (transportation modes, human activities and associated environment contexts) based on non-location based sensors. As a result, a less reliant solution to location-based sensors (e.g., GPS sensor) can be leveraged for low-energy mobile sensing applications in dynamic urban environments. Inherently, this type of solutions will be an energy efficient option as GPS sensors typically consume more energy in most mobile sensing applications.

Hence, the contributions of this chapter include:

1. More accurate inference of transportation mode and human activities without relying on the locations of mobile users.

2. Simultaneous inference of transportation modes, human activities and mobile user's environmental contexts (e.g., whether the user is in a moving environment or not) based on the dependency of contextual labels of user's annotations.

3. A robust system for mobile context recognition on data streaming from low-energy ubiquitous sensors.

## 3.2 Related Works

For many years, major works have been performed in order to understand the human behaviour and mobility patterns in various interdisciplinary studies. To understand the human behaviour in social interactions, previous study in [Wang et al., 2011] includes empirical analysis and prediction that incorporates the essence of human mobility patterns to network measure such as connectivity in social networks. They derived the mobility patterns from similarities in movements and interactions via trajectories and communication records, which subsequently be correlated with the social connectedness in order to produce models for social link prediction.

In regards to human activity recognition, Su et al. [2014] presented a general overview of techniques and challenges in performing activity recognition from the mobile phone sensors. Mainly, the research in activity recognition is often treated as a classification problem, such as [Chen et al., 2012a, Riboni and Bettini, 2011, Hemminki et al., 2013, Feng and Timmermans, 2013, Kim et al., 2014]. In most cases, there are class labels associated with certain human activities for training and testing phases. Another issue in this domain is related to the adaptability to perform real-time activity recognition on continuous streaming of sensor data. The challenges are predominant in such environments due to the variability and evolving user's behaviours. These changes can be affected by the presence of concept drift [Widmer and Kubat, 1996]. Therefore, adaptive approaches such as [Wang et al., 2012, Abdallah et al., 2015] are recommended to perform incremental learning in order to adjust with the evolving sensor streams and user's behaviours.

Furthermore, Lane et al. [2010] has produced a survey in regards to the significance of the sensors embedded in mobile phones to be integrated within the space of personal, group and community sensing applications. It is aligned with the fact that mobility patterns can be inferred from these powerful mobile phones through variety of built-in sensors. In the transportation domain, it is known that traffic remains a major problem, which is visible in our current society in the last decades. This issue signifies greatly given the growing population and limited services that are able to satisfy the transport needs. Therefore, congestion becomes prominent, which affects the processes in urban planning and traffic management. Hence, it brings a significant motivation for our study to then help in extending the horizon of human mobility research. As one of the benefits, the congestion can be reduced to a certain extent with improved quality of service delivery and transport resource allocation. Since our communities have been emerged with the usage of mobile phones in their daily operations, the potential is limitless to extract more knowledge from human behaviour and mobility patterns. Many works such as [Sohn et al., 2006, Lu et al., 2010] included the tracking of user positioning characteristic in conjunction to activities in order to offer location-based model of mobility patterns.

In the context of human activity recognition based on mobile sensors, segmentation on live streaming data is often required in order to extract the summary for further analysis and prediction tasks. Within the field of activity recognition, relevant solutions such as [Okeyo et al., 2014, Wan et al., 2015, Cho et al., 2015] have been proposed for time segmentation purposes. However, these did not address the problem where each of the sensors can be sampled at a different rate. Thus, it is feasible to have a sparse segment when the time window size is small. This problem was then addressed in the previous work [Liono et al., 2016] for optimal windowing of multivariate sensor data, especially in the scenario of multi-activity recognition. In a larger scenario of human activities in daily life, multi-context recognition is a prominent problem, especially when different environments have their intrinsic sensing patterns. Therefore, the immediate challenges that we face in multi-context recognition in an uncontrolled environment are associated with the noise of streamed sensor data that are subject to different types of user's environments and their activities. Such in-the-wild sensing settings [Vaizman et al., 2018b] require non-trivial consideration of many factors associated with the user's daily life, which increases the magnitude of real challenges and needs in multi-context recognition exponentially. For example, the users may turn off the location tracking (e.g., GPS sensor) in order to preserve the operational time of smartphone in their daily life (i.e., minimising battery usage as highlighted in [Vaizman et al., 2018b]). Hence, another immediate challenge is addressed in this chapter to design and construct a mobile context recognition system that relies less on the location data of users. Consequently, our hypothesis initially assumed that these multiple contextual label sets are associated with each other, reflected by multi-faceted contexts of human activities in the daily life of these mobile users.

The scope of this chapter includes the study of human activity recognition from the mobile sensors that are induced within the context of transport mobility patterns. Therefore, the problems addressed in our study are related to transportation mode discovery and the associated human activities within the temporal domain. The mobility patterns presented are derived from overlapping labels in temporal segmentation of sparse sensor data streams. To the best of our knowledge, there is a lack of study in terms of inferring mobility patterns for recognising both transportation mode and human activity.

## 3.3 Problem Definition

In this section, we present the problem definition based on the applications of smart sensing to infer multiple contextual labels simultaneously from raw activity annotations in daily life. Inherently, these multiple contextual labels are related to mobile user contexts such as transportation modes, human activities and their environmental contexts (i.e., mobility of corresponding transportation modes).

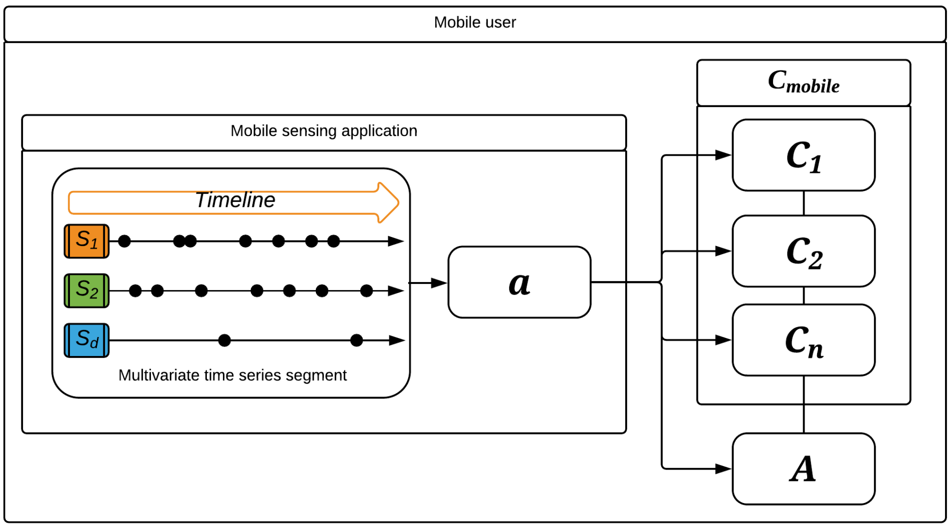



Figure 3.1: Inferring multiple contexts (human activity and related mobile contexts) on multivariate time series data of mobile sensors.

Let $S = \{S_1, S_2, ..., S_d\}$ be the set consisting of $d$ number of mobile sensors of a user and $S_i$ $(1 \leq i \leq d)$ is a sensor identifer within the range of $d$ sensor streams. Each arrival of sensor reading instance point $I_{ji}$ is associated to unique timestamp $t_j$ that is naturally continuous in a sensor stream $S_i$. It should be noted that sensor streams set $S$ correponds to a segment dedicated to a given raw user annotation $a$. In this case, the raw user annotation refers to the short description (typically in a diary-based study) of a human activity and its related contexts (e.g., transportation mode and the type of environment). Given the temporal set $T = \{t_1, t_2, t_3, t_4, ..., t_m\}$ in a segment, the timestamp $t_j \in T$ corresponds to a particular sensor data instance $I_{ji}$. Each instance $I_{ji}$ consists of a collection of values corresponding to multiple channels of a given sensor $S_i$.

Given the feature vectors that can be computed from all segments of human activities in daily life, we formulate the following problem to infer multiple contexts simultaneously from a raw user annotation.

**Definition 1.** *Let a be the instance of a set of raw human annotations* $Annotations = \{a_1, a_2, a_3, a_4, ..., a_r\}$ *in a mobile sensing application.*

For a general overview of the inferring mobile user contexts, Figure 3.1 describes that an instance of user annotation $a$ can be decomposed into multiple contextual labels consisting of human activity $A$ and its related mobile contexts $C_{mobile}$ (including transportation mode).

**Definition 2.** *Therefore, we can define* $a = \{A, C_{mobile}\}$ *where A is the instance of human activity and* $C_{mobile}$ *is the mobile contexts set associated to A. In this case, a transportation mode trans is a member of* $C_{mobile} = \{C_1, C_2, ..., C_n\}$. *Hence, the aim of a mobile sensing application is to infer A and each member in* $C_{mobile}$ *simultaneously, assuming there are dependencies between the instances of A, and members in* $C_{mobile}$.

## 3.4 Mobile Context Recognition System

In this chapter, our mobile context recognition system is used to tackle the problem of simultaneous inference of human activities, transportation modes (including their environments) of mobile users in their daily life. Therefore, we designed this system to be generic for mobile context recognition purposes. In overall, this system is composed of two overarching sub-modules (refer to Figure 3.2): contextual modelling and multi-contexts inference.

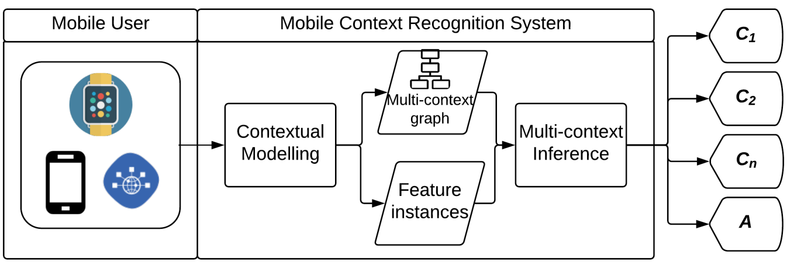



Figure 3.2: General overview of mobile context recognition system.

In the first module, the process of constructing structured hierarchical contexts is performed using the proposed modelling approach (defined in Section 3.4.1) by generalising, decomposing and extracting the contexts associated with human activities in a mobile sensing application. The second module is dedicated to a simultaneous inference of the human activity and related contexts of the mobile user.

In the contextual modelling module (refer to Figure 3.3), there are two major components associated: 1) feature construction and extraction, and 2) modelling human activities, transportation modes and their related contexts in a hierarchical structure. The first component relates to typical process to produce important features of a segment, which is related to a raw user annotation *a*. The second component will be elaborated in Section 3.4.1.

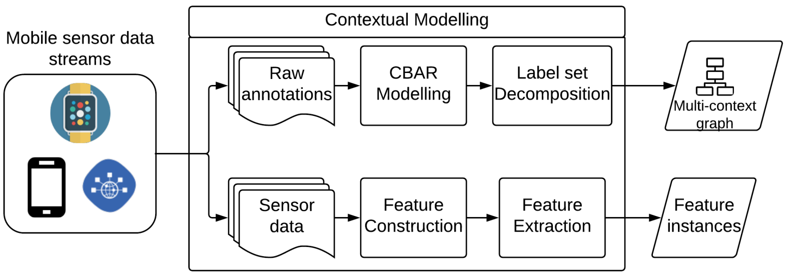



Figure 3.3: Process workflow in contextual modelling component of mobile context recognition system.

### 3.4.1 Context-based Activity Recognition (CBAR) Modelling

Raw annotations can be noisy and inconsistent in a diary-based study, especially the ones collected through Experience Sampling Method (ESM) [Csikszentmihalyi and Larson, 2014]. In this section, context-based activity recognition (CBAR) modelling is presented as a conceptual modelling approach that can be used to interlink the raw annotations to human activities and mobile contexts defined in Definition 2. In other cases, entity recognition [Schweizer and Schmidt, 2014] can be performed on textual data of the raw user annotation, which can then be automatically associated to a relevant human activity and its mobile contexts (including transportation mode).

Since the human activities are being contained in a space, there are several common properties that can characterise the transportation modes and environments of the mobile users. In this chapter, we refer these as the **properties of sensing space**, which consist of:

1. **Mobility of the sensing space.** For instance, the environment can move in spatial and temporal domains. In other words, the mobile users will experience dynamic mobility extrapolated by their sensing space (environment) by moving from one location to another location. In this chapter, the nature of a sensing space can defined whether it is: 1) moving or 2) stationary.

2. **Motorisation of the sensing space.** In this case, the environment (space) where mobile users are contained in can be characterised whether it is motorised or not. Consequently, further categorisations (e.g., vehicle vs. non-vehicle) of the environments can be derived from this property.

As a result, the human activities can be modelled based on the above properties, which could produce a significant distinction of where the mobile users are contained in (including the environmental contexts associated to the contained space). It should be noted that the above properties can be expanded further to characterise human activities in the wild. In this chapter, these properties are presented in our initial framework for the purpose of multi-context recognition of mobile users.

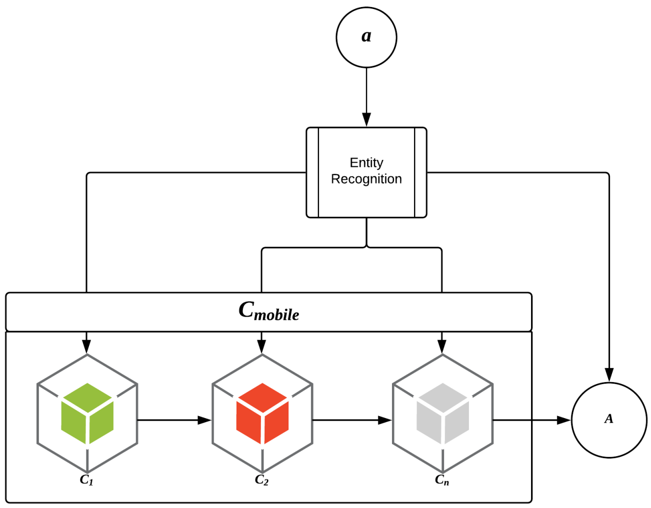



Figure 3.4: CBAR Modelling from raw user annotation.

As shown in Figure 3.4, a given raw user annotation $a$ can be decomposed through a pre-defined entity recognition process. The objective of this conceptual process is to first derive relevant transportation modes and human activities from all user annotations. In $C_{mobile}$ set, the last item (i.e., $C_n$) should correspond to the transportation mode (i.e., $C_n = trans$) where a human activity $A$ is performed in. Consequently, the commonalities (based on **properties of sensing spaces**) of all transportation modes in the system can be derived into a sequence of $C_1$ to $C_{n-1}$. Since there is an assumption of dependencies between all mobile contexts and human activities, the modelling can be expressed as: $C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_n \rightarrow A$.

For a typical implementation of simultaneous recognition of all contextual labels $(C_1, C_2, ..., C_n, A)$, a simplistic approach is to build an independent classifier per each member of the contextual labels. This typical approach is used as the baseline in our experiment.

Based on our conceptual modelling approach, there are two approaches to address the simultaneous inference issue:

1. Multi-stage inference from multiple classifiers that are modelled hierarchically. In other words, inference is made from $C_1$ sequentially to $C_n$ (transportation mode), then $A$.

2. Inference from a multi-target classifier. In particular, multi-target classification [Last, 2016] is a special case of multi-label classification in terms of modelling the dependencies between target classes from multiple label sets.

As a result of CBAR modelling process in our system, all raw annotations can be then be forwarded to the label set decomposition process, where multi-context graph (schema) is produced as an output for the next phase (models construction in multi-context inference module of our mobile context recognition system).

### 3.4.2 Multi-context Inference of Transportation Mode and Human Activity

Given the multi-context graph (schema) and feature instances produced from the previous module, various machine learning models can be constructed (refer to Figure 3.5).

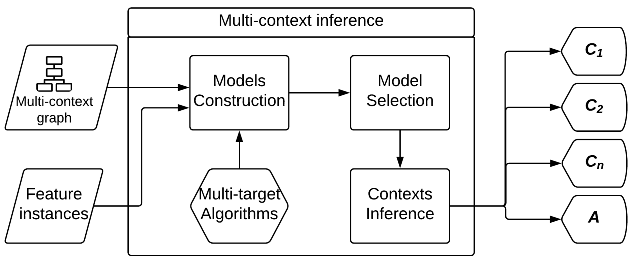



Figure 3.5: Process workflow in multi-context inference of mobile context recognition system.

The construction of models are based on two approaches that have been previously mentioned in Section 3.4.1. The first approach refers to construction of a classifier set, where inferred context (output) becomes the input for next inference stage. In other words, the classifiers are modelled in a hierarchical structure until the final inference is achieved. The second approach of constructing a model is by training a classifier based on multi-target algorithms. Each model is constructed based on the input of training set (from feature instances). Ultimately, contexts inference sub-component takes the best model to perform simultaneous recognition of $C_1, C_2, ..., C_n$ and $A$. Consequently, the best model is selected through an internal evaluation process in our system based on appropriate metrics (e.g., the proportion of exact match).

## 3.5 Experiments and Evaluation

### 3.5.1 CrowdSignals Dataset

In order to validate our mobile context recognition system, CrowdSignals dataset [Welbourne and Tapia, 2014] is used in the experiment. This dataset consists of rich crowdsourced mobile sensor data from smartphones and wearable devices through in-the-wild data collection campaign. Hence, each annotation is labelled by real participants in their daily life. In particular, we leverage the time-interval labels (refer to the ground truth annotations of scripted behaviour) in this dataset, due to the presence of exact start and end time of specific and ongoing activities, events and situations. Although this dataset consists of the lifelog mobile sensor data from more than 30 participants, we use the Android smartphone data of five representative participants for the experiment and evaluation in this chapter. From these five mobile users, we identified the following unique raw annotations:

1. Walking
2. Riding in a car
3. Bus riding
4. Playing video game
5. Stairs
6. Light rail riding
7. Escalator
8. Elevator
9. Drinking
10. Riding scooter

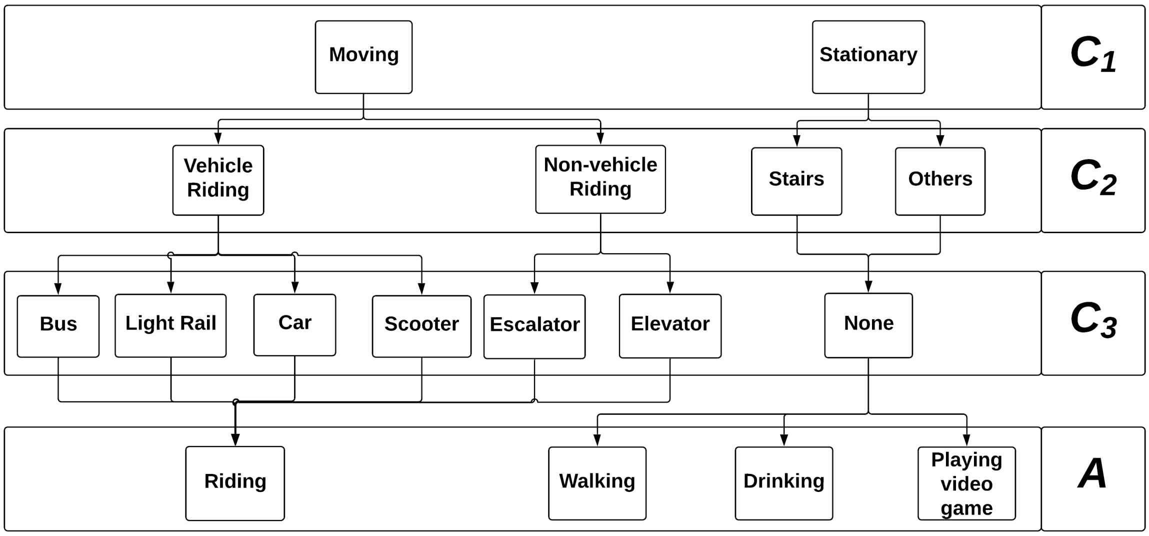



Figure 3.6: Multi-context graph from CBAR modelling of CrowdSignals dataset.

For a typical solution of activity recognition, these ten annotations can be used directly to classify the human activity based on single-label-classification mechanism. However, it should be noted that from the recognised human activities, their transportation modes need to be inferred further. Through the CBAR modelling approach, we can decompose these annotations to three different mobile context label sets $C_{mobile} = \{C_1, C_2, C_3\}$ and an activity label set $A$ as shown in Figure 3.6. In other words, each raw annotation $a$ above will be associated to all decomposed contextual labels (i.e., $a = \{C_1, C_2, C_3, A\}$). In this case, the ultimate aim of a mobile sensing application is to infer the all decomposed labels accurately.

Although the decomposition of a raw user annotation can be performed through entity recognition, the schema graph (refer to Figure 3.6) is manually defined in our experiment. Hence, a mobile context recognition system should be able to perform an accurate simultaneous inference of mobile user contexts from the following label sets derived from CrowdSignals dataset:

1. ***Environment type of transportation mode ($C_1$)*** label set:

$$C_1 = \Big\{\text{Moving, Stationary}\Big\}$$

2. ***Type of sensing space ($C_2$)*** label set:

$$C_2 = \left\{\begin{array}{l}\text{Vehicle Riding, Non-vehicle Riding,}\\ \text{Stairs, Others}\end{array}\right\}$$

3. ***Transportation mode*** *($C_3$)* label set:

$$C_3 = \begin{Bmatrix} \text{Bus, Light Rail,} \\ \text{Car, Scooter,} \\ \text{Escalator, Elevator} \end{Bmatrix}$$

4. ***Human activity*** *(A)* label set:

$$A = \begin{Bmatrix} \text{Riding,} \\ \text{Walking,} \\ \text{Drinking,} \\ \text{Playing video game} \end{Bmatrix}$$

Table 3.1: Activity segments and their annotations for sampled mobile users of CrowdSignals dataset.

| **User ID** | **Smartphone** | **Annotated segments** | **Annotations** |
|:---:|---:|---:|---:|
| **A** | Asus Zenfone 2 | 56 | 1, 2, 3, 4, 5 |
| **B** | Samsung Galaxy S7 | 20 | 1, 2, 3, 5, 6, 7, 8, 9 |
| **C** | Samsung Galaxy S7 Edge | 131 | 1, 3, 4, 5, 6, 7, 8, 9, 10 |
| **D** | Samsung Galaxy S6 Edge | 212 | 1, 2, 3, 4, 5, 6, 7, 8, 9 |
| **E** | Samsung SM-A800F | 325 | 1, 2, 3, 4, 5, 6, 7, 8, 9 |

In terms of the coverage of activity segments, the annotations are not always given by the participants at all time. This evidence is also shown in Table 3.1 and Figure 3.7.

As described in Table 3.1, the sampled mobile users have different types of Android smartphones. Consequently, the accuracy of multi-context inference could also be affected by the quality of mobile sensors assembled by the device manufacturers. Moreover, each smartphone model has different capabilities for mobile sensing due to availability of sensor streams and limited types of embedded mobile sensors. To build a generic model (i.e., person-independent model), we leverage the following sensor streams that exist for the five sampled mobile users:

1. Accelerometer
2. Magnetic field
3. Light
4. Screen status

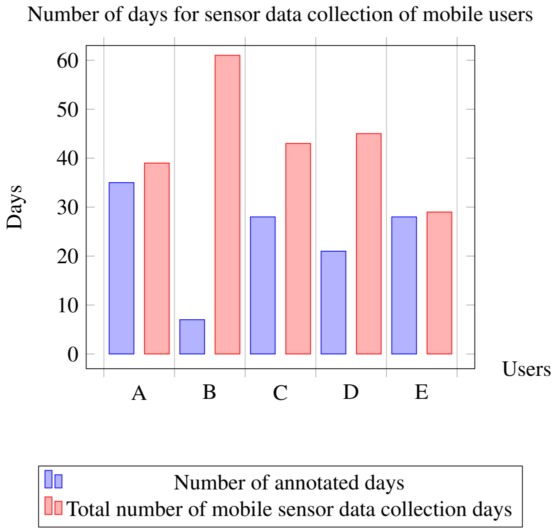



Figure 3.7: Number of annotated days and total days for sampled mobile users of CrowdSignals dataset

In the CrowdSignals dataset, these sensors may be sampled at different frequencies. In particular, accelerometer, magnetic field and light sensors are sampled with **SENSOR_DELAY_FASTEST** setting during the mobile data collection on Android smartphones. This setting corresponds to the default setting of the associated smartphone model for reading the sensor data as fast as possible. Hence, the frequency might also be affected by the quality of sensors embedded in these smartphones. For instance, an accelerometer sensor may be sampled at 20Hz for a particular smartphone while another model would be sampled at 50Hz. For the screen status, the smartphone listens to the callback events when the screen is on or off. Consequently, the granularity of the data used in our experiment is shown in Table 3.2 in terms of counts of sensor data points.

Table 3.2: Granularity (count) of raw sensor data points from Android smartphones.

| **User ID** | **Accelerometer** | **Magnetic field** | **Light** | **Screen status** |
|---|---:|---:|---:|---:|
| **A** | 10,270,550 | 10,149,507 | 51,896 | 371 |
| **B** | 1,044,436 | 247,243 | 24,981 | 49 |
| **C** | 17,318,445 | 4,786,078 | 447,907 | 460 |
| **D** | 22,554,947 | 9,627,477 | 521,762 | 1,214 |
| **E** | 40,040,845 | 20,814,032 | 40,755,734 | 1,664 |

### 3.5.2 Feature Construction and Extraction

For each annotated segment, we applied time-interval based temporal segmentation (1 second window size $\tau$ with 50% overlap) and extract the features for each sensor stream. In this case, we extracted the statistical features (mean, median, maximum, minimum, standard deviation, interquartile range, root-mean-square) from each window of sensor streams. In other words, a statistical feature is computed via a given function for all feature values in a window, bounded by $\tau$.

For accelerometer and magnetic field sensor streams, the features are extracted from the magnitude value computed from tri-axial sensor readings. In other words, the magnitudes for accelerometer $mg_{acc}$ and magnetic field $mg_{mag}$ are calculated as follows:

1. $$mg_{acc} := \sqrt{x_{acc}^2 + y_{acc}^2 + z_{acc}^2}$$

   where $x_{acc}$, $y_{acc}$, $z_{acc}$ are the tri-axial sensor values of phone's acceleration.
2. $$mg_{mag} := \sqrt{x_{mag}^2 + y_{mag}^2 + z_{mag}^2}$$

   where $x_{mag}$, $y_{mag}$, $z_{mag}$ are the tri-axial sensor values of magnetic field measurement.

The purpose of computing and leveraging the magnitude values is to construct a more robust model, which is invariant towards smartphone's orientation [Yurtman and Barshan, 2017]. For light sensor stream, the statistical features are extracted from the raw reading of illuminance (measured in lx) within a given window. On the other hand, the screen status corresponds to whether the screen on the smartphone is on or off, which may indicate how users are actively engaged with their smartphones. Furthermore, it is known that there are distinct variations of accelerometer and magnetometer reading inside various indoor spaces [Susi et al., 2013]. Hence, it is justified that a robust and effective model could be build to infer the transportation mode of a mobile user at a given point in time. Unlike the systems such as [Reddy et al., 2010, Byon and Liang, 2014] that require GPS sensor to determine the transportation mode, our mobile context recognition system do not need to rely on the user's locations. In fact, sampling from the GPS sensor is known to have significant energy consumption compared to other sensors (also proven in the study by Hemminki et al. [Hemminki et al., 2013]). It should be noted that the gyroscope sensor stream is excluded during the modelling stage due to its unavailability for user E's smartphone model.

### 3.5.3 Experiment Setup

A pre-defined schema (refer to Figure 3.6) is used to map the predicted raw annotation to the associated environment type ($C_1$) of transportation mode, further categorisations ($C_2$) of $C_1$, transportation mode ($C_3$) and human activity ($A$).

Our experiment is composed of two distinct sets. The first set refers to inference made by independent classifiers for each contextual label (shown in Figure 3.8). In this case, the classifiers are independently trained based on decomposed label sets (defined in Section 3.5.1). The first experiment set is used as the baseline of a typical mobile sensing application.

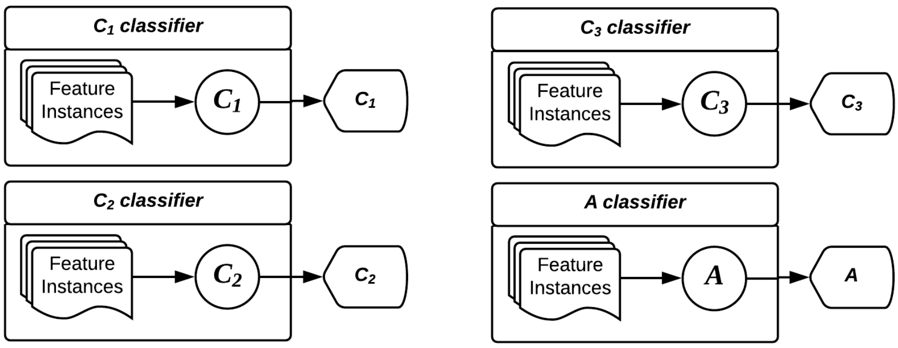



Figure 3.8: Inference from independent classifiers.

The second set of our experiment is associated with the inference of all contextual labels using our mobile context recognition system. We implemented two approaches of models construction of multi-context inference. The first one is referred as **multi-stage inference** where a classifier is trained on each target label set. Subsequently, the output of inference will be appended to the feature instances that are used the next stage of inference (refer to Figure 3.9). For instance, the training process of classifier for $C_2$ relies on feature instances (consist of feature vectors and ground-truth vector of $C_1$).

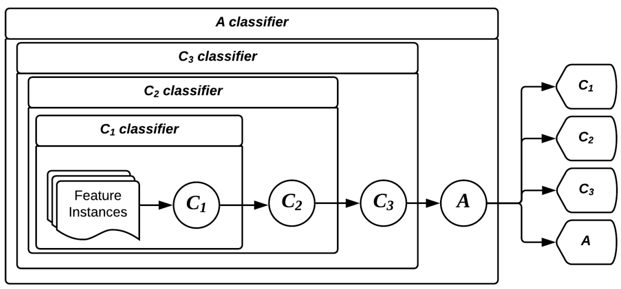



Figure 3.9: Multi-stage inference from classifiers.

The second approach of models construction is based on the pre-defined multi-target algorithms (refer to Figure 3.10). Hence, the multi-target classifiers are deployed and evaluated on the following algorithms:

1. Class Relevance (CR), which is the multi-target version of the Binary Relevance (BR) method for multi-label classification.
2. Classifier Chains (CC)
3. Nearest Set Replacement (NSR), which is the multi-target version of Pruned Set (PS) method for multi-label classification.
4. Ensemble of Classifier Chains of boosted classifiers (EN-CC-AdaBoost). For the boosting implementation, AdaBoost algorithm is used. The setup of EN-CC-AdaBoost could require significant memory resources in order to operate. This fact is also proven during our experiment.

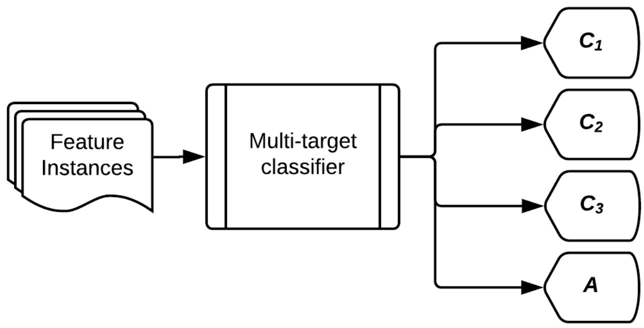



Figure 3.10: Multi-target inference from a classifier.

For both experiment sets, we leverage the following algorithms as the base classifiers:

1. Naïve Bayes (NB).
2. Support Vector Classifier (SVC).
3. Multilayer Perceptron (MLP).
4. Decision Tree (DT).
5. Random Forests (RF).
6. k-Nearest Neighbor with k=1 (1NN).

For the baseline (first experiment set) and multi-stage classification experiment (in our mobile context recognition system), we implemented the inferences using WEKA (version 3.8.1) data mining software library [Hall et al., 2009]. It contains the various functions needed for data exploration and data mining tasks. For base classifiers, the default configurations are used imperatively in WEKA. In this case, decision tree classifier is named as J48 in WEKA. Default parameter for k-Nearest Neighbor classifier is set where $k = 1$. In particular, we use the SMO algorithm [Keerthi et al., 2001] for SVC in WEKA. Furthermore, the maximum number of trees for Random Forests is set to 100.

For the multi-target inferences (in our mobile context recognition system), they are evaluated based on our experiment using MEKA (version 1.9.0) [Read, 2012–2015], a multi-label/multi-target extension to WEKA.

### 3.5.4 Evaluation

In order to validate the accuracy of multi-context inferences, 10-fold cross-validation was applied to all runs in our experiment.

As mentioned previously, the experiment is evaluated on following two experiment sets: inference from independent classifiers (baseline) and mobile contexts inference in our system (using both multi-stage and multi-target algorithms approaches in models construction process).

For the evaluation, we leverage exact match and accuracy metrics. Accuracy is computed for each inference from the corresponding contextual label set, while exact match measure refers to the proportion of correct prediction of all contextual labels (simultaneous inference of all contextual labels). Table 3.5 shows the overall performance for independent inference, given the base classifiers. Although the accuracy performance between independent inference and multi-stage inference is competitive, the exact match (simultaneous inference of mobile contexts) through multi-stage models shows a significant difference, proving that the dependencies between human activities and transportation mode (including related contexts) are not to be neglected for a reliable inference in a mobile sensing application.

Table 3.3: Performance of independent classifiers (baselines).

| **Metric** | **Baseline** | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 1.4% | 45.5% | 34.5% | 70.8% | **89.4%** | 88.7% |
| Accuracy: $C_1$ | 66.0% | 66.9% | 68.5% | 88.1% | **93.4%** | 92.1% |
| Accuracy: $C_2$ | 11.6% | 65.3% | 67.3% | 87.5% | **93.0%** | 91.6% |
| Accuracy: Transportation Mode ($C_3$) | 5.4% | 49.4% | 56.4% | 83.5% | **91.7%** | 89.3% |
| Accuracy: Activity | 16.8% | 67.0% | 69.2% | 87.5% | **93.0%** | 91.6% |

Table 3.4: Performance of Multi-stage models.

| Metric | Multi-stage classification | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 2.6% | 48.0% | 49.7% | 81.0% | **90.7%** | 88.7% |
| Accuracy: $C_1$ | 66.0% | 66.9% | 68.5% | 88.1% | **93.4%** | 92.1% |
| Accuracy: $C_2$ | 12.3% | 65.2% | 67.0% | 87.2% | **92.8%** | 91.6% |
| Accuracy: Transportation Mode ($C_3$) | 5.8% | 48.2% | 52.7% | 82.4% | **91.1%** | 89.3% |
| Accuracy: Activity | 22.4% | 66.8% | 65.9% | 87.1% | **93.1%** | 91.6% |

Table 3.5: Performance of Class Relevance (CR) models.

| Metric | Class Relevance | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 1.4% | 45.3% | 36.7% | 70.8% | **92.9%** | 88.7% |
| Accuracy: $C_1$ | 66.0% | 66.9% | 68.8% | 88.1% | **93.5%** | 92.1% |
| Accuracy: $C_2$ | 11.5% | 65.3% | 66.8% | 87.5% | **93.1%** | 91.6% |
| Accuracy: Transportation Mode ($C_3$) | 5.4% | 49.6% | 56.0% | 83.4% | **91.8%** | 89.3% |
| Accuracy: Activity | 16.8% | 67.0% | 68.5% | 87.5% | **93.2%** | 91.6% |

Table 3.6: Performance of Classifier Chains (CC) models.

| Metric | Classifier Chains | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 2.6% | 48.8% | 49.1% | 82.1% | **91.4%** | 88.7% |
| Accuracy: $C_1$ | 66.0% | 67.2% | 67.4% | 88.3% | **93.7%** | 92.1% |
| Accuracy: $C_2$ | 12.2% | 65.7% | 65.9% | 87.4% | **93.3%** | 91.6% |
| Accuracy: Transportation Mode ($C_3$) | 22.4% | 49.6% | 56.0% | 83.4% | **91.8%** | 89.3% |
| Accuracy: Activity | 5.8% | 66.6% | 61.2% | 87.3% | **93.4%** | 91.6% |

From the performance comparison in Table 3.3, 3.4, 3.5, 3.6 and 3.7, we can conclude that best base classifier is dedicated to Random Forests in most cases (except the performance in Table 3.8), thus maintaining itself as the state-of-art algorithm for a real mobile sensing application. Although the basic tree based classifier (i.e., Decision Tree) has the best performance in Table 3.8, its overall performance is yet lower than our proposed multi-stage inference models (Table 3.4). Hence, this result suggests that the applications of ensemble and boosting on multi-target algorithms may not increase the performance of a mobile context recognition system. However, this approach can still be used for model selection (in

Table 3.7: Performance of NSR models.

| Metric | NSR | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 11.4% | 49.5% | 50.4% | 82.4% | **91.6%** | 88.9% |
| Accuracy: $C_1$ | 38.1% | 68.1% | 69.0% | 88.2% | **93.8%** | 92.4% |
| Accuracy: $C_2$ | 32.7% | 66.5% | 67.4% | 87.3% | **93.4%** | 91.9% |
| Accuracy: Transportation Mode ($C_3$) | 27.3% | 49.8% | 56.2% | 83.3% | **91.8%** | 89.6% |
| Accuracy: Activity | 22.5% | 67.9% | 63.8% | 87.4% | **93.6%** | 91.9% |

Table 3.8: Performance of mobile context recognition system using En-CC-AdaBoost.

| Metric | EN-CC-AdaBoost | | | | | |
|---|---|---|---|---|---|---|
| | **NB** | **SVC** | **MLP** | **DT** | **RF** | **1NN** |
| Exact Match | 1.6% | 48.7% | 50.1% | **90.1%** | 89.8% | 86.9% |
| Accuracy: $C_1$ | 65.9% | 67.4% | 70.9% | **93.1%** | 92.9% | 91.3% |
| Accuracy: $C_2$ | 10.5% | 65.8% | 69.4% | **92.7%** | 92.4% | 90.7% |
| Accuracy: Transportation Mode ($C_3$) | 4.9% | 49.2% | 57.1% | **90.8%** | 90.4% | 87.9% |
| Accuracy: Activity | 18.0% | 67% | 67.5% | **92.8%** | 92.5% | 90.6% |

the internal evaluation process), in case if it outperforms other inference approaches. Inherently, the applications of both ensembles and boosting consume significant resources (in terms of training time, memory and space requirements).

Through our mobile context recognition system, its performance for the exact match is increased by 2.2% (from the baseline), which is ultimately achieved by a multi-target inference approach of NSR algorithm on RF classifier. Although the exact match measurement of NSR is lower than CR algorithm, we selected NSR as the best model due to better independent prediction result for $C_1$, $C_2$, transportation mode and human activity. If those criteria were not considered, the increased exact match would be 3.5% over the baseline for RF classifier. On the other hand, the most significant increase of 19.3% can be noticed (for exact match in Table 3.10) through the usage of ensemble, multi-target classifier chains and boosting strategy on decision tree (as base classifier). Furthermore, the significant improvement is also shown for the accuracy of each contextual label (e.g., inference of transportation mode depicted in Table 3.9).

By being able to recognise multiple contexts of mobile users simultaneously, many interactive and ubiquitous applications can be enabled. Let us consider the scenario of a smart device to be able to recognise the user contexts in daily life ubiquitously. Given the accurate simultaneous recognition of

Table 3.9: Accuracy of transportation mode inference for baseline and mobile context recognition system.

| System | NB | SVC | MLP | DT | RF | 1NN |
|---|---|---|---|---|---|---|
| Baseline | 5.4% | 49.4% | **56.4%** | 83.4% | 91.7% | 89.3% |
| Proposed system | **27.3%** (NSR) | **49.8%** (NSR) | 56.0% (CR, CC) | **90.8%** (EN-CC-AdaBoost) | **91.8%** (NSR) | **89.6%** (NSR) |

Table 3.10: Exact match for baseline and mobile context recognition system.

| System | NB | SVC | MLP | DT | RF | 1NN |
|---|---|---|---|---|---|---|
| Baseline | 1.4% | 45.6% | 34.5% | 70.8% | 89.4% | 88.7% |
| Proposed system | **11.4%** (NSR) | **49.5%** (NSR) | **50.4%** (NSR) | **90.1%** (EN-CC-AdaBoost) | **91.6%** (NSR) | **88.9%** (NSR) |

both user environments and activities, it can be used by the assistive technologies to improve the quality of life for these mobile users. For example, intelligent reminders of user activities and notifications for major transportation delays due to the current situation of the users. This outcome can also be leveraged for the applications of discovering user routines [Sadri et al., 2017, Rahaman et al., 2017a, 2018] based on personal contexts of mobile users. In an intelligent healthcare scenario, a robust and simultaneous recognition of multiple user contexts would be important to be considered for elderly and disabled people, while travelling through various accessible paths [Rahaman et al., 2017b]. In a large and pervasive sensing scenario (e.g., [Liono et al., 2018b]), mobile context recognition would be beneficial for crowdsensing from smart devices (e.g., Internet of Things) if the model can be used for robust and simultaneous user contexts recognition. In smart city applications, simultaneous contexts recognition can be used to enable smarter citizen violation prediction based on real-time inference from multiple sources, such as parking violation prediction [Shao et al., 2018] that can be improved by considering the ubiquitous contexts from both parking officer and IoT devices on the parking spaces. In the case of analysing the space utilisation (e.g., [Arief-Ang et al., 2016, 2017, 2018a,b]), simultaneous recognition of contexts from multiple rooms can be leveraged for better occupancy counting and recommendation of places to occupy, also by considering the dynamic environment and personal user contexts. In a smart home environment, intelligent electricity consumption prediction (e.g., [Song et al., 2017]) can be improved by considering multiple contexts recognition of tenant activities.

## 3.6 Conclusion

In this chapter, a mobile context recognition system is presented for simultaneous inference of transportation modes and human activities, including the environmental contexts of transportation modes. Our system and study aim to enable smarter mobile sensing applications by being able to provide intelligent assistance from the accurate inference of related mobility contexts associated with human activities in daily life. Through our hierarchical and contextual modelling approach (combined with multi-target algorithms), our system is capable of outperforming the traditional approach of independent inference for multiple contexts of mobile users. Through the exact match of all contextual labels of the mobile users, 19.3% improvement is noticeable for a solution with decision tree as the base classifier. On the other hand, the best performance of exact match is acquired by random forests based solution in multi-context recognition by far.

Furthermore, our system is robust towards the intelligent inference for the data streamed from low-energy sensors, which reduces the overall reliance on location-based sensors. Nevertheless, many challenges should be addressed to improve our system further. For example, one of the immediate challenges in real-world is related to coping context recognition with subjective and in-situ user annotations (e.g., in a mobile crowdsensing scenario). Given that multiple contexts of mobile users can be sensed seamlessly and accurately, it provides a greater avenue for situation inference from mobile sensing perspective, which also enables the future intelligent mobile applications to be more context-aware in providing proactive assistance for their users.

# Chapter 4

# Predicting User Annotation in Interactive Mobile Data Sensing and Collection

## 4.1 Introduction

The inherent noise that is related to sensor data and human annotations heavily influences the accuracy of context recognition. Moreover, human annotations may include contextual information on their activities and environments. The acquisition of human annotations is a non-trivial task in mobile sensor data collection, which is typically a crucial step before every mobile sensing experiment. In this chapter, we provide the avenue to improve the interactivity of annotation acquisition process of mobile sensing by performing intelligent user-driven annotation prediction.

The Experience Sampling Method (ESM) [Csikszentmihalyi and Larson, 2014] 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 [Fuller-Tyszkiewicz et al., 2015]. However, it has offered significant benefits for ubiquitous computing research in recent years, for example, emotion recognition [Ghosh et al., 2015], mobile user intimacy and smartphone usage [Gustarini et al., 2016, van Berkel et al., 2016], human activity recognition [Bao and Intille, 2004], lifelogging [Gurrin et al., 2014, Li et al., 2010, Gouveia and Karapanos, 2013, Atz, 2013] and mobile sensor data collection [Welbourne and Tapia, 2014, Vaizman et al., 2018a,b, Berkel et al., 2017].

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 [Chen et al., 2006] and human activity recognition [Junker et al., 2008, Cornacchia et al., 2017].

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 [Gershuny, 2004]. To perform a specific task in daily life, retrospective memory [Brewer et al., 2017] 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 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.

## 4.2 Related Work

**Experience Sampling Method**. The ESM is a prevalent approach used in many domains [Bao and Intille, 2004, Kahneman et al., 2004, Froehlich et al., 2007, Krueger and Schkade, 2008, Froehlich et al., 2009] to recall recent or past activities of a user. Its reliability and validity have been empirically studied in [Csikszentmihalyi and Larson, 2014], 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 [Csikszentmihalyi and Larson, 2014], 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 [Ferreira et al., 2014] and break-points between a user's activities [Obuchi et al., 2016]. According to [Ghosh et al., 2015], the experience sampling could be triggered for the mobile users from the signal, event or time (at regular intervals). Moreover, the users in [Bao and Intille, 2004] 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 [Lane et al., 2008] 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 chapter, 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 [Minor and Cook, 2017] 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 [Zhou and Xu, 2007], 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 [Blum and Mitchell, 1998], 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 [Guan et al., 2007]. 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 [Jaques et al., 2015], 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 [Tan et al., 2014] 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 [Hossain et al., 2017], 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.

## 4.3 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?*

### 4.3.1 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 $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_n\}$ be the set of sensors available during the collection of data on a mobile user, where $i$ is the index of $i$-th sensor, $\boldsymbol{1} \leq \boldsymbol{i} \leq \boldsymbol{n}$ 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 $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a_d}\}$, where $\boldsymbol{d}$ is the number of unique experience sampling labels $\boldsymbol{a}$ in $\boldsymbol{A}$.

Let $\boldsymbol{S}_{ia} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s_m}\}$ 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., $\boldsymbol{j} = \boldsymbol{m}$), where $j$ is the index of $j$-th instance in the observed time series $\boldsymbol{S}_{ia}$, $\boldsymbol{1} \leq \boldsymbol{j} \leq \boldsymbol{m}$ and $\boldsymbol{m}$ is the length of $\boldsymbol{S}_{ia}$ within a certain time-interval boundary $\boldsymbol{t}_\Delta \leq \boldsymbol{t}_\delta$ 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 4.1). 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., $\boldsymbol{t}_\Delta = \boldsymbol{t}_\delta$). However, the duration of 'Light Rail Riding' is less than 30 minutes (i.e., $\boldsymbol{t}_\Delta = \boldsymbol{t}_\delta - \boldsymbol{z}$ ) 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 $t_\delta$ 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.

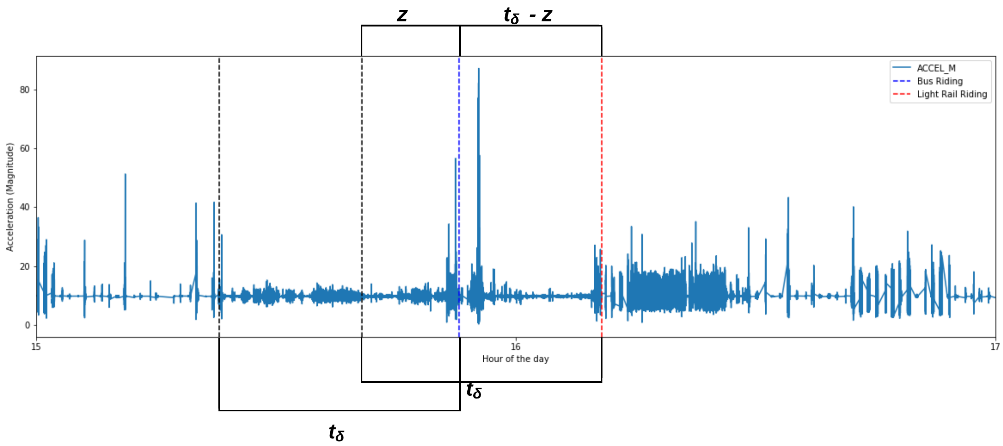



Figure 4.1: Point-based experience sampling label problem for user annotations. Blue ('Bus Riding') and red ('Light Rail Riding' dashed lines are the ESM data points (i.e., point-based experience sampling labels).

Therefore, the problem of annotation prediction is formulated as follows: given an unlabelled time series set for all sensors $\boldsymbol{S_u} = \{\boldsymbol{S}_{1u}, \boldsymbol{S}_{2u}, \ldots, \boldsymbol{S}_{nu}\}$ 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 $\boldsymbol{A}$).

### 4.3.2 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 4.2. 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 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).

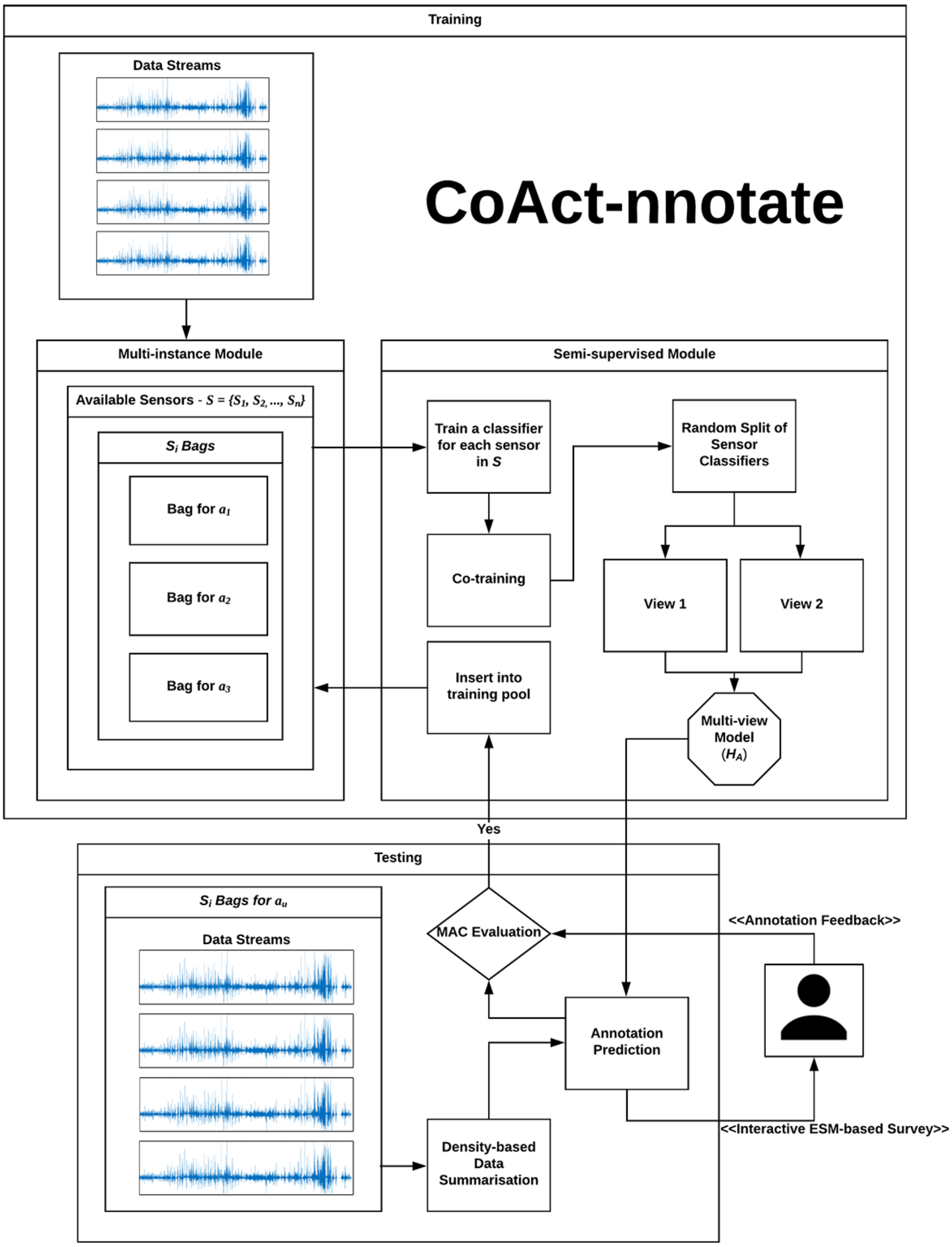



Figure 4.2: **CoAct-nnotate**: user-driven annotation prediction framework for mobile experience sampling labels.

### 4.3.3 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 4.3, 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.

A sensor feature bag for label $a$ is represented as $S_{ia}$ throughout this chapter. 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 $t_{\Delta}$.

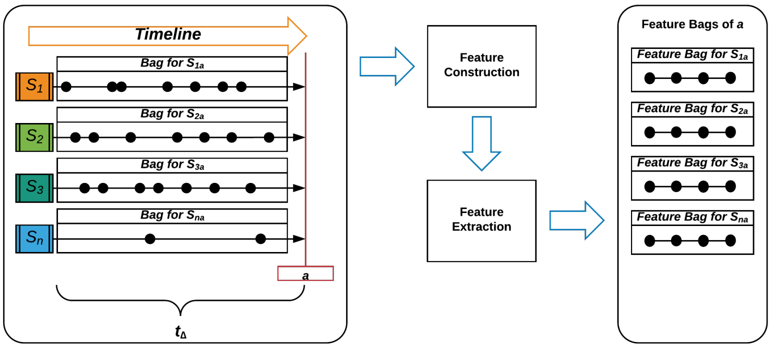



Figure 4.3: Workflow of multi-instance learning of instance streaming from multiple mobile sensors.

In this chapter, 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 4.3). Hence, we define the MIL prediction problem as follows:

Let us define unlabelled feature bags $S_u$, where $\boldsymbol{Su} = \{\boldsymbol{S}_{1u}, \boldsymbol{S}_{2u}, \ldots, \boldsymbol{S}_{nu}\}$, $i$ is the index of $i$-th feature bag for $i$-th sensor, $\boldsymbol{1 \leq i \leq n}$ 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 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_l\}$, $k$ is the index of $k$-th feature instance in a bag, $\boldsymbol{1 \leq k \leq l}$ 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 $\boldsymbol{H_A} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \ldots, \boldsymbol{H}_n\}$, $i$ is the index of $i$-th classifier for $i$-th sensor, $\boldsymbol{1 \leq i \leq n}$ 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 $t_\Delta$.

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 chapter. 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.

### 4.3.4 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 4.2). In theory, co-training (also known as co-regularisation) [Yu et al., 2011] 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 4.2).

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.

### 4.3.5 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 [Liono et al., 2018b], 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.

### 4.3.6 Multi-view Annotation Prediction

Once the sets of summarised bags ($summarisedBags_{first}$ and $summarisedBags_{second}$) are acquired through the process in Section 4.3.5, 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 $\beta$ 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:

$$macEvaluated = M_A \cdot ceil\left(\frac{ceil(\beta - Pr(y_{agreed}|X)) + D_A}{2}\right) \tag{4.1}$$

where $M_A$ is the binary value indicating a mutual agreement occurrence (i.e., $M_A \in \{0,1\}$), $\beta$ 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 4.3.7 below).

### 4.3.7 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 4.3.6). 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., $summarisedBags$), up-sampling (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 up-sampling 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 $\delta$ (i.e., Poisson($\delta$)). The method of upsampling is not restricted to instance replication because other forms, such as generative approaches of sampling (also known as generative oversampling [Liu et al., 2007, Das et al., 2015]), 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.

## 4.4 Experimental Evaluation

### 4.4.1 Dataset

We use the CrowdSignals dataset [Welbourne and Tapia, 2014] 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 $t_\delta$ is set to 30 minutes. Thus, each sensor bag in the MIL phase contains at least the data points within the duration of $t_\Delta$. 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*.

### 4.4.2 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 [Birant and Kut, 2007, Shao et al., 2016] and [Liono et al., 2018b] 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 4.1 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 [Pedregosa et al., 2011]):

- Naive Bayes (**NB**)
- Support Vector Classifiers (**SVC**)
- Multilayer Perceptron (**MLP**) with 0.00001 as the L2 penalty (regularisation parameter), L-BFGS [Andrew and Gao, 2007] 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 $TrainingPool$.
- 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 each time window, statistical features are extracted, such as *mean, median, maximum, minimum, standard deviation, interquartile range* and *root mean square*. 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:

$$Correctness = \frac{\sum_{u=1}^{v} annotation^{u}_{correct}}{v} \tag{4.2}$$

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.

### 4.4.3 Results

As shown in Figure 4.4, we leverage nine sensor channels (mentioned in Section 4.4.1) 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 $t_\delta$ (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 4.4, the time lengths of data points are varied with fewer outliers, as shown in Figure 4.5. 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 $S_{ia}$.

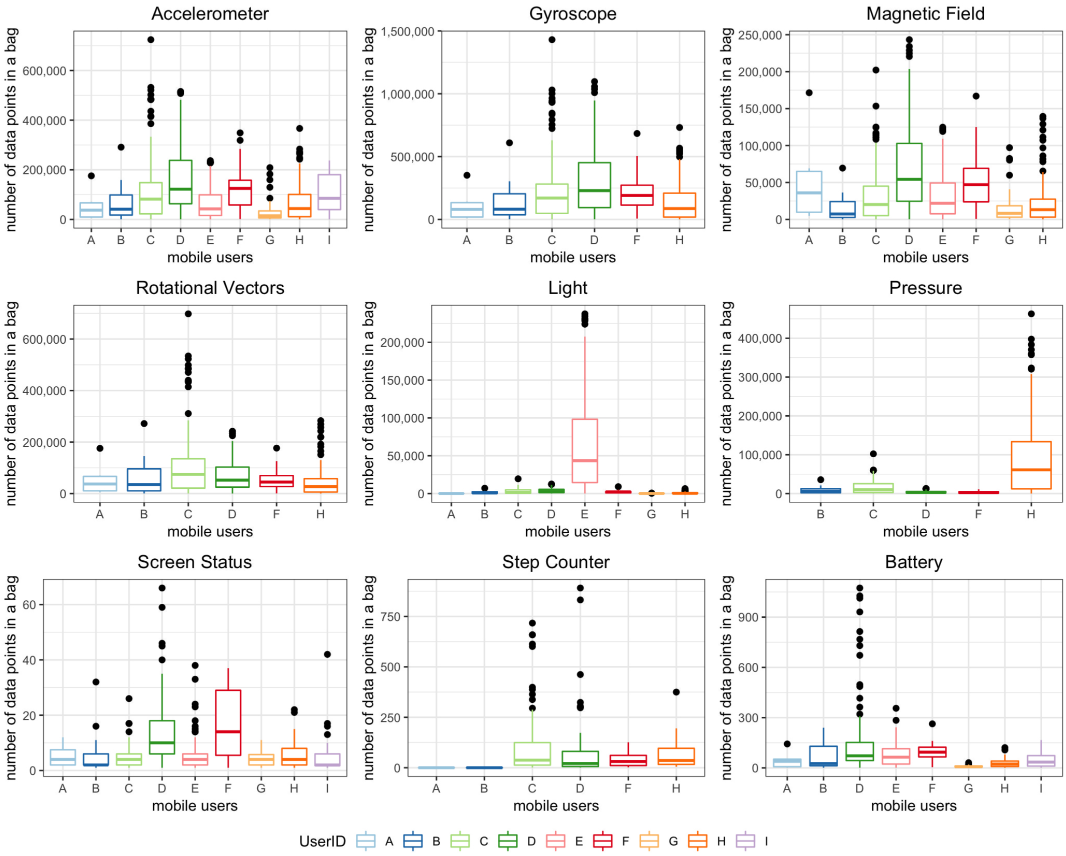



Figure 4.4: The granularity of data points in sensor bags.

Table 4.1: 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** |

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 4.1, 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 classifier

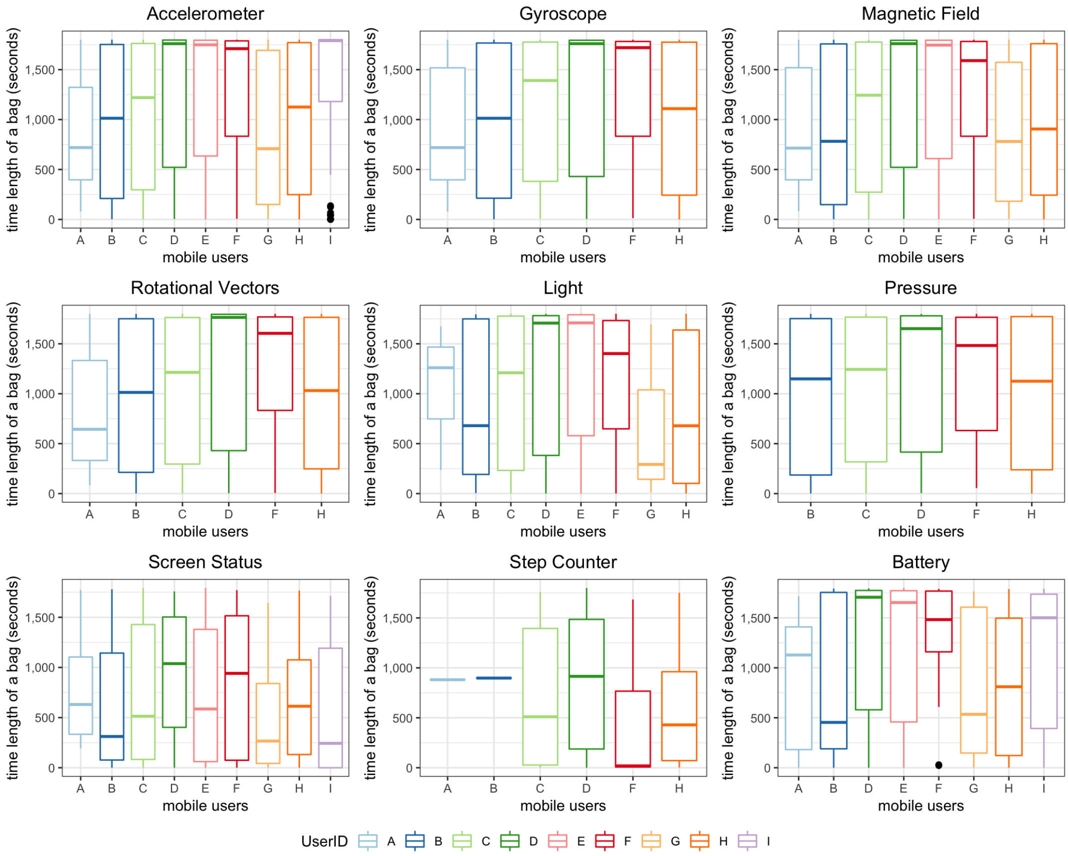



Figure 4.5: The time length of data points in sensor bags (seconds).

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 4.6).

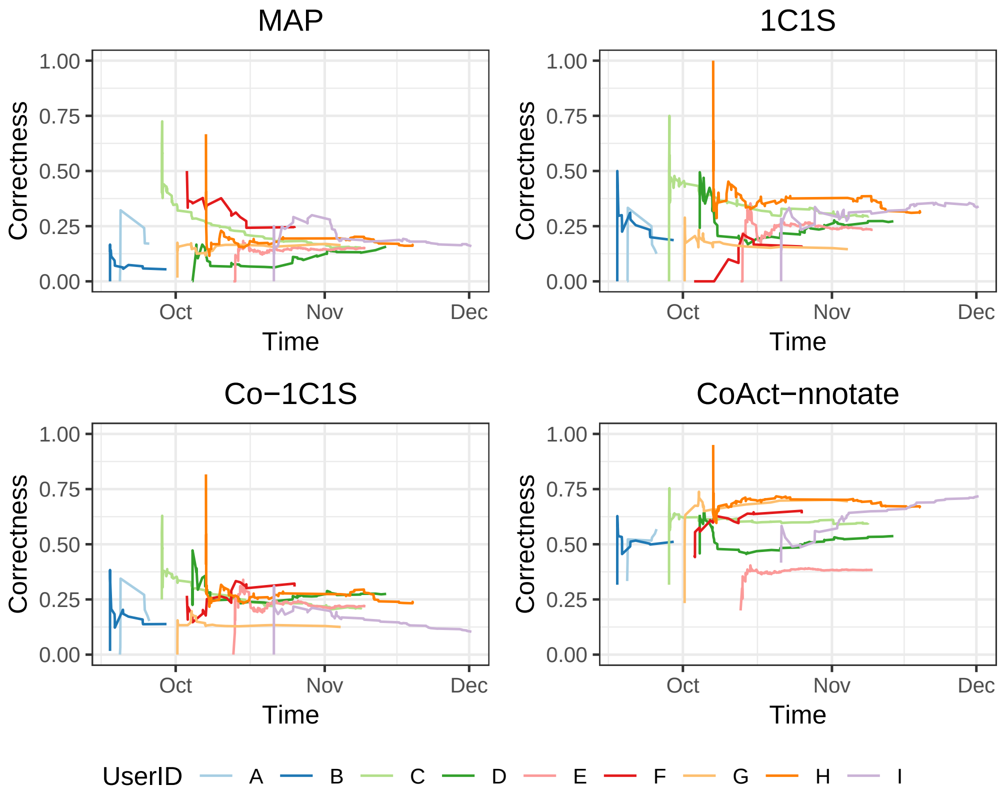



Figure 4.6: The average progression of correctness over the time for sampled users.

As shown in Figure 4.6, 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 4.7. 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 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 chapter 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 [Liono et al., 2018b] 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 chapter is to improve the model of annotation prediction over time by using both concepts of co-training and active learning.

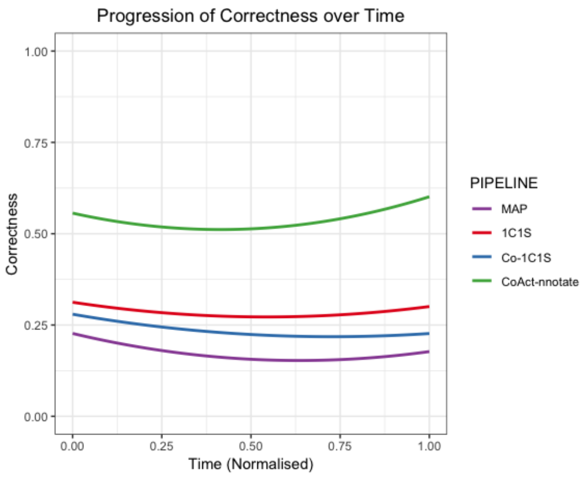



Figure 4.7: The progression of correctness over the time for MAP, 1C1S, Co-1C1S and CoAct-nnotate.

## 4.5 Conclusion

This chapter 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 chapter 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 [Liono et al., 2019b, Trippas et al., 2019] 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.

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. By having mobile devices to be more context-aware, annotations can be acquired seamlessly for the purpose of situation inference on intelligent mobile sensing applications. As previously mentioned in Chapter 3, these raw annotations can be composed of various contexts that are relevant to the users. Once a system is capable of performing an accurate context recognition, they can be then utilised for situation inference. Thus, the next chapter presents the intelligent task recognition based on the modelling of cyber, physical and social activities of mobile users in daily mobile sensing.

# Chapter 5

# Recognising Tasks of Mobile Users from Continuous Contextual Signals

## 5.1 Introduction

Understanding the user's situation is paramount to enable evidence-based insights for daily decision-making purposes. Previously, we have covered the prediction of in-situ annotations based on activities and recognition of their decomposed contexts in mobile sensing environments. Leveraging the contextual information of related mobile users is crucial in order to make sense of their actual situation at a given time in-the-wild. In this chapter, we focus on recognising the tasks of mobile users based on the cyber, physical and social signals that are derived from multi-sourced sensor data.

The rise of digital assistants in recent years is evidenced by the popular usage of voice assistants such as Siri or Alexa, and the growing uptake of smart speakers such as Google Home or Amazon Echo. These voice assistants are not only embedded in computing devices such as smartphones, tablets, wearables, laptops, desktops, but also becoming more ubiquitous. From the home to the office, the kitchen and to the meeting room, everyday appliances, like fridge, television, microwave, and conference facilities, have become voice-assistant enabled.

Existing AI-powered assistants, however, still fall short of its ultimate vision to truly complement, support, and empower users in their daily lives. Recent findings on a large user study of smart-speaker assistants by Bentley et al. [2018] highlight the main user interactions with these voice assistants mainly are to issue basic commands for specific domain applications, such as queries for music mainly for smart speakers, and commands related to communication or messaging applications on smartphone-based assistants.

In order to truly enable a richer user interaction with digital assistants, it is essential to enable these assistants to serve wider range of user queries while they are performing specific tasks. Supporting task progression and completion is the last mile in search interactions [White, 2018], and more generally in supporting digital assistant applications. Characterising and modelling tasks are the first steps to enable this support.

Let us imagine a future personal digital assistant that can identify and track our tasks ubiquitously, whenever and wherever it is required based on mobile user contexts. Such intelligent applications would be beneficial not only to track human tasks in daily life but also for assisting users to complete the tasks, providing recommendations of actions to improve the work productivity and the overall life experience and well-being.

If tasks can be recognised and tracked, support systems, such as digital assistants, recommender systems or search engines, can be adapted to better help humans complete their tasks. An intelligent assistant could then monitor the progress of a task, understand when a task is complete, or even encourage someone to switch from their current task to one that is currently critical.

In the field of psychology, the terminology of task was first defined by Leont'ev in 1978 [Leont'ev, 1978]. Based on the literature of activity theory mentioned by Bedny and Meister [2014], Leont'ev [1978] defined a task as *"a situation requiring achievement of a goal in specific conditions"*. Bedny and Karwowski [2006] states that a goal is a *"conscious mental representation of humans' own activity in conjunction with a motive"*. In other words, humans perform specific activities in order to progress and complete a task. We form a hypothesis, therefore, that we can better recognise a task when the underlying activities to achieve a task can be inferred.

Recognising human tasks in daily life is non-trivial. While the research into Human Activity Recognition (HAR) in the Ubiquitous Computing community has matured in recent years [Sigg et al., 2014, Khalifa et al., 2015, Liono et al., 2016, Abdallah et al., 2018], there has been little to no research on using intelligent sensing applications to recognise human tasks in daily life. Existing research mainly focuses on specific activities to be recognised, such as to identify the current or next app usage or email activities from weblogs; or the well-established study of HAR for simple and more complex locomotive activities (e.g., from walking, running, climbing stairs to travelling in cars, buses, trains, etc.). Research into tasks in Information Retrieval has been mainly used to improve search engine performance, as the knowledge of the task is used as a context to provide better results to user queries [Jones and Klinkner, 2008, Hua et al., 2013, Li et al., 2014, 2016]. The main challenges in recognising human tasks in daily life from a ubiquitous sensing perspective are due to the noisy environment and the dynamics of human activities. The immediate challenge for modelling mobile user behaviours come from the user annotations that are acquired from different users with a wide range of tasks and professions, and multiple the underlying contexts. This, if not addressed, can significantly impact the performance of task recognition.

A task can be characterised by many factors in the ubiquitous sensing scenario, such as human activities, actions, mobility, social encounters, and online behaviours. We, therefore, approach task recognition with the Cyber-Physical-Social (CPS) modelling paradigm [Ren et al., 2017, 2018a,b]. We hypothesise that tasks can be recognised from modelling the underlying signals of cyber activities (e.g., app and online activities), physical activities (e.g., mobility [Rahaman et al., 2018] and locomotive activities) and social activities (e.g., interaction with others during a task). Specifically, this research recognises five different categories[1] of tasks: work-related, personal, social-exercise-entertainment, caring, and civil obligation related. We capture these task categories through a task entity recognition process. For each task category, we further expand the annotations to tasks that the users were engaged in. The list of tasks may include travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, finance, IT (software or hardware-related tasks), customer care and problem-solving. These tasks were labelled leveraging the idea of task-taxonomy by Trippas et al. [2019].

For this research, we have generated a task behaviour dataset, which includes hourly and daily logs of user tasks over a four-week period. Participants include non-professionals (e.g., students) and busy professionals with a wide range of occupations, from office workers to business owners. We collect participants' task behaviours by continuously logging the smartphone sensor and app foreground and background data, and their laptop/desktop running application logs, along with the annotations of performed tasks based on the recall mechanisms of both *in-situ* (e.g., recalling a recent task in the past one hour) and *retrospective* (e.g., recalling all tasks at the end of the day which were performed throughout the day).

To our knowledge, this is the first work aimed at recognising tasks in daily life, utilising cyber, physical, and social activity signals. Our key research contributions include:

- A new problem formulation for task recognition in daily life.
- A novel framework to capture and recognise tasks in daily life.
- The proposed CPS activity modelling to derive intrinsic and specific characteristics of a wide range of daily tasks, from personal, social, and caring to work-related tasks.

[1] American Time Use Survey (ATUS): https://www.bls.gov/news.release/pdf/atus.pdf

- Analysis and discussion of the underlying cyber, physical, and social activity signals in modelling task behaviours across different cohorts of busy professionals and students.

## 5.2 Related Work

The proliferation of ubiquitous computing has enabled humans to perform daily tasks, depending on their activities, mobility, and online and social factors, in both spatial and temporal dimensions. For an efficient system interface design, a human task can be leveraged as the reference to guide the design process to allow successful human-computer interactions, especially in a ubiquitous scenario. The urgency of such research was first mentioned by Miller in 1976 [Miller, 1976], which led to further studies in computer graphics and visualisations [Pfautz, 2002, Chen and Thropp, 2007, Peng et al., 2010, Lin and Kuo, 2011].

A machine-learning-based approach to task boundary identification was presented by Khabsa et al. [2018], who trained a binary classifier to decide whether two consecutive interactions are part of the same task. However, task recognition is not considered in that research. Moreover, the study is restricted to the realm of cyberspace for human tasks. In recent work, Stisen et al. [2017] proposed task phase recognition and task progress estimation by modelling highly mobile workers in a large hospital complex, via the approximation of localisations from WiFi access points and human activities from workers' smartphone accelerometer sensors. However, their sensing scenario is only focused on a set of tasks that have clear objectives (achievement criteria), for example, the completion of tasks based on the delivery of orderlies. Furthermore, their experiment is conducted in a restricted and closed environment, i.e., within the coverage of WiFi access points inside a hospital. In a broader view of context-aware computing, ubiquitous sensing should be addressed in uncontrolled environments, where users' activities and mobility change over time and space. In this case, task recognition is a non-trivial and difficult challenge to be studied in the daily life of a mobile user.

The discovery of contextual information from daily routines has been previously addressed by Nguyen et al. [2016] by simultaneous context (i.e., latent patterns) and community extraction using a unified Bayesian nonparametric framework. Such structure and group discovery of mobile users are important for context-aware computing, especially to recognise human tasks from ubiquitous sensing “in the wild". In other words, this approach of social grouping could be crucial in distinguishing typical tasks being performed by mobile users, based on their daily behaviours and common profiles (e.g., users who work in the academic and educational sector). Another ubiquitous sensing application is presented by Sarker and Salim [2018] to mine behavioural rules from smartphones in managing incoming calls. Discovering such behavioural rules would provide greater impact if a user's on-task behaviour can be identified prior to receiving potentially distracting calls. Sensor data and machine learning have been studied to construct attention management systems and interruptibility models [Anderson et al., 2018], i.e., to predict when, or in which device [Mehrotra et al., 2019], to send a notification in order to minimise the negative impact on user experience. The prediction of user tasks using human-computer interaction and machine learning is presented by Stumpf et al. [2005]. They identify a task by observing an activity sequence, e.g., opening a file, saving a file, sending an email, cutting and pasting information. Recently, Ren et al. [2018b] investigated how to predict users' demographics by considering their CPS behaviours. However, the CPS aspects that could further describe a user task were not considered.

Our work falls under the umbrella of *anticipatory mobile computing* [Pejovic and Musolesi, 2015], where different types of sensors in mobile phones are used to feed a machine-learning algorithm to recognise the task the user is performing. Sensors and behavioural data have been used to address different data mining problems, including the modelling and prediction of time-based reminders [Graus et al., 2016], the estimation of the duration of tasks [White and Hassan Awadallah, 2019], and the modelling of context and intent of users for context-aware recommender systems [Sun et al., 2016]. To the best of our knowledge, the recognition of tasks from ubiquitous sensing in daily life is yet untapped, especially by incorporating CPS contexts in behaviour modelling of mobile users. Therefore, the contributions of this chapter will enable future intelligent and assistive applications from ubiquitous sensing to support daily and time-consuming human tasks, powered by smart devices in the personalised and ubiquitous environments of mobile users.

## 5.3 Problem Formulation

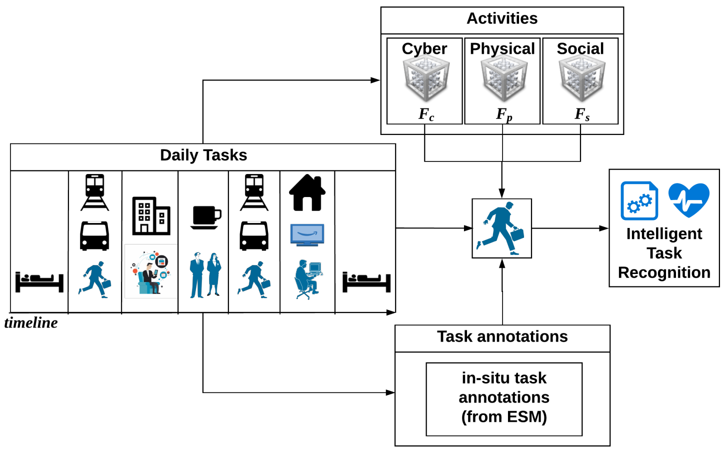



Figure 5.1: Task recognition problem in daily life.

Table 5.1: Symbols used throughout the description and analysis.

| Symbol | Description |
|---|---|
| $u$ | the user |
| $t$ | the time stamp |
| $A = \{a_i\}$ | the set of tasks $a_i$ |
| $A_{esm} = \{a_{esm}\}$ | the set of raw annotation in ESM |
| $F_c = \{f_c\}$ | the set of cyber features $f_c$ |
| $F_p = \{f_p\}$ | the set of physical features $f_p$ |
| $F_s = \{f_s\}$ | the set of social features $f_s$ |
| $z(\cdot,\cdot)$ | presence-based task boundary construction function |
| $g(\cdot)$ | function to mapping CPS features and tasks |

In this section, we formulate the problem of recognising daily tasks based on CPS activities of human participants. The CPS activities are derived from the sensing log signals (refer to Figure 5.1). Note that these signals can be sourced from smart devices (e.g., smartphones, tablets, wearables, desktop computers and the Internet of Things). Moreover, the ground truth labels of daily tasks performed by the participants are also captured through in-situ task annotations. Our aim is to derive the task annotations by characterising the sensing log signals. Table 5.1 shows the notations used throughout this chapter.

**Cyber Activities:** These include users' involvement in different cyber activities, such as emailing, Web browsing, social networking, entertainment and many other applications. Each cyber activity contains the cyber content and the timestamp, e.g., web browsing activity contains the website visited and the corresponding timestamp of visitation. In this research, we investigate the following cyber activities: *social Networking, utilities, communication & scheduling, news & opinion, entertainment, design & composition, business, reference & learning, software development* and *shopping*. We define each kind of cyber activity as a binary variable $f_c$ to represent whether users are involved in the corresponding activity at a certain time. Thus, a user's cyber activity is defined as a set of records: $< u, F_c, t_i >$, where $u$ is the user, $t_i$ is the timestamp, and $F_c$ is the set of cyber features, denoting the user's involvement in the above mentioned cyber activities.

**Physical Activities:** These contain mainly the physical activities of users in the spatio-temporal domain. These activities are captured through the readings of *accelerometer, gyroscope, magnetometer sensors, transport mode*, and even the semantic labels of *visited locations* (e.g., home, office and train stations). Thus, a user's physical activity is defined as a set of records: $< u, F_p, t_i >$, where $F_p$ is the set of physical features, denoting the users' physical activities as mentioned. The details of each feature $f_p \in F_P$ are presented in Section 5.5.

**Social Activities:** These contain information about *social interactions* during the progress of users' daily tasks including direct interactions with other people. These activities (i.e., interactions) are captured through *the presence of WiFi/Bluetooth access points, the surrounding noise levels, and in-situ annotations, all of which characterise the types of environment and degree of social encounters with other individuals*. A user's social activity is defined as a set of records: $< u, F_s, t_i >$, where $F_s$ is the set of social features, denoting the users' social environment as mentioned above. Again, details of the social features used in this study are presented in Section 5.5.

**Tasks:** These denote the *daily tasks* performed by users. There are various categories of tasks [Trippas et al., 2019] including travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, finance, IT (software- or hardware-related tasks), customer care and problem solving. We claim that each task $a$ can be characterised by the associated combination of cyber, physical and social activities. To evaluate the completeness and accuracy of our claim, all tasks under each task category are also obtained via the **E**xperience **S**ampling **M**ethod [Csikszentmihalyi et al., 1977, Csikszentmihalyi and Larson, 2014]. For ESM-based annotation acquisition, the description of the most recent tasks are requested in-situ from the users. The users are required to recall the most recent tasks performed by them, with responses entered using a survey-style smartphone app.

**Task Boundary Construction:** Since the techniques and queries of task description (including their CPS activities) are distinct from user to user, the boundary of a task and granularity of contextual information can be different for the same experienced task when a mobile user is providing a corresponding annotation. The annotation that the user provides can be associated with its relative perception upon answering the short questionnaire. These associations can be inferred from both temporal and spatial user contexts. In this study, we consider the relative user contexts on the temporal domain where the recall of recent user task is based on restricted time range (e.g., recent one-hour time slot). Defining the task boundary is an essential process, in order to reconstruct the inferred time-slot of user annotation for intelligent task recognition purpose.

**Task Recognition:** This is formulated as follows: Given the CPS activities of a user $u$ at time $t$, the recognition of the task $a$ currently being undertaking is defined as:

$$g(F_c, F_p, F_s) \rightarrow a, \tag{5.1}$$

where $g(\cdot)$ is a function that establishes a mapping between a task and its CPS activities denoted as $F_c, F_p$ and $F_s$, respectively.

## 5.4 Proposed Framework

In this section, we propose a framework to recognise daily tasks for mobile users based on their CPS activities. As shown in Figure 5.2, this consists of the following key components: task annotation, presence-based task boundary construction, CPS feature construction, CPS-based modelling, and CPS-based learning.

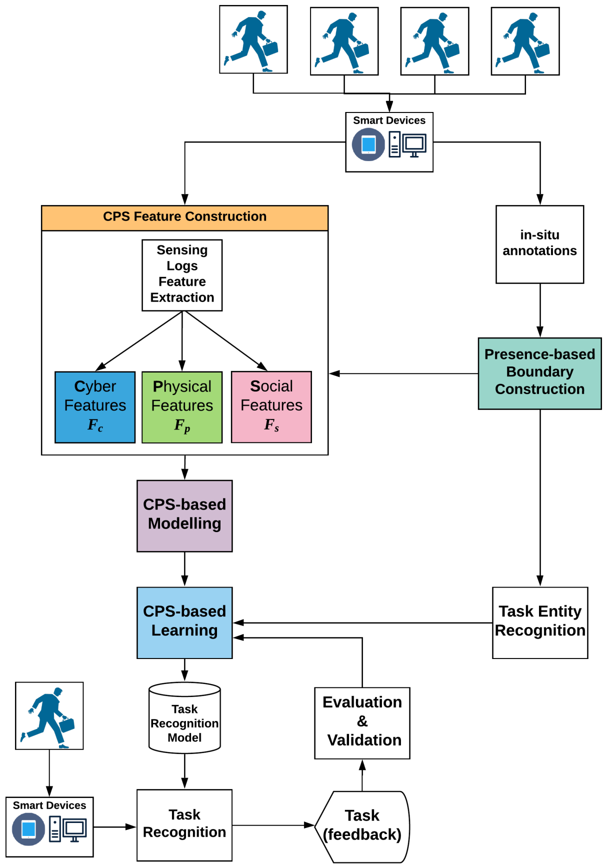



Figure 5.2: Conceptual framework of intelligent task recognition.

### 5.4.1 Task Capture: In-situ Annotations

In a typical ESM-based study, the annotations are acquired from timepoint-based experience sampling. In this case, the acquisition of task annotations (i.e., ESM-based annotations) is achieved through in-situ surveys that can be triggered by notifications through an app.

**ESM-based annotations:** These are based on a quick questionnaire in order to minimise interruption to daily activities and tasks. Therefore, the questions in relation to the performed tasks should not be too long and must be straightforward. ESM aims to minimise human cognitive bias while reducing the reliance on participants' abilities to accurately recall earlier experiences [Berkel et al., 2017] (i.e., human tasks). Hence the ESM process typically does not include a question that asks the mobile user about the actual task start time. In this study, the annotations acquired from the ESM process are defined as in-situ annotations. As shown in Figure 5.3, an in-situ survey is displayed upon a mobile app notification at $t$; the annotation is conducted at time $t_\delta$, corresponding to the task and its contextual information within the boundary of $t_\beta$ and $t$; $\beta$ is the estimated boundary of a performed task that can be inferred from the ESM annotation process. For example, the question of "What kind of task did you engage in between 10:00 AM and 11:00 AM, that you spent most of your time on?" corresponds to $t$ = 11:00 AM when the ESM was requested through a mobile app notification, $\beta = 60$ minutes and $t_\beta$ = 10:00 AM. On the other hand, the question of "What kind of time-consuming task are you currently engaged in?" corresponds to $t$ = 11:15 AM when the ESM was requested through a mobile app notification and $\beta = 30$ minutes according to a predefined parameter in the system.

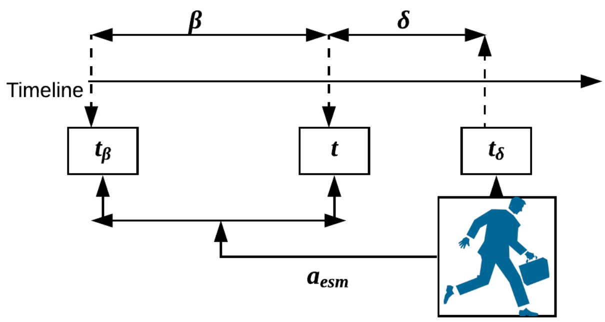



Figure 5.3: Task annotation acquisition through ESM.

### 5.4.2 Presence-based Task Boundary Construction

For in-situ annotations acquired through ESM, a mobile user can answer the questionnaire the moment an app notification comes. In a real-world scenario, the nature of the user's engaged tasks can be complex. Therefore, the user could be very busy on a complex task and answer it later when the chance occurs, either before or after the next hourly mobile app notification for ESM survey.

Defining such a task boundary from ESM annotations is an essential process for our proposed intelligent task recognition framework due to subjective perceptions of mobile users. In many instances, point-based ESM annotations are predominant in the field of ubiquitous computing and intelligent mobile sensing, where no exact start and end time are defined concretely by the user. The intuitive approach to address this issue is to assign a task to a specific time segment based on its presence.

A rule-based function can be applied to all in-situ (i.e., $A_{esm}$) annotations for a particular day to establish the boundaries of these tasks. A simplistic approach applied in our framework refers to the following rules. Any timestamped in-situ annotations correspond to previous hour time segments, where each segment has $\beta$ time length. This condition is only valid for an annotation that has its timestamp recorded before the maximum time segment of the day. Any annotation outside the range of defined time segments should be allocated to the maximum time segment of the day. For example, a user can answer the ESM survey at 11:00 PM where the questionnaire would refer to what task the user was performing during the 06:00 PM to 07:00 PM time slot (i.e., the maximum time segment of the day). Furthermore, this rule-based function should be robust towards the shifting temporal context (i.e., timezone shift) of the mobile user. In fact, the user could be moving between countries (e.g., for holiday or business purposes) during the course of mobile sensing data collection, which is also evident from the participants in this study.

The boundary of a task can be expanded to the predefined time segments of a day. In this case, the function for presence-based task boundary construction can be expressed as follows:

$$z(a_{esm}) = L_{a_{esm}} \oplus TZ_{a_{esm}}, \tag{5.2}$$

where $L_{a_{esm}}$ denotes the set of words from the raw annotations of $a_{esm}$, $TZ_{a_{esm}}$ corresponds to set of epoch-timestamps (including their timezones) when the ESM surveys are answered, and $\oplus$ refers to the element-wise operator that links the two sets (i.e., $L_{a_{esm}}$ and $TZ_{a_{esm}}$) for presence-based task boundary construction process. It should be noted that $L_{a_{esm}}$ and $TZ_{a_{esm}}$ must have equal item length $n_{a_{esm}}$. Inherently, the $z$ function generates a set of items (with $n_{a_{esm}}$ length) consisting of the time boundaries of the in-situ task annotations.

Moreover, the result of task boundary construction can be used as the input for task entity recognition. In this case, the purpose of this entity recognition module (refer to the conceptual framework in Figure 5.2) is to define the categorisation of common tasks from textual annotations. The output of the entity recognition process would be the set of class labels to be used for task recognition. For instance, our running example "meeting friends social get together" will be labelled as "social/exercise/entertainment" task by the entity recognition module.

We note that the result from task boundary construction can also be used for constructing CPS features. For example, the indication of a task being performed collaboratively with the presence of other people as one of the social features in a meeting (e.g., attending a meeting with colleagues or in a group discussion), which is allocated to specific time segment (inferred task boundary).

### 5.4.3 CPS Feature Construction

This module utilises sensing logs from users' smart devices to construct CPS features associated with a task. Since raw sensing signals are timestamped and can be streamed from these smart devices, we define the following general process for CPS feature construction.

All features (i.e., CPS feature sets) can be constructed based on the alignment of user tasks and the time segments defined within the scope of annotation fusion. Therefore, several functions can be applied to these raw signals to construct the following feature sets:

- $F_c$: A cyber feature set consisting of the features that are related to a user's cyber activities, such as smartphone app usage patterns, categories of visited web domains and application usage.
- $F_p$: A physical feature set consisting of the features that are related to a user's physical movement, locations (including their semantics), and mobility, such as the magnitude of the accelerometer, gyroscope and magnetometer signals of user's smartphone or wearable device, transportation mode, change of location clusters, and transportation hotspots.
- $F_s$: A social feature set consisting of the features that are related to a user's ambient sensing environment, social profiles and interactions with other individuals on the tasks, such as the relative noise level surrounding the user, indication of proximity to other individuals or sensing devices, direct interaction with an individual, or the number of people involved in completing a task.

Table 5.2: CPS feature sets used in modelling.

| Feature Set | Features |
|---|---|
| Cyber Activities | Binary features of uncategorized, social networking, utilities, communication & Scheduling, news & opinion, entertainment, design & composition, business, reference & learning, software development, shopping within the scope of one hour before the task and during the task. |
| Physical Activities | Statistical features from sliding window model on magnitudes of accelerometer, gyroscope, and magnetometer readings. |
| Social Activities | The count of unique ID of wireless access points (i.e., BSSID) and statistical features from sliding window model on noise level. |
| **Task Labels [Trippas et al., 2019]** | **Travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, finance, IT (software or hardware-related tasks), customer care, or problem-solving.** |

Note that the statistical features about CPS activities (as described in Table 5.2) correspond to the following temporal features extracted from a sliding window model with window size $\delta = 300$ seconds and 50% overlap: *mean, median, maximum, minimum, standard deviation, interquartile range (IQR),* and *root mean square (RMS).*

Specifically, in each window for $F_p$ construction, the magnitudes of accelerometer, gyroscope and magnetic field are computed according to the following equation:

$$magnitude := \sqrt{x_{sensor}^2 + y_{sensor}^2 + z_{sensor}^2} \tag{5.3}$$

where $x_{sensor}$, $y_{sensor}$, $z_{sensor}$ are the tri-axial sensor values of a smartphone's sensor (i.e., accelerometer, gyroscope or magnetometer). Moreover, the noise level and magnitude values from accelerometer, gyroscope and magnetometer readings are normalised using min-max normalisation.

### 5.4.4 CPS-based Task Modelling and Learning

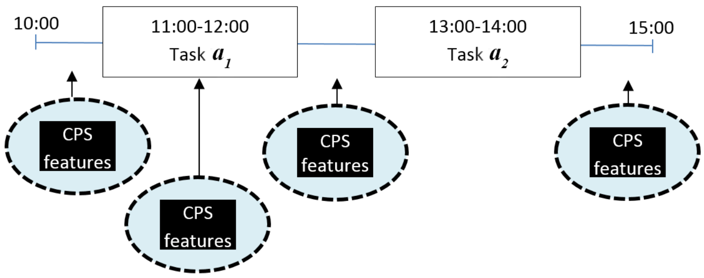



Figure 5.4: An example of CPS-based task modelling, illustrating the computation of CPS features for two tasks ($a_1$ and $a_2$) in different time segments.

In this module, features constructed from the three different feature spaces are integrated together to build a CPS-based task model. Figure 5.4 shows an example of a task timeline between 10:00 and 15:00 and consisting of two tasks: $a_1$ and $a_2$. Ideally, the CPS-features can be seen in anywhere in the task timeline (before, during and after a specific task). To build the CPS-based task model, the temporal dependency of feature sets *before*, and *during* the task were considered in our experiment.

CPS-based modelling can be applied to any of the CPS feature sets (i.e., $F_c$, $F_p$ or $F_s$). As shown in the previous section, we expanded $F_c$ to include the cyber features one hour before and during a task, while a sliding window model is applied to extract statistical features from smartphone sensors for $F_p$ and $F_s$. The combination of $F_c$, $F_p$ and $F_s$ will produce the final set defined as the CPS feature set, which will be used for learning purposes (i.e., building classifiers for intelligent task recognition).

The CPS feature set is then used to build a set of classifiers. In this module, the learning process includes training, testing and internal evaluation processes. Hence, the best classifier can be selected (based on certain metrics, such as accuracy and $F_1$-score) and used for real-world task recognition on a mobile user. In our conceptual framework, a real-world scenario is also considered where a mobile user may provide the feedback for the actual tasks and the portion of sensing logs can be used to retrain the model (i.e., improve the performance of task recognition model from CPS-based learning using semi-supervised approach). However, the scope of this research is currently limited towards the recognition of tasks based on the CPS activities of mobile users.

## 5.5 Experiment and Evaluation

### 5.5.1 Mobile Data Collection and Task Capture

To evaluate our task recognition framework, we collected a task dataset from 17 participants over a maximum of 20 week-days.[2] In other words, each participant had to dedicate her time to provide annotations during a one month period of data collection (from Monday to Friday).

The data collection was performed using Android smartphone apps (RescueTime[3] and our sensor data collection app, denoted as sensing-app) and a desktop app (i.e., RescueTime[4], to collect cyber data, from visited web domains and their categorisations). Our sensing-app recorded sensor data with the following reading frequency settings:

- Accelerometer: 50 Hz.
- Magnetometer: 50 Hz.
- Gyroscope: 50 Hz.
- Noise level: 1 second.

It should be noted that the actual number of data points are subject to the hardware capabilities of user's phones and resource availability, although the frequencies of these sensors were prepared with these settings. For example, if the user's phone has the maximum capability of 20 Hz, then the Android operating system will choose 20 Hz instead of 50 Hz. Moreover, the phone may also run out of battery during the period of data collection in a day.

To minimise the battery usage of the data collection Android app, we collected these sensor data within 30 seconds timeframes, and a one minute gap between frames for no data collection mode. The

[2]Data collection protocol reviewed and approved by the Human Research Ethics Committee at RMIT University (see Appendix C).
[3]https://play.google.com/store/apps/details?id=com.rescuetime.android
[4]https://www.rescuetime.com/download

collection of sensor data was only scheduled from 06:00 AM to 07:00 PM. Moreover, we provided the flexibility for the users to pause the sensing-app during the day. The participants reported their timestamped tasks through the ESM (triggered through hourly app notifications, from 08:00 AM to 07:00 PM). In our proposed framework, a task entity recognition process can be used on the in-situ annotations after their boundaries are constructed (refer to Section 5.4.2) to assign each reported task to one of the following task categories (compact categorisation derived from American Time Use Survey[5]):

- **Work-related tasks:** Tasks related to the participant's roles. A participant can have multiple roles (e.g., many jobs).
- **Personal tasks:** Tasks related to personal organisation, reflection or care (includes commuting, cleaning and house improvement).
- **Social/exercise/entertainment tasks:** Tasks related to social events, exercise and relaxation (entertainment).
- **Caring tasks:** Tasks related to taking care of household or non-household members.
- **Civil obligations:** Tasks related to civil obligations of the participant (e.g., "voting (election), signing the petition, or participating in a strike (protesting against government policies)".

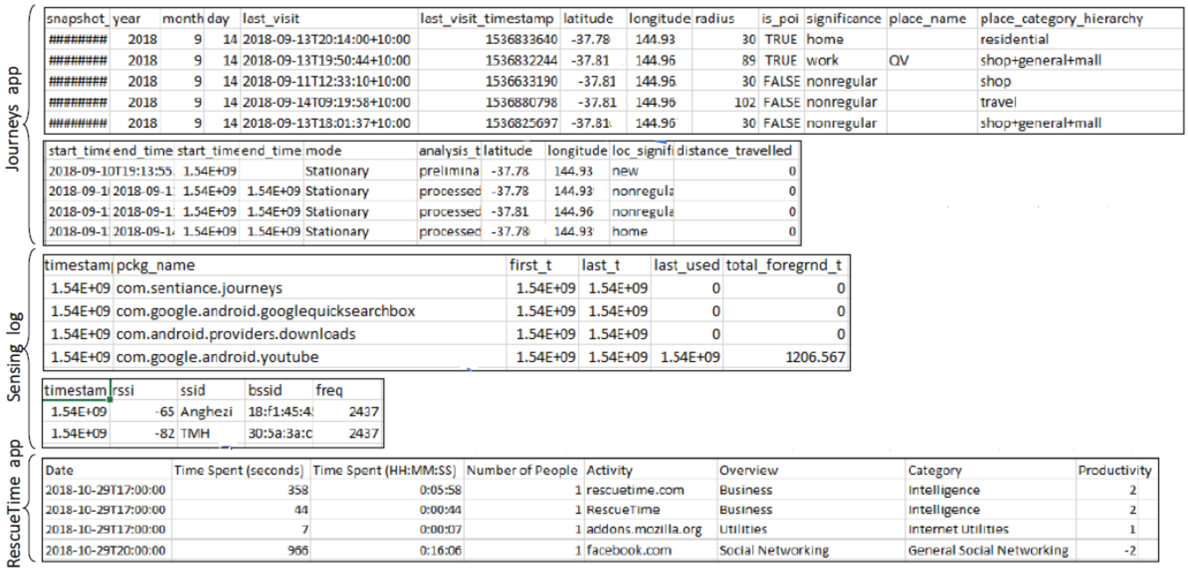
### 5.5.2 Collected Mobile Sensing Data and Task Annotations

In our data collection campaign, the protocol for logging mobile sensing data and its task-capture survey design are reproduced (and adjusted) from [Liono et al., 2019b] on task intelligence to collect rich pervasive sensing data. The collection of sensor data is built upon two independent approaches in acquiring task annotations from participants. The first approach was to use ESM on hourly surveys (see Appendix A) through the mobile app notifications from 08:00 AM to 07:00 PM. Our sensor data collection also acquires user annotations via a Daily Reconstruction Method (DRM) approach, to chronologically recall the tasks that the user has performed in a day (see the survey in Appendix B). It should be noted that the experiments in this chapter are only targeted on the ESM-based task annotations since their boundaries are not strictly defined (as previously mentioned in the formal problem formulation).

Hence, utilising the data that we collect (including personal lifestyles, movement behaviours and progress of tasks according to user perception) from real participants will expand the task understanding beyond this study, which could lead to many interesting research directions on ubiquitous computing. Figure 5.5 illustrates the sample snapshot of logged mobile sensing data except raw logs of accelerometer, gyroscope, magnetometer and noise level for a de-identified participant. Note that the data shown in

[5] American Time Use Survey (ATUS): https://www.bls.gov/news.release/pdf/atus.pdf

Figure 5.5 is an obfuscated version of the original logs to preserve the privacy (including GPS locations) of the participant.




| snapshot_ | year | month | day | last_visit | last_visit_timestamp | latitude | longitude | radius | is_poi | significance | place_name | place_category_hierarchy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ######## | 2018 | 9 | 14 | 2018-09-13T20:14:00+10:00 | 1536833640 | -37.78 | 144.93 | 30 | TRUE | home | | residential |
| ######## | 2018 | 9 | 14 | 2018-09-13T19:50:44+10:00 | 1536832244 | -37.81 | 144.96 | 89 | TRUE | work | QV | shop+general+mall |
| ######## | 2018 | 9 | 14 | 2018-09-11T12:33:10+10:00 | 1536633190 | -37.81 | 144.96 | 30 | FALSE | nonregular | | shop |
| ######## | 2018 | 9 | 14 | 2018-09-14T09:19:58+10:00 | 1536880798 | -37.81 | 144.96 | 102 | FALSE | nonregular | | travel |
| ######## | 2018 | 9 | 14 | 2018-09-13T18:01:37+10:00 | 1536825697 | -37.81 | 144.96 | 30 | FALSE | nonregular | | shop+general+mall |

| start_time | end_time | start_time | end_time | mode | analysis_t | latitude | longitude | loc_signifi | distance_travelled |
|---|---|---|---|---|---|---|---|---|---|
| 2018-09-10T19:13:55. | | 1.54E+09 | | Stationary | prelimina | -37.78 | 144.93 | new | 0 |
| 2018-09-1 | 2018-09-1 | 1.54E+09 | 1.54E+09 | Stationary | processed | -37.78 | 144.93 | nonregula | 0 |
| 2018-09-1 | 2018-09-1 | 1.54E+09 | 1.54E+09 | Stationary | processed | -37.81 | 144.96 | nonregula | 0 |
| 2018-09-1 | 2018-09-1 | 1.54E+09 | 1.54E+09 | Stationary | processed | -37.78 | 144.93 | home | 0 |

| timestam | pckg_name | first_t | last_t | last_used | total_foregrnd_t |
|---|---|---|---|---|---|
| 1.54E+09 | com.sentiance.journeys | 1.54E+09 | 1.54E+09 | 0 | 0 |
| 1.54E+09 | com.google.android.googlequicksearchbox | 1.54E+09 | 1.54E+09 | 0 | 0 |
| 1.54E+09 | com.android.providers.downloads | 1.54E+09 | 1.54E+09 | 0 | 0 |
| 1.54E+09 | com.google.android.youtube | 1.54E+09 | 1.54E+09 | 1.54E+09 | 1206.567 |

| timestam | rssi | ssid | bssid | freq |
|---|---|---|---|---|
| 1.54E+09 | -65 | Anghezi | 18:f1:45:4 | 2437 |
| 1.54E+09 | -82 | TMH | 30:5a:3a:c | 2437 |

| Date | Time Spent (seconds) | Time Spent (HH:MM:SS) | Number of People | Activity | Overview | Category | Productivity |
|---|---|---|---|---|---|---|---|
| 2018-10-29T17:00:00 | 358 | 0:05:58 | 1 | rescuetime.com | Business | Intelligence | 2 |
| 2018-10-29T17:00:00 | 44 | 0:00:44 | 1 | RescueTime | Business | Intelligence | 2 |
| 2018-10-29T17:00:00 | 7 | 0:00:07 | 1 | addons.mozilla.org | Utilities | Internet Utilities | 1 |
| 2018-10-29T20:00:00 | 966 | 0:16:06 | 1 | facebook.com | Social Networking | General Social Networking | -2 |

Figure 5.5: A partial snapshot of the raw sensing logs.

We conducted the task annotations surveys based on the ESM method since the idea of ESM is to minimise human cognitive bias while reducing the reliance on the participants' ability to accurately recall earlier experiences. Specifically, a short questionnaire is sent through push notification on an hourly basis aiming to minimise the interruption to daily activities and tasks. In this study, the annotations acquired from the ESM process are defined as in-situ annotations. Figure 5.6 illustrates a partial snapshot of captured task annotations through ESM.

| UserId | Sampled Time | Task | Task_Category | Micro-task | Task_Presence | Time_Spent | Freshly_Started | Progress | Cyber | Physical | Social |
|---|---|---|---|---|---|---|---|---|---|---|---|
| U001 | Wednesday, 13-03-2019 08:12 | commuting to work | personal | travel | multi-activity | 2 hour | NO | complete | NO | NO | NO |
| U001 | Wednesday, 13-03-2019 19:01 | ate dinner | personal | meals_breaks | one-activity | 1 hour | NO | complete | NO | NO | NO |
| U001 | Friday, 01-03-2019 19:03:44 | catching up with frien | social-exercise | communicati | one-activity | 2 hour | NO | 80-99% | NO | NO | YES |
| U001 | Wednesday, 06-03-2019 19:42 | Gym workout | social-exercise | physical | one-activity | 2 hour | NO | 40-59% | YES | YES | NO |
| U001 | Tuesday, 19-02-2019 04:03:31 | Investigation on Email | work-related | it | one-activity | > 2 hour | NO | 40-59% | YES | YES | YES |
| U001 | Wednesday, 20-02-2019 12:14 | Compliance Gap analy | work-related | it | one-activity | 1 hour | NO | 20-39% | YES | YES | YES |

Figure 5.6: A partial snapshot of the captured tasks using ESM.

### 5.5.3 Task Annotations: Non-professionals and Busy-professionals

In this subsection, we perform an exploratory analysis of in-situ task annotations collected from ESM-based hourly surveys. The 17 participants in our study are diverse, consisting of two distinct cohorts: twelve non-professionals (e.g., full-time or part-time students) and five busy-professionals.

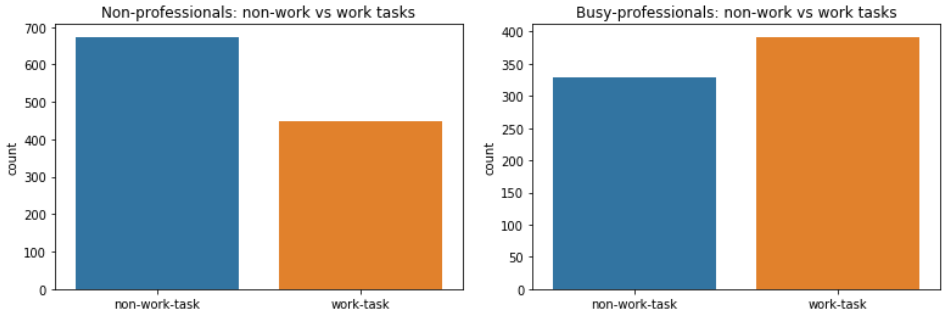



Figure 5.7: Frequency of non-work and work-related tasks: non-professionals and busy-professionals.

Figure 5.7 shows the frequency distribution of non-work and work-related tasks. It is evident from the independent groups that the busy-professionals have a relatively higher proportion of work tasks than non-work tasks, and vice versa for non-professional participants. Non-work tasks comprise the other four big task categorisations: personal, social/exercise/entertainment, caring and civil obligation related tasks.

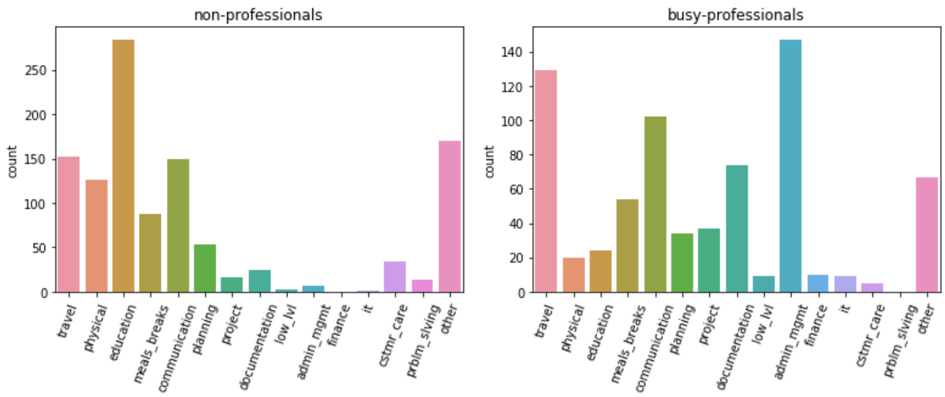



Figure 5.8: Frequency of tasks: non-professionals and busy-professionals.

We further expand each task category into more granular tasks based on the recent study performed by Trippas et al. [Trippas et al., 2019] on work-tasks. Specifically, we leverage their taxonomy to assign task annotations given by the users to one of the following tasks: travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, finance, IT (software- or hardware-related tasks), customer care and problem solving. Any task annotations that do not belong to any of these task categories are relabelled as "other". Given the task taxonomy, an independent research annotator manually assigned each raw task annotation to any of these task categories (denoted as tasks throughout this chapter). The distribution of these tasks are shown in Figure 5.8 for both non-professionals and busy-professionals. In this chapter, we perform task recognition based on CPS activities of daily human life in various pervasive sensing environments.

### 5.5.4 Non-professionals vs. Busy-professionals by Sub-categories

#### 5.5.4.1 Work-related Tasks

For work-related tasks, Figure 5.9 evidently shows a large number of education tasks for non-professionals. it is usual since the majority of these users has the main role of being a student. The following communication and customer care appeared with total numbers of annotations (below 50 in-situ annotations). This drastic drop of work-related non-education tasks could be caused by inconsistencies in their other roles (such as part-time work, which typically starts from late afternoon or evening). We discovered this from the direct interaction with the users (through weekly meetings between a researcher and corresponding participant). It should be remembered that all the hourly surveys are subject to the time slots from 07:00 AM to 07:00 PM. On the other hand, the top-five work-related tasks of busy-professionals consist of admin and management, communication, documentation, project and planning. This distribution is inherently aligned with busy-professional demographics, where the majority of those in our study are office-workers.

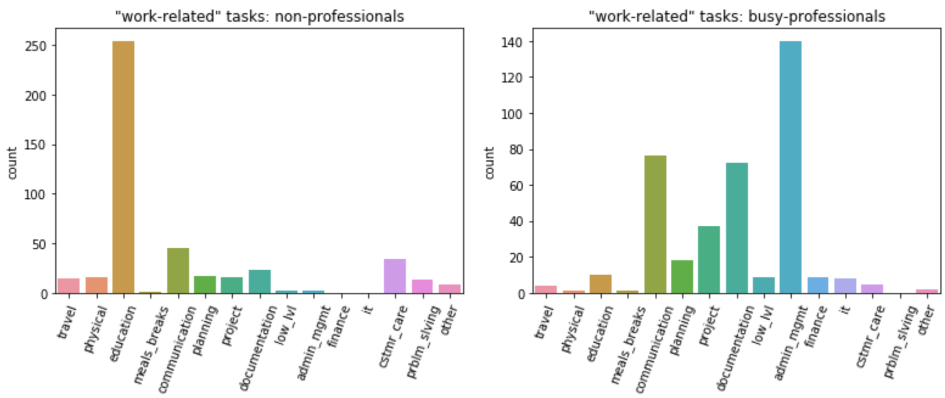



Figure 5.9: Frequency of tasks in the *work-related* category, for non-professionals and busy professionals.

#### 5.5.4.2 Personal Tasks

For personal tasks (shown in Figure 5.10), it is interesting that the majority of in-situ annotations reported for both non-professionals and busy-professionals are associated with travel. From the raw annotations that the user answered on the ESM-based hourly surveys, we discovered that most of the travel related tasks are associated with the act of travelling to the locations where they spend most of the time in their

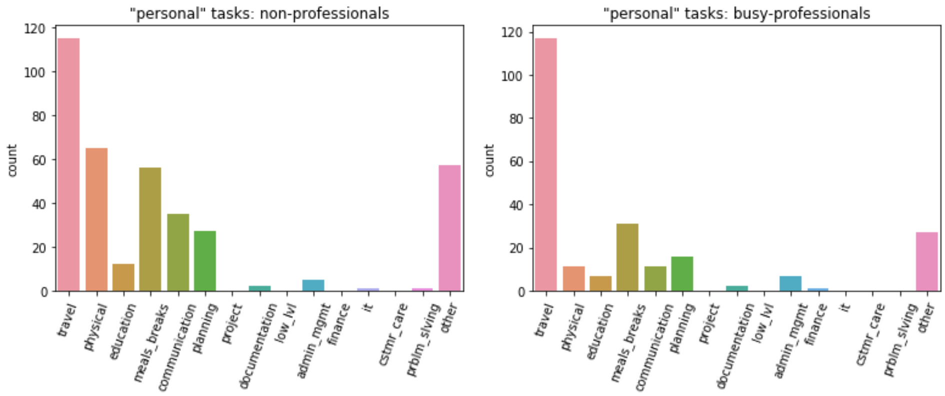



Figure 5.10: Frequency of tasks in the *personal* category, for non-professionals and busy professionals.

main roles or occupations (e.g., commuting to university or going to the office in the morning). On the other hand, the act of leaving those locations, e.g., heading back home after work, are likely to be marked as personal tasks by these participants. For busy-professionals, there are limited numbers of physical tasks (personal tasks) reported during the weekdays. This outcome is relatively aligned with the fact that office workers could be progressing on their tasks without much physical movement, and be seated for the work-related tasks, such as admin and management, communication, documentation, project and planning tasks.

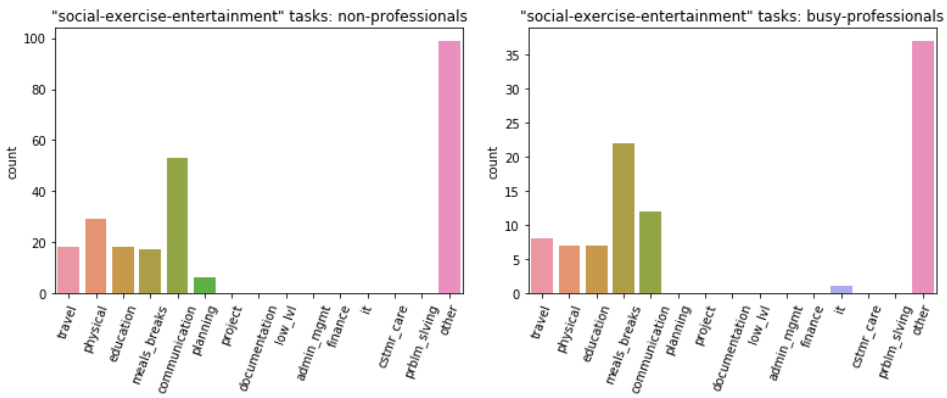



Figure 5.11: Frequency of tasks in the *social/exercise/entertainment* category, for non-professionals and busy professionals.

#### 5.5.4.3 Social/exercise/relaxation Tasks

Social, exercise and relaxation (entertainment) related tasks may not be completely aligned with the taxonomy of tasks. This is shown in Figure 5.11, where the majority of in-situ annotations could not be associated with any of the information workers' task categories, therefore being relabelled as "other". In the top-five tasks reported (except "other") by both non-professionals and busy-professionals, these three tasks are included: communication, travel and education. For social activities, a person would communicate with another in a gathering event. An activity such as travelling outside the CBD area can be relaxing and is typically marked as travelling for relaxation. The appearance of travel tasks (relaxation) is likely to be prominent during public holidays.

#### 5.5.4.4 Caring Tasks

Busy-professionals are more likely to focus on work-related tasks during the days. Therefore, tasks such as "caring" were less likely to be engaged on weekdays (from morning to afternoon). However, both cohorts agree that at least communication and physical were usually involved in caring tasks (as shown in Figure 5.12).

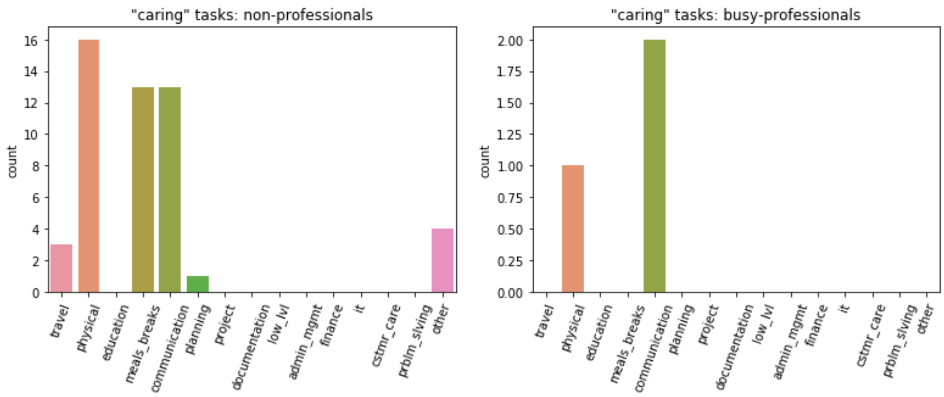



Figure 5.12: Frequency of tasks in the *caring* category, for non-professionals and busy professionals.

#### 5.5.4.5 Civil-obligation related Tasks

On the other hand, the task such as communication is typically more dominant for civil obligation related tasks (refer to limited sample data in Figure 5.13). In an event such as signing a petition or participating in a protest, communication is important to complete the task goal. However, it shows this task category occurs rarely for the two cohorts. For busy-professionals, caring and civil obligation related tasks would be sporadic during a typical working day.

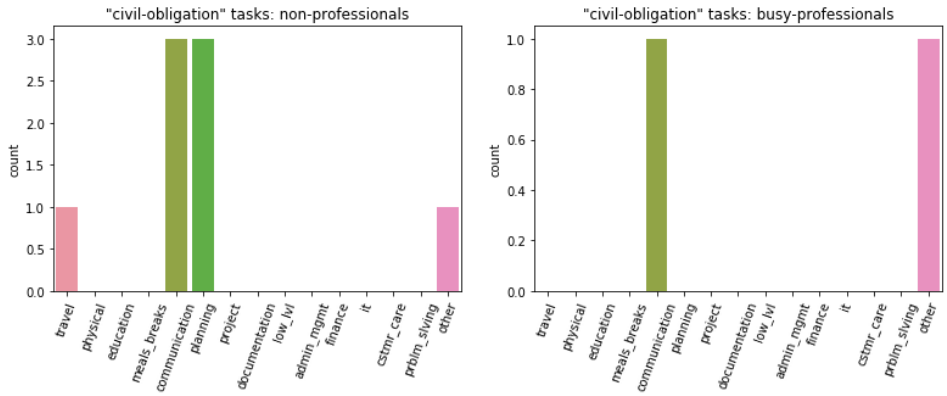



Figure 5.13: Frequency of tasks in the *civil obbligation* category, for non-professionals and busy professionals.

After data pre-processing and feature extraction (as detailed in Section 5.4.3), we use a five-minute sliding window model with 50% overlap for our experiment in the next subsection), the CPS feature sets contain a total of 7,653 instances for non-professional cohort (on all reported 1,121 in-situ annotations) and 5,271 instances for busy-professional cohort (on all reported 721 in-situ annotations), respectively. Consequently, 62 features are extracted corresponding to each task label of the instances, consisting of 22 features of $F_c$, 21 features of $F_p$ and 8 features of $F_s$.

### 5.5.5 Experimental Setup

In order to signify our contributions for CPS activity modelling, we conduct our study over three different experiment sets:

1. **Work-related tasks:** In this experiment set, the annotations categorised as “work-related” are included to perform task recognition (i.e., tasks associated with the main roles/occupations of corresponding users).
2. **Social/exercise/entertainment tasks:** In this experiment set, the annotations that belong to tasks related to social events, exercise and relaxation are included for recognising the associated tasks.
3. **Personal/caring/civil tasks:** As shown in the previous subsection, the annotations of “caring” and “civil obligation” related tasks are sparse. Therefore, these two task categories would be grouped together with personal tasks in the experiment set due to the similar nature of these tasks (to be completed personally by the corresponding user).

For each experiment set defined above, we conduct an empirical performance evaluation of intelligent task recognition using these following settings on both cohorts of non-professionals and busy-professionals:

- task recognition based on cyber set (i.e., $F_c$).
- task recognition based on physical set (i.e., $F_p$).
- task recognition based on social set (i.e., $F_s$).
- task recognition based on CPS contexts (i.e., combination of $F_c$, $F_p$ and $F_s$).

Separate task recognition using the CPS feature sets discretely are defined as the baselines in our experiment. In our implementation of intelligent task recognition, we deployed the following classifiers:

- Support Vector Machine (SVM).
- Naive Bayes.
- $k$-Nearest Neighbour ($k$-NN).
- Logistic Regression Classifier with Restricted Boltzmann Machine feature extractor, denoted as LRC (RBM).
- Decision Tree.
- Random Forests.

These classifiers are instantiated using the scikit-learn [Pedregosa et al., 2011] machine learning tools in Python. To build a Decision Tree classifier, information gain (entropy function) is used in the tree splitting process. One hundred trees are used to construct the ensembles for a Random Forests classifier. For building a classifier based on the SVM algorithm, a tolerance parameter of 0.001 is used with a radial basis function (RBF) kernel. For the Naive Bayes classifier, a variant algorithm Gaussian Naive Bayes is used. For LRC (RBM), we leverage a Bernoulli Restricted Boltzmann Machine model to perform effective non-linear feature extraction from CPS feature sets with a 0.06 learning rate, 100 hidden units, and 20 iterations. For $k$-Nearest Neighbour ($k$-NN) classification a setting of $k = 5$ is used.

Before we include the data for training and testing, all annotations that belong to "other" tasks are removed from our experimentation to reduce the performance bias in the classification result. Therefore, the intelligent system (e.g., a digital assistant) can focus on recognising essential tasks to provide necessary supports towards their task completion.

### 5.5.6 Evaluation

In order to evaluate and validate the performance of task recognition, stratified five-fold cross-validation was applied to the instances (produced by the five-minute sliding window model with 50% overlaps) of CPS feature sets. For the independent experiment of only the cyber feature set ($F_p$), there are 1,605 instances (951 instances for non-professional and 654 instances on busy-professional cohorts) comprised of binary features (refer to Table 5.2). These instances correspond to the total number of reported tasks. This means that different instances (task-windowed instances for $F_p$, $F_s$ and CPS feature sets and task instances for the $F_c$ feature set) for the same user can be in both train and test sets. The model for intelligent task recognition is built based on a person-independent approach. In other words, our proposed intelligent task recognition framework aims to discover and distinguish the general tasks for all mobile users, based on CPS contexts.

In our framework, the internal evaluation process is based on $F_1$-score. $F_1$-score refers to the harmonic mean of precision and recall of task recognition:

$$F_1 - score_a = \frac{2 * precision_a * recall_a}{precision_a + recall_a} \tag{5.4}$$

Here $precision_a$ refers to the number of correctly recognised tasks divided by the number of all recognised tasks of a particular class label, and $recall_a$ refers to the number of correctly recognised tasks divided by the total number of existing tasks in a test set, for a particular class label.

The results in Figures 5.14 and 5.16 ($F_1$-score and confusion matrix of best classifier) show the example of the imperative performance on work-related task recognition when our application is trained using all CPS features. From the outcome of our empirical evaluation on the three experiment sets, it is evident that incorporating all CPS feature sets together in the process of building a classifier, will provide better overall predictive performance.

#### 5.5.6.1 Recognition of Work-related Tasks

For work-related tasks, the experiment result is validated on the recognition of the following tasks:

- **Non-professionals:** travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, customer care and problem solving.
- **Busy-professionals:** travel, physical, education, meals and breaks, communication, planning, project, documentation, low-level, admin and management, finance, IT (software- or hardware-related tasks) and customer care.

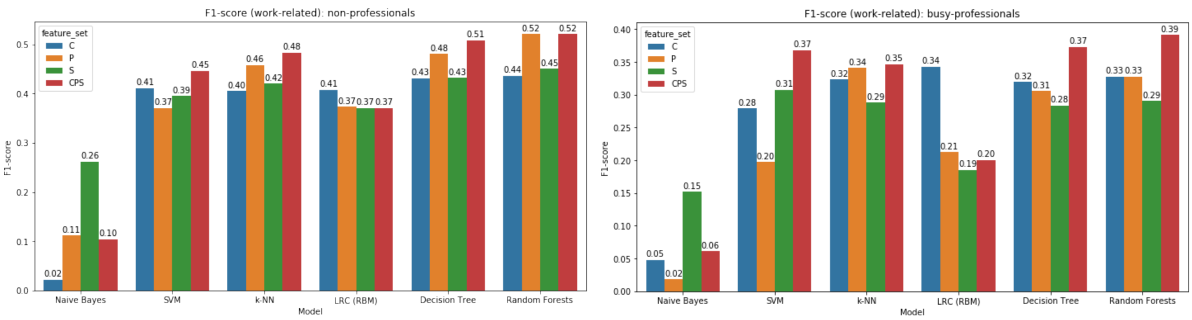



Figure 5.14: F1-score of task recognition (work-related): non-professionals and busy-professionals.

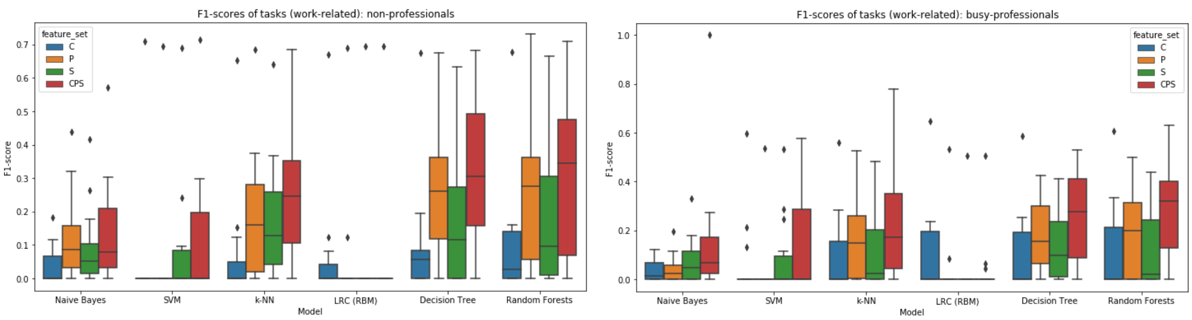



Figure 5.15: Boxplots of $F_1$-scores on work-related task recognition: non-professionals and busy-professionals.

Figure 5.14 shows the overall result of work-related task recognition based on the $F_1$-score (weighted average over all $F_1$-scores of tasks). The RF model for the non-professionals cohort achieves the best classifier performance (with $F_1$-score of 52.06%). On the other hand, the RF model is also suggested as the best classifier (with $F_1$-score of 39.13%) for task recognition. The best models for both cohorts are attained when they are trained on CPS feature set. Here, RF models are suggested based on the consideration of $F_1$-scores of all tasks (see Figure 5.15). Although the performance of the best classifier

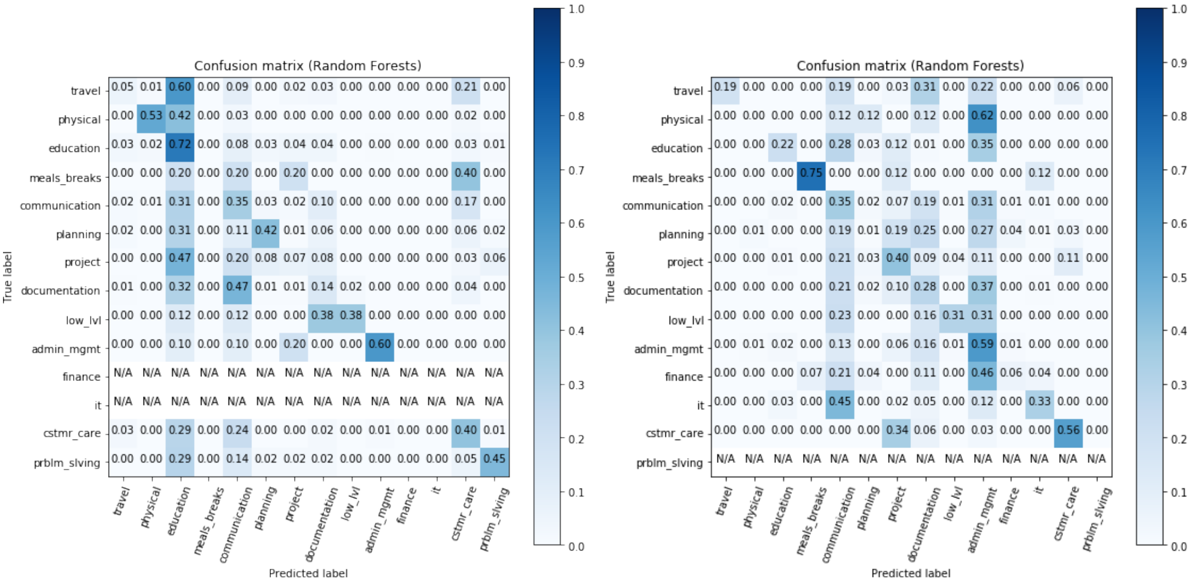



Figure 5.16: Confusion matrix of task recognition (work-related): non-professionals and busy-professionals.

for busy-professionals is relatively lower than non-professionals, this presents an on-going challenge for an intelligent assistant to optimise work-related task recognition. Moreover, the side-by-side confusion matrix is detailed in Figure 5.16, which respectively shows the exact proportion of prediction across all class labels (excluding "other" tasks) of RF models on non-professionals and busy-professionals.

Based on the misclassification we identify from the confusion matrix, we propose possible improvements of work-related task recognition for intelligent assistants on the following associated tasks under these conditions:

- Proportion of misclassifications on tasks is at least 10%, and
- Accuracy of classification for "true" class label (e.g., "Travel" task) is below 70%.

From our analysis on non-professionals, the majority of work-related tasks are often being misclassified as "education" task, despite the stratified cross-validation approach applied to the model. This result can be caused by the broad task category that is highly relevant to the professions of these non-professional participants, who are mainly students. On the other hand, physical tasks have a distinctively large proportion of instances being predicted as "admin/management" (62%), "documentation" (12%), "planning" (12%) or "communication" (12%) for busy-professionals. This result could be caused by limited physical activities being performed by typical office workers in daily life. For "education" related tasks, busy-professionals may require a certain environmental condition (e.g., similar to their office environment) in order to progress/complete such tasks, given the misclassifications of instances on "admin/management" (35%), "communication" (28%) and "project" (12%). Evidently, there are repeated patterns of misclassification between all these tasks of "admin/management", "documentation" and "communication", which may explain the association of these tasks in a typical office setting. Grouping these tasks (depending on the cohort) could enhance the performance of work-related task recognition. However, it is a growing challenge for an intelligent assistant to be able to distinguish these tasks to seamlessly support the mobile users in carrying their everyday work-related tasks.

#### 5.5.6.2 Recognition of Social/exercise/entertainment Tasks

For social/exercise/entertainment tasks, the experiment result is validated on the recognition of the following tasks:

- **Non-professionals:** travel, physical, education, meals and breaks, communication and planning.
- **Busy-professionals:** travel, physical, education, meals and breaks, communication and IT (software- or hardware-related tasks).

From our observation over the result of five-fold cross-validation, RF model also has shown the best performance with an $F_1$-score of 36.99% for professionals. However, the DT model is suggested for busy-professionals since it achieves overall $F_1$-score of 56.44%, which outperforms the RF model by an absolute difference of 4.2%. The highest performance can still be achieved by these models based on the CPS feature set. This result suggests that social/exercise/entertainment tasks could be more predictable to work-related tasks for busy-professionals, and vice-versa for non-professionals.

By understanding the confusion matrix of tasks prediction on social/exercise/entertainment tasks, it is evident that such degradation of $F_1$-score for non-professionals are mainly affected by substantial misclassification of "education" (45%) and "meals/breaks" (42%) tasks as "communication". However, the misclassification of "meals/breaks" tasks as "communication" (13%) is substantially lower than "communication" tasks as "meals/breaks" (38%) for busy-professionals. High numbers of misclassification are also noticeable for travel and physical tasks for busy-professionals, where they are often being recognised as "meals/breaks" (23%-37%) and "communication" (17%-28%). Although the number of class labels for social/exercise/entertainment tasks is limited, it is still important for intelligent assistants to support mobile users in daily life. Certain types of professions may require high concentration on work-related tasks. In between those work-related tasks, the users may be engaged on non-work tasks that may affect user's activities or even the performance of their tasks on that day. It should be noted that the progression of their tasks can be affected by user situations in-the-wild.

#### 5.5.6.3 Recognition of Personal/caring/civil Tasks

For personal/caring/civil tasks, the experiment result is validated on the recognition of the following tasks:

- **Non-professionals:** travel, physical, education, meals and breaks, communication, planning, documentation, admin and management, IT (software- or hardware-related tasks) and problem solving.
- **Busy-professionals:** travel, physical, education, meals and breaks, communication, planning, documentation, admin and management and finance.

Our comprehensive evaluation revealed that RF model for busy-professionals achieves the highest classifier performance (with $F_1$-score of 51.19%) while RF is also selected as the best model for non-professionals (with $F_1$-score of 30.43%). Although RF models provide lower overall performance for non-professionals, a substantial improvement is still noticeable when the model is trained using all CPS features.

By understanding the confusion matrix of tasks prediction on personal/caring/civil tasks, it is evident that many of the tasks are misclassified as “travel”, especially for the following tasks: 1) “physical”, 2) “education”, 3) “meals/breaks”, 4) “communication”, 5) “planning”, 6) “documentation” and 7) “admin/management” for both non-professionals and busy-professionals. For non-professionals, an intelligent assistant would need deeper insights on how users can be supported on both “planning” and “documentation” tasks. It is generally known that non-professionals have issues with personal organisation tasks, especially with the fact that the majority of them are students. Nevertheless, an intelligent assistant should support more personal tasks because users could be more productive and have more time to spend on other complex work-related tasks.

## 5.6 Exploratory Insights

The evaluation in the previous section has shown a dominating performance of CPS task recognition in daily life. In order to understand how informative these CPS signals on intelligent task recognition, the performance evaluation alone is insufficient. This section presents additional feature analysis and exploratory insights on the CPS aspect of user perception in performing or completing a task at a given time.

### 5.6.1 Feature Importance for Task Recognition

By fitting sensing data altogether (i.e., $F_c$, $F_p$ and $F_s$) into the best classifiers (according to task categories), the feature importance scores can be computed. Once these scores are ranked in descending order, they can be used to infer which features play crucial roles for intelligent task recognition.

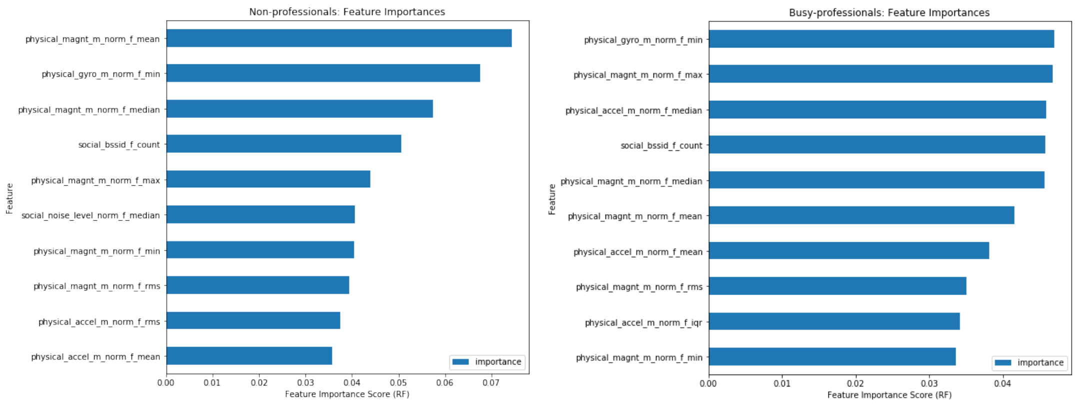



Figure 5.17: Top-10 features for work-related tasks: non-professionals and busy-professionals.

Figure 5.17 shows the example of top 10 features to recognise work-related tasks for non-professionals and busy professionals, respectively. The machine learning algorithm that performed best overall (random forests) has been used to compute the feature importance scores. Scores are ranked in descending order.

We can see that the count of unique access points "social_bssid_f_count" is always included in the top 10 features in most experiment sets (except social/exercise/entertainment tasks for non-professionals). At least 80–90% of top 10 features for task recognition are in fact, dominated by physical features. For busy-professionals, the analysis shows that the recognition of social/exercise/entertainment tasks is mainly affected by the importance of whether a user is engaged in "Business", "Design & Composition" or "Communication & Scheduling" cyber activities in the previous hour, and is currently engaged in "Shopping" based cyber activities. This is aligned with the results reported in Section 5.5.6.2, where the predictability for social/exercise/entertainment tasks is higher than work-related tasks on busy-professionals. In fact, the presence of cyber activities does not emerge in top-10 features for work-related task recognition on both cohorts. Our experiment also shows that no cyber activities are included in top-10 features for both cohorts to recognise personal/caring/civil tasks. This could be aligned with substantial needs for physical and social activities in carrying the tasks, such as “preparing myself to go to work (personal care or organisation)” or “taking care of my nephew”.

### 5.6.2 Participant Perception of CPS Activities

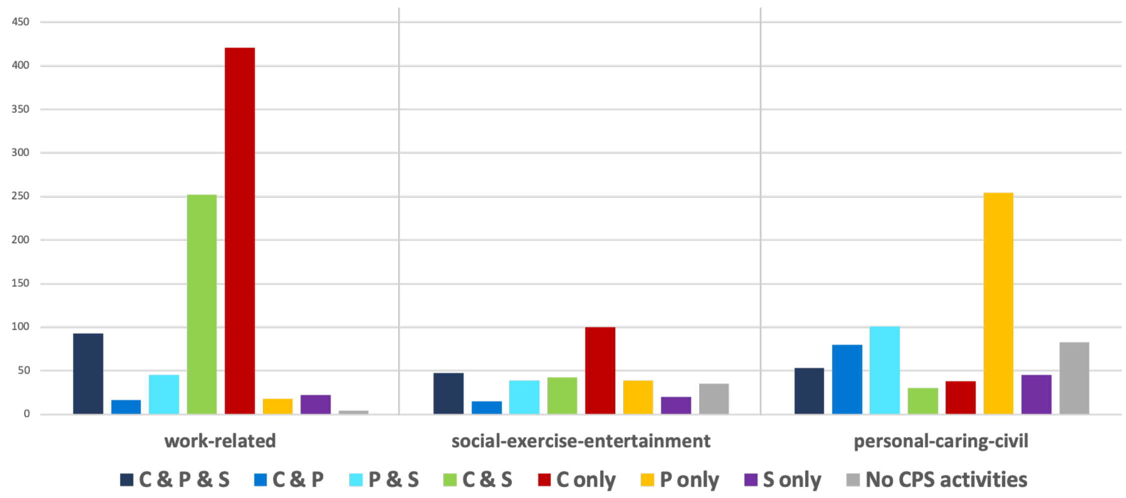



Figure 5.18: Overall participant perception of CPS activities on their tasks.

Figure 5.18 shows the general overview of participant perceptions of CPS activities on their engaged tasks, reported in the hourly surveys. From the survey answers, it is evident that users perceive that the majority of work-related tasks (up to 673 tasks) require either cyber activities only or both cyber and social activities. At least 93 tasks were reported requiring all CPS activities in order to progress or complete the work-related tasks. For social/exercise/entertainment-related tasks, the majority of tasks also required cyber activities only. Examples for this set of tasks would be “watching/browsing youtube videos”, “read online news”, “watching a tv series on Netflix”, “listening to podcast” and “listening to music”.

On the other hand, such social/exercise/entertainment-related tasks that users perceive to require all CPS activities can be complex and more abstract by their nature, for example: “shopping for swimwear in the CBD”, “Sport, fitness and calling with a friend” and “relaxing at home”. For personal/caring/civil tasks, a total of 254 tasks were reported as majority tasks that required physical activities only. However, the lowest count of 30 tasks was reported to require both cyber and social activities, followed by 38 tasks reported to require cyber activities only. In both cases, the participants may perceive less needs for physical activities to complete personal-related tasks such as “making a phone call to parents”, “asking a friend to follow up with a certain application (administrative related)”, “market research”, “checking emails”, “waiting for the train” and “personal admin tasks at home”. Interestingly, participants reported that any of CPS activities may not be needed for 83 personal/caring/civil, 35 social/exercise/entertainment tasks and four work-related tasks. These tasks can be subjective depending on the in-situ mental state of the participants, which explains the higher concentration of the “other” category of tasks identified, followed by education, meals/breaks, travel, and planning-related tasks.

### 5.6.3 CPS and Participant Perception Models

As described previously, users have their own perception on the needs of CPS activities in progressing or completing their tasks. It is natural to wonder whether the type of signals indicated by participants may be useful to select which set of features to use to feed a machine learning model. Hence, we extend our exploration by comparing models that use the set of signals (i.e., a combination of C, P, and S signals) based on participant perception against the combination of features that performs best overall: using all CPS signals. Based on the needs of CPS activities reported/perceived by the participants, the combination of binary answers (i.e.,“YES”/“NO”) can be used to construct participant perception models. There are six possible combinations of feature sets to construct a participant perception model, consisting of: C & P, P & S, C & S, C, P, and S.

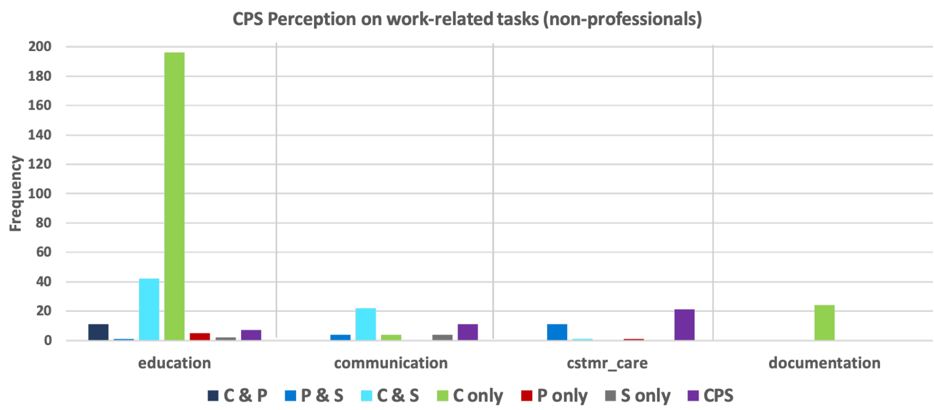



Figure 5.19: Perception of CPS activities on top 4 work-related tasks (non-professionals).

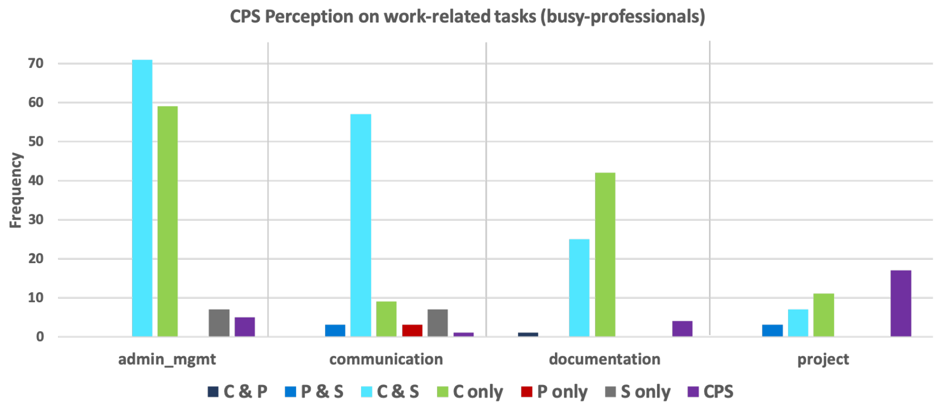



Figure 5.20: Perception of CPS activities on top 4 work-related tasks (busy-professionals).

Figures 5.19 and 5.20 show the overview of participant perception on the needs of CPS activities for top four work-related tasks (ranked by the frequency of tasks for each cohort). For each task type, we create the baseline task recognition model, which selection of feature sets is decided by the mode of participant perception inputs. For example, the recognition model for admin/management tasks would be built by taking in C and S feature sets.

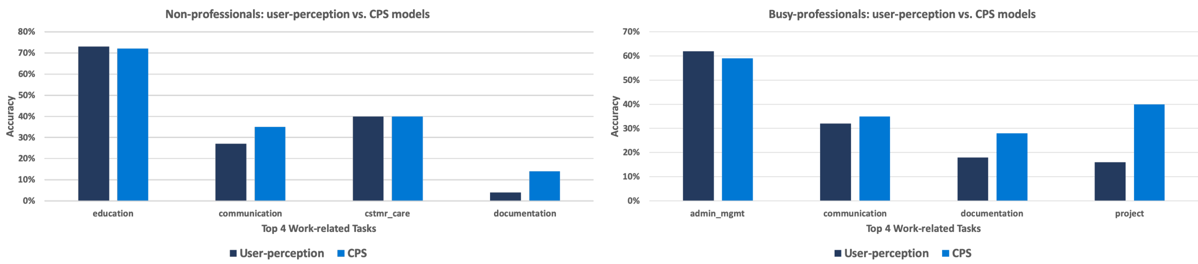



Figure 5.21: User-perception vs. CPS models: top-4 work-related tasks.

Our analysis reveals the following accuracy comparisons on the recognition of top-four work-related tasks (shown in Figure 5.21):

1. Non-professionals.
    - Education (CPS model: 72% vs. C model: 73%)
    - Communication (CPS model: 35% vs. C&S model: 27%)
    - Customer care (CPS model: 40% vs. P&S model: 40%)
    - Documentation (CPS model: 14% vs. C model: 4%)

2. Busy-professionals.
    - Admin/management (CPS model: 59% vs. C&S model: 62%)
    - Communication (CPS model: 35% vs. C&S model: 32%)
    - Documentation (CPS model: 28% vs. C model: 18%)
    - Project (CPS model: 40% vs. C model: 16%)

Based on the above comparisons, the participant perception model slightly outperforms the top tasks for both cohorts. The greatest improvement by 3% can be noticed on admin/management recognition, where the model is constructed based on the C & S feature set. In most cases, no substantial improvement is prominent in these comparisons. In fact, some features sets informed by the participant perception (especially when the amount of annotations is low) may lead to lower performance than the model that uses all CPS signals. In summary, all CPS signals seem to be informative for recognising all types of tasks. Our analysis also suggests that the perception of CPS signals reported by participants may be subjective and not necessarily accurate when used to select features for task recognition models.

## 5.7 Implications and Limitations

The development of accurate task recognition methods such as those described in this chapter has several implications for the design of intelligent and assistive technologies. By being able to reliably recognise tasks from users' CPS activities, a range of supports can be provided to users as the task is ongoing, e.g.,

- *Recommend resources or skills/actions that are relevant and/or required for task completion.* Figure 5.22 shows an overall distribution of application types started over a randomly selected task "documentation" for the busy professional participants in our study.

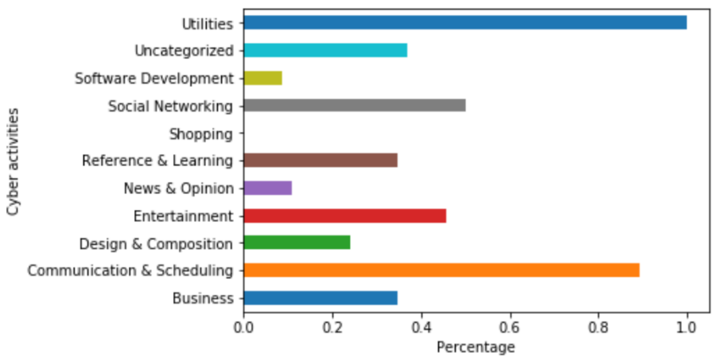



Figure 5.22: Top-10 features for work-related tasks: non-professionals and busy-professionals.

Note that this application types are automatically generated by the RescueTime app. Though the association between this particular task type and the applications is not guaranteed, some interesting patterns were observed which may suggest some personalised resource recommendation for task completion. For instance, a major portion of documentation tasks required 'entertainment' applications to be run concurrently. The future intelligent assistant could recommend appropriate entertainment contents based on the user profile which would enhance the work productivity.

- *Task progression measurement.* Task recognition can also be applied retrospectively after the task is performed. By understanding the context of a task could help measure task progression and task completion;
- *Task-related content recommendation.* Task recognition can be utilised to recommend contents that promote serendipity and information discovery related to the current task;
- *Guided tasks.* The accurate recognition of a particular task would offer direction on next steps (e.g., next task/sub-tasks).
- *Notification management.* Some tasks may require uninterrupted attention for an efficient completion. Task recognition could enable maintaining singular focus on the task at hand (e.g., to suppress notifications unrelated to the task);
- *Detect on-task/off-task behaviours.* Task recognition could further facilitate the understanding of on-task/off-task behaviours of users which may be affected by digital/physical intrusions (e.g., receiving e-mails/calls from clients, someone entering the office).
- *Task reporting.* Retrospective task recognition could be leveraged to generate summaries / periodic reports related to the completed tasks. It would be utilised to understand the tasks users perform in different situations and to generate contextual reminders if required.

It should be noted that the study described in this chapter is constrained by the time frame for data collection: over four consecutive weeks, only during working days, and from 7am to 7pm. Collecting data for a longer period, including weekends, and more time per day, would allow to analyse seasonal behaviour of tasks and activities, as well as a broader set of tasks (e.g., more social/exercise/entertainment tasks for busy professionals).

In reality, collecting longitudinal life sensing data involves significant challenges for ubiquitous computing research due to the natural course of human behaviours (e.g., disengagement). For instance, we had multiple cases of early abandonment during our data collection (e.g., participants dropping out after the first week). One area for future investigation is the cross-validation that we have performed, In our study, all instances are stratified for different partitions according to the discretisation of tasks, similar to the approach in the activity recognition study (e.g., [Liono et al., 2016, Abdallah et al., 2018, Liono et al., 2018a]). For the progression of tasks over time (e.g., weeks to months on a certain task related to a particular project), the stratification of instances based on the discretisation of tasks for different partitions may no longer valid, due to the potential loss of sequential information that may be crucial for recognising the task. Different validation strategies such as stratification by days per user may lead to more accurate recognition of tasks. Moreover, the current experimental results are limited to task recognition using the historic sensing logs collected from participants. Real-time task recognition may require sequential contextual information from the sensing data. The labelled sensing logs used in this chapter can aid automated analysis and labelling of future unlabelled data sets in an unsupervised manner. Eventually, this data could be used to analyse a broader variety of task behavior.

## 5.8 Conclusions and Future Work

Task recognition has many applications in digital assistants, productivity applications, and other intelligent and assistive technologies. This includes interruption support, task management, and generating task-relevant recommendations to help users make progress on their tasks. In this chapter, we presented methods to recognise user tasks based on sensing logs of CPS activities. We conducted a detailed analysis and showed that we can recognise tasks more accurately by fully utilising a range of CPS features.

Future work includes experimenting with richer feature sets, performing additional user studies (with more users, a broader portfolio of tasks, and different user cohorts), and integrating our task recognition methods into technologies to help boost user task efficiency and effectiveness, among other goals.

## 5.9 Acknowledgements

The research in this chapter was partially supported by Microsoft Research.

# Chapter 6

# Managing and Scaling Mobile Sensing and IoT Data with Quality-Driven Summarisation

## 6.1 Introduction

Previously, we have discussed progressive steps of our contributions from optimal windowing (for temporal segmentation), multi-context recognition, annotation prediction, to human task recognition. All of these chapters are presented based on the notion of in-the-wild mobile sensing. Consequently, we expand our contributions further to scale up mobile sensing data collection and experiment effectively. In this chapter, we formalise the problem of extracting compact representations of raw data streamed given the data quality measure that can be derived from multi-sources (i.e., large number of smart devices acting as "raw data providers").

Internet of Things (IoT) is an emerging area where everyday objects can be included as an integral component of the Internet and are able to communicate and share the knowledge between each other, and with users [Atzori et al., 2010, Georgakopoulos and Jayaraman, 2016]. In the IoT age, we envision that there exists a need for multiple data-as-a-service providers (we term these providers as data brokers such as [Perera et al., 2014]) who perform the collection, aggregation and sharing (for monetary benefits) of IoT data. Cloud computing has been used in the past to store IoT data. However, with the rapid increase of IoT device deployments and applications the cost of storage per unit can increase exponentially over time.

New techniques and approaches are required to cope with the growing amounts of data that are collected from IoT devices (and the supporting ecosystem) to serve the needs of various applications and users (e.g., data analytics applications that rely on IoT data to arrive at critical decisions). For example, in ubiquitous computing research, smart data-analytics applications rely on robust, yet compact machine learning models which are essential for the functioning of such applications. It is clear that providing unlimited storage for the raw data stemming from IoT in the cloud is impractical.

An important aspect of storage management of IoT data concerning such applications is the quality of data (QoD). Quality of data is defined as a metric that influences the performance (i.e., accuracy, recall and validity) of a machine learning model used by data analytics applications. Therefore, QoD is an essential factor to meet requirements such as accuracy of the machine learning and data mining models employed by the smart applications. According to the Data Warehousing Institute [Wayne, 2004], data quality problems result in estimated costs of $600 billion a year for US businesses. In addition, US economy receives significant impact that costs $3.1 trillion a year due to poor data quality within businesses and government [Wikibon, 2012]. However, determining QoD without prior knowledge of the application, nor any feedback from data consumers is a difficult challenge.

The problem of data quality (especially in the IoT domain) is challenging to address due to the noise and uncertainty induced by the large amounts of data produced by heterogeneous IoT devices (we term these devices that produce data as data providers). Moreover, the existing work (e.g., [Batini et al., 2009, Wang et al., 1993, Wang and Strong, 1996, Lee et al., 2002, Cai and Zhu, 2015, Zaveri et al., 2016]) on data quality often requires domain knowledge, annotations, or semantic labels. To the best of our knowledge, there is no study that has addressed the reliable measurement of data quality from large amounts of IoT data, regardless of their data types and application domains. Consequently, it is critical to address this issue of data quality to cope with the emerging problem of storing and managing IoT data in cloud.

In this chapter, to achieve the goals of developing a domain agnostic approach for QoD-aware storage management in IoT, we propose:

1. A formal model to represent data quality for data stemming from IoT;
2. A novel technique to define and compute the QoD without requiring any feedback from users of the data (we term them consumers) nor any prior knowledge of the application domain;
3. A smart data summarisation mechanism for managing storage based on computed quality of data (QoD);
4. A novel framework named QDaS that enables management of big data and incorporates the novel QoD estimation and data summarisation techniques.

To the best of our knowledge, this work is a pioneering effort that addresses the ability to perform storage management autonomously based on the quality of data inferred from the use of the acquired data (we call this data utility). Consequently, it requires no direct feedback and domain knowledge of the data consumers. Most importantly, our proposed system offers the capability to provide the indicative measurement that can be leveraged by data providers in improving the quality of their datasets.

The rest of this chapter is organised as follows. Section 6.2 provides the background and related work. This is followed by problem definition in Section 6.3. Section 6.4 describes QDaS, the domain agnostic framework for storage management. Section 6.5 presents a technique to generate data usages when the data utility is unavailable. Finally, Section 6.6 presents the outcome of extensive evaluations to validate the robustness and effectiveness of QDaS using real-world datasets.

## 6.2 Related Work

### 6.2.1 Data Collection from Internet of Things

Many research experiments such as Reality Mining [Eagle and Pentland, 2006], Mobile Data Challenge [Laurila et al., 2012], Device Analyzer [Wagner et al., 2014], and StudentLife [Wang et al., 2014a] exemplify analytics-based applications based on data sensed from IoT devices such as smartphones. The data collected from the IoT ecosystem (called data providers) would require large storage capacity in cloud repositories. Inherently, it needs to be easily accessed with high availability and be consumed seamlessly for further analysis by data consumers such as data analytics platforms.

Another inherently significant challenge with IoT data collected for data analytics applications is the difficulty in establishing the notion of quality of data, especially if they are provided by multiple IoT data providers. For example, consider a smart city scenario. There are many environmental and experimental conditions that can affect the quality of collected data, such as malfunction of sensors within the smart city networks and infrastructure. Moreover, the signal patterns that are captured by the ubiquitous sensors may change due to a particular environmental condition such as [Ye et al., 2014], which results in the needs for calibration and smoothing techniques in order to produce a better data quality. Consequently, these data quality issues are strongly related to the study of "veracity", one critical element of the Big Data concept [Assunção et al., 2015].

### 6.2.2 Quality of Services (QoS) and Quality of Data (QoD)

In service-oriented research such as [Hachem et al., 2014, Neiat et al., 2014], the focuses were directed towards composition and quality service deliveries that are often associated with certain measurements such as coverage, service time, certainty and freshness. Other related work on cloud computing such as [Beloglazov and Buyya, 2010, Nathuji et al., 2010] provided energy-efficient solutions while maintaining high QoS for cloud data centres. In [Wang et al., 2014b], a solution to evaluate the quality of cloud services accurately was proposed by using fuzzy logic control based on the uncertainty of services. However, this study is restricted to the performance of cloud service providers (i.e., service uncertainty), which is mainly measured based on the response time of service transactions. According to the basic QoS measure models proposed in 2007 [Ardagna and Pernici, 2007], data quality is included as one of the important aspects of web service composition. Unfortunately, there is a lack of study relating to data quality of service providers.

The notion of QoD is prevalent in the systems that provide Data as a Service (DaaS). Inherently, QoD is often associated with data privacy where data is exposed through web services based on certain criteria (in [Truong and Dustdar, 2010], they are called data concerns). At the theoretical level [Batini et al., 2009], the dimensions of data quality consist of the following basic components: accuracy, completeness, consistency and timeliness. In recent years, the dimensions of QoD have been expanded through the studies in many domains [Wang et al., 1993, Wang and Strong, 1996, Lee et al., 2002, Even and Shankaranarayanan, 2007, Lane et al., 2010, García-Recuero et al., 2013, Cai and Zhu, 2015, Zaveri et al., 2016] (e.g., availability, validity, integrity and relevancy). In IoT applications, it is important to ensure high QoD is maintained progressively according to certain Service Level Objective (SLO). In service-oriented computing, SLO is enabled through the Service Level Agreement (SLA) between service providers and consumers. Essentially, this concept is also applied to the distribution of data in DaaS applications, which can be described through provider-consumer interaction model [Wu et al., 2013]. In [García-Recuero et al., 2013], the fulfilment of high QoD is achieved through data replication across different geo-located clusters, which requires high resource availability for data storage. In this chapter, we are particularly concerned with the limited cloud storage space for data providers to retain continuous streaming of high quality of data from the IoTs.

Despite the comprehensive research to achieve high data quality from data collection tasks (e.g., [Ganti et al., 2011, Sherchan et al., 2012, Jayaraman et al., 2014, Kawajiri et al., 2014, Maharjan et al., 2016, Cardone et al., 2016, Hu et al., 2017], there is a lack of study on the estimation and maintenance of data quality based on contributions from the crowds. In this chapter, we envision that once the quality can be defined concretely, it can be enhanced further by motivating data providers to review their datasets.

To the best of our knowledge, there is no related work (including a utility model) that has addressed the contributions provided by data consumers that can be used to motivate data providers to enhance their data quality.

In the field of data management, Fan [2015] claimed data consistency, data deduplication, information completeness, data currency and data accuracy as foundations of data quality. However, service-oriented computing provides a different paradigm to deliver information based on a user's criteria, which focus on coverage, accuracy and freshness of application services. In fact, poor data quality often results in the degradation of predictive models that are produced by machine learning algorithms. In this chapter, we present a domain agnostic framework that can be used to assess data quality in order to perform quality driven summarisation of data for efficient storage management.

### 6.2.3 Data Trading and Resource Provisioning

In recent years, there have been many research efforts focusing on the policy design and governance of data trading [Iyilade and Vassileva, 2013, Aïmeur et al., 2016] to resource provisioning [Calheiros et al., 2011, Chaisiri et al., 2012, Laatikainen et al., 2014, Cheng et al., 2015] to retain these data. Most of these studies have focused on the area of cloud computing research given its benefits in terms of optimising the costs of ephemeral resources and preservation of user privacy for the given datasets.

In this domain, data trading is the core process that enables the collaboration between peers and knowledge sharing between data providers and data consumers. Consequently, the notion of integrity, security and privacy of data have to be established - a common theme in recent studies [Iyilade and Vassileva, 2013, Aïmeur et al., 2016]. Moreover, the emergence of the data marketplaces (such as [Cao et al., 2016]) has become increasingly significant due to the needs of data owners to share and sell their datasets. However, our study involves the design of data brokers as the general representations of data marketplaces or data exchange services. In a real-world scenario, data brokers may have limited resources to store data from the providers and provide reasonable QoS for consumers. Despite the related works in resource provisioning, there is a lack of study on storage management based on data quality that is estimated from data utility. In addition, a domain agnostic solution is required to manage diverse datasets that are supplied by data providers.

## 6.3 Problem Definition

The main challenge in this study is to define the Quality of Data (QoD) in a broader context, where the data can be shared collaboratively between the users and their data usage can be used as an indicative measurement of their usefulness. Therefore, we formalise the problems based on a traditional data provider-consumer interaction model. In general, data providers would require on-going subscriptions from the data consumers (government, data analytics platform, etc.). Since permanent storage is not feasible to retain all datasets supplied by the data providers, the main research challenges addressed in this chapter are:

- How to assess the quality of heterogeneous datasets provided by multiple IoT data sources?
  In this chapter, dataset refers a collection of data from a provider over a certain time period, for a specific application domain.
- How can we summarise the datasets for effective storage and management while maintaining a high quality of data?

### 6.3.1 Provider-Consumer Interaction Model

In a practical application of a provider-consumer interaction model [Wu et al., 2013], data is distributed by the providers through a data broker to meet consumer queries and requirements. The data broker is responsible for storing and managing the IoT data. There have been many research efforts in this area, including the exchange of data agreements [Truong et al., 2011] and the management of data stores [Sellami et al., 2016]. However, the related work on the quality of data itself have shortcomings. Therefore, the clear gap that can be found in this research is related to the value-based quality estimation. In this chapter, value-based quality estimation refers to the quality estimation that can be derived from the data consumers, which can also be beneficial to them. Consequently, the quality of data is essentially determined by the value it holds for the consumers. This can be measured by the frequency and the relevancy of data consumption from the consumers' point of view.

Figure 6.1 shows a simple representation of a responsible entity (namely a data broker) in distributing the data between providers $P$ and consumers $C$. In the proposed formal modelling of quality of data, we use the term datasets to represent IoT data to be used by data analytics applications.

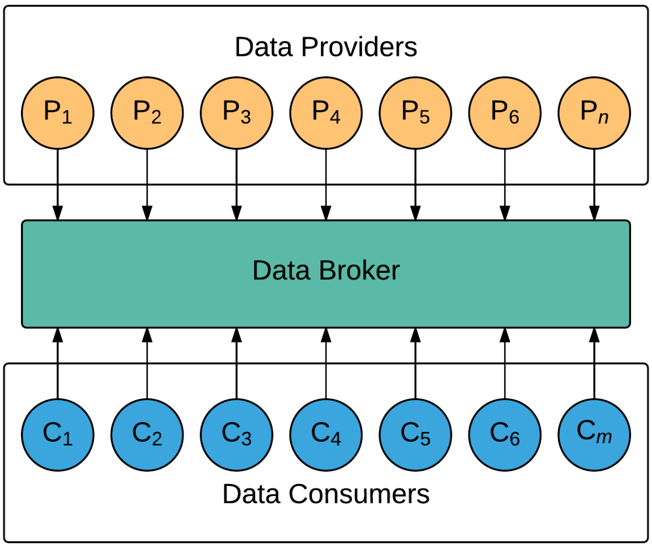



Figure 6.1: Data broker layer in provider-consumer interaction model.

### 6.3.2 Storage Management for Data Brokers

Let $P = \{P_1, P_2, ..., P_n\}$ be the set of providers with $n$ as the index of $P$ and $C = \{C_1, C_2, ..., P_m\}$ with $m$ as the index of $C$, $r$ is the index of $r$-th dataset provided by $P_n$ (i.e., $D_{nr}$). Each $D_{nr}$ contains the instances composed of $< t, F >$ where:

- $t$ is the unique timestamp of the corresponding instance.
- $F$ is the vector of $d$-dimensional feature values, where $F = \{f_1, f_2, ..., f_d\}$. In this chapter, we describe the attributes of a given dataset $D_{nr}$ (excluding $t$) as features. Each dataset $D_{nr}$ provided by $P_n$ may have different $d$ feature dimensions. In this case, the type of application, data and dataset dimension are not known a priori. Moreover, $D_{nr}$ may be updated by $P_n$ over a period of time and be accessible through data subscriptions of $C_m$ (refers to Figure 6.2).

In the given example of Figure 6.1, the data broker has several issues in terms of storing all the data in the given $D_{nr}$ and the capacity to deliver these datasets to $C_m$ due to storage limitation. In cloud computing, this can also be affected by external factors such as the budget (costs) for having the resources to store the data.

Since the storage is limited, the problem can be formalised to reduce the number of instances in the given dataset. In other words, compact representation of datasets should be stored instead of permanently retaining all the raw data given by $P_n$. Hence, storage management should be performed when a certain

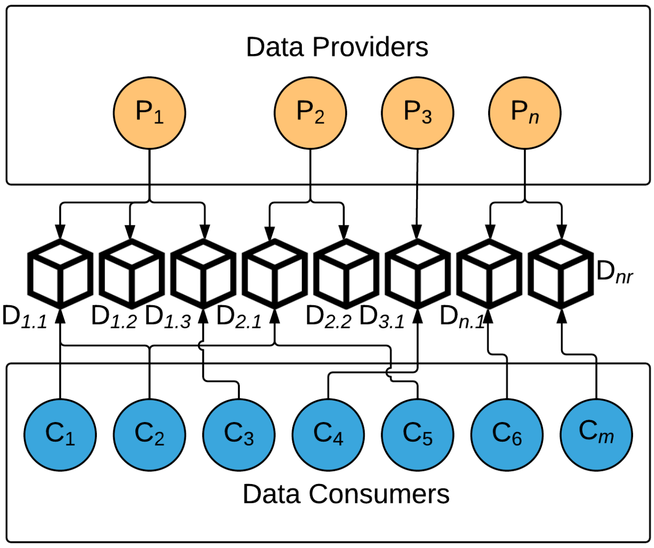



Figure 6.2: An example of data flow from data providers $P$ to data consumers $C$.

threshold $\alpha$ is reached in terms of storage capacity. To extract a compact representation of dataset instances for each $D_{nr}$, the quality of data needs to be measured.

Essentially, quality of data can be inferred based on the intelligence that can be acquired from the data consumers. In other words, the value of a dataset can be determined based on the utilisation by $C_m$. Inherently, high quality of data can be inferred as the product of high data utility. In short, frequent access/usage of data results in a higher utility value.

## 6.4 QDaS: a QoD-driven Framework for IoT Storage Management

In a typical IoT application, the data streamed from the real world are often found with noise and inconsistencies. The variability of data is dominant and thus the application domains, feature dimension, data size and data predictability are unknown in a priori. Therefore, there is a significant need to provide a smart solution for storage management in the cloud, since there can be constraints in terms of operational costs and availability of storage resources. In this chapter, we seek to solve this problem by selectively summarising data based on the notion of estimated data quality which can be inherently derived from the consumption of corresponding datasets. To address these challenges, our QDaS framework incorporates a mechanism for smart data summarisation based on QoD that are derived from usage patterns of the data.

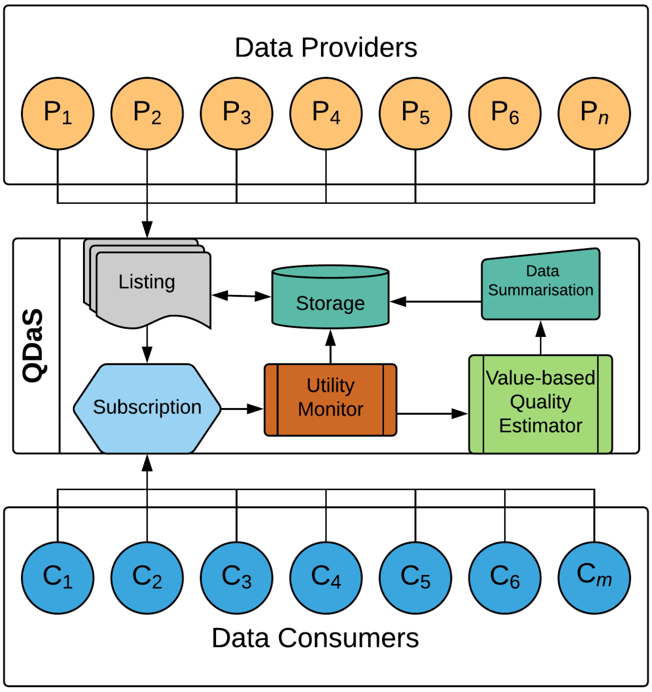



Figure 6.3: General architecture of QDaS.

The general architecture of QDaS is depicted in Figure 6.3. It is designed to be domain agnostic and hence does not require direct feedback or additional metadata from data consumers. Instead, QDaS assumes a typical setting where data consumers need to have continuous subscriptions to the datasets given by the data providers. The details of QoD estimation and technique for smart data summarisation will be elaborated in the following subsections.

### 6.4.1 Quality of Data Estimation

#### 6.4.1.1 Quality Definition based on Data Usages

The utility monitor component (as shown in Figure 6.3) in QDaS should capture the usage metrics (such as frequency of download/access) of subscribers (data consumers) for each dataset over the time. These metrics are then used as indicative measures to define the quality of data of a given dataset $D_{nr}$. In short, the quality of a dataset is defined based on the data utility (i.e., data usage) within a certain temporal duration (i.e., from a start time $t_{start}$ to end time $t_{end}$ as shown in Figure 6.4).

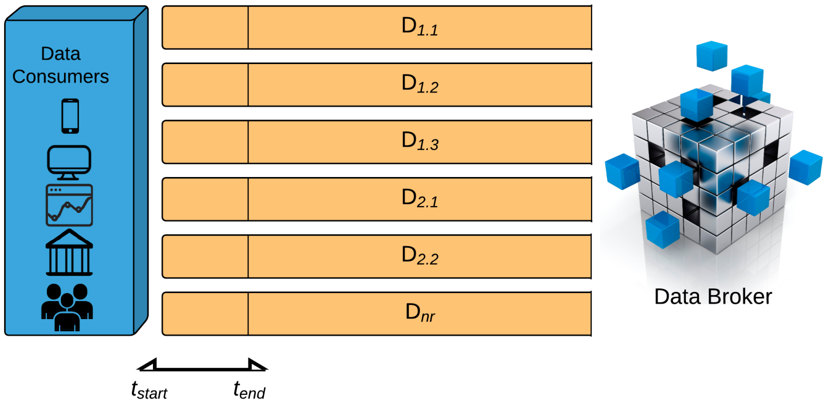



Figure 6.4: Data slice for quality estimation and data summarisation.

In this case, let us define the frequency of usage as a metric for data utility. For every non-empty data slice of $D_{nr}$ from $t_{start}$ to $t_{end}$ (referred as $Sliced_{D_i}$), the quality of data $Q_{utlty}$ of $Sliced_{D_i}$ can be computed as follows:

$$Q_{utlty} = usage(Sliced_{D_i}) / \sum_{j=1}^{b} usage(Sliced_{D_j}) \tag{6.1}$$

where $usage(X)$ is the usage function that produces measurement for data usage of data slice $X$, $b$ is the total number of all non-empty data slices (denoted as $Sliced_{D_{all}}$), $Sliced_{D_i} \in Sliced_{D_{all}}$. $Sliced_{D_{all}}$ is the list of data slices that is derived from all datasets $D_{nr}$ having $t$ of its data point within the range of $t_{start} \leq t \leq t_{end}$. In short, the numerator in Equation 6.1 is the usage of data slice extracted from the target dataset for quality estimation, while the denominator is the sum of usages for all data slices obtained from all datasets of the listed providers. In Figure 6.3, the process to compute $Q_{utlty}$ refers to **Value-based Quality Estimator** component.

#### 6.4.1.2 Categorisation of QoD

Once the quality of data utility $Q_{utlty}$ is computed for each data slice, they can be grouped according to their quality measures. In this case, the categorisation of high and low QoD can be established by simply setting a threshold $\delta$, which can be derived by a given function for all $Q_{utlty}$ of non-empty data slices as a data-driven approach. In our case, the median function is used to compute $\delta$, where $\forall\ Q_{utlty} > 0$. Hence, the categories can be divided into the following:

- **Low Quality** $LowQ_{utlty}$ where $min(Q_{utlty}) \leq LowQ_{utlty} \leq \delta$
- **High Quality** $HighQ_{utlty}$ where $\delta < HighQ_{utlty} \leq max(Q_{utlty})$

It should be noted that the categorisation of QoD is crucial due to fact that data slices are likely to be summarised if they belong to $LowQ_{utlty}$ category.

### 6.4.2 Coin Toss Test for Data Summarisation

In this section, we explain the process that we included before the actual execution of data summarisation. All data slices will have chances to be selected for summarisation. However, the decision to summarise a given data slice should be determined by a random chance (with a certain probability) that is influenced by its data quality (i.e., $Q_{utlty}$). Hence, we propose a generative approach (refers to Section 6.4.2.2) that can be integrated with the coin toss process. Consequently, the result of the coin toss process will determine if the data slice needs to be summarised. The result of a coin toss is either **Head** or **Tail**. In this case, no summarisation would be performed on a given data slice if the result of a coin toss is **Head**. Hence, the probability of getting a **Head** (denoted as $P(M)$) is redefined based on the following approach.

#### 6.4.2.1 Chinese Restaurant Process (CRP)

CRP is a stochastic process that is derived from the concept of customer seating allocation to potentially infinite number of tables [Aldous, 1985]. In the machine learning research, it is often associated with many applications of nonparametric bayesian methods (e.g., Dirichlet process [Ferguson, 1973]). The overall intuition is driven where the customer will likely to sit at a table with more customers in comparison to the tables that have fewer customers. In CRP, a concentration parameter $\alpha$ is used to control the computation of probability for each table. Therefore, first customer will be allocated to the first table with probability $\frac{\alpha}{\alpha} = 1$. Subsequently, the following customer $c_i$ will be either allocated to an empty table with probability $\frac{\alpha}{i-1+\alpha}$ or occupied table with probability $\frac{n}{i-1+\alpha}$. In this context, $n$ refers to the number of customers that have already been allocated to the corresponding occupied table.

#### 6.4.2.2 Monte Carlo Chinese Restaurant Process (MC-CRP)

In the Monte Carlo approach, the simplest form of the test can be performed where the probability $P(M)$ is derived from the repeated $n_{trial}$ number of trials, which is also based on the quality $Q_{utlty}$ of a data slice. Hence, we propose to incorporate CRP in every trial where each table is labelled as **Head** with the probability $P(H)$. In this chapter, we call this method as Monte Carlo Chinese Restaurant Process (MC-CRP).

$P(H)$ is derived based on the proportion of the region (category of QoD in Section 6.4.1.2) that $Q_{utlty}$ belongs to.

If $Q_{utlty}$ of $Sliced_{D_{nr}}$ is $LowQ_{utlty}$, then

$$P(H) = \frac{median(Q_{utlty}) - min(Q_{utlty})}{MAX_Q - min(Q_{utlty})} \tag{6.2}$$

where $MAX_Q$ is the upper bound of QoD. $MAX_Q$ is defined as:

$$MAX_Q = \begin{cases} max(Q_{utlty}) + \gamma, & \text{if } (max(Q_{utlty}) + \gamma) < 1 \\ 1, & \text{otherwise} \end{cases} \tag{6.3}$$

where $\gamma$ is an additional parameter introduced to control the output $MAX_Q$.

If $Q_{utlty}$ of $Sliced_{D_{nr}}$ is $HighQ_{utlty}$, then

$$P(H) = 1 - \frac{median(Q_{utlty}) - min(Q_{utlty})}{MAX_Q - min(Q_{utlty})} \tag{6.4}$$

From our observations, high $P(H)$ resulted in a higher probability $P(M)$ from repeated trials of Monte Carlo simulation. On the other hand, low $P(H)$ resulted in a lower probability $P(M)$. In this case, $P(H)$ signifies the survivability of $Sliced_{D_{nr}}$. Hence, a result of **Tail** will let $Sliced_{D_{nr}}$ to be summarised, given a biased chance of having **Head** with probability $P(M)$.

### 6.4.3 Smart Data Summarisation Technique

This subsection will discuss the technique of **Smart Data Summarisation** employed by QDaS. One algorithm of the smart data summarisation is described below: Algorithm 2.

**Algorithm 2** Data Summarisation Algorithm

**procedure** $DS(D_{nr}, t_{start}, t_{end})$
  $count_{ori} \leftarrow Count(D_{nr})$ ▷ Original count of $D_{nr}$
  $Q_{utlty} \leftarrow QualityForDataUtility(D_{nr}, t_{start}, t_{end})$ ▷ refers to Equation 6.1
  $P_n \leftarrow GetDataProvider(D_{nr})$
  $Rules_{P_n} \leftarrow RetrieveRules(P_n)$ ▷ Retrieve the rules of data summarisation for $P_n$
  $Clus_R \leftarrow DC(D_{nr})$ ▷ Extract clusters using density based methods
  $Sliced_{D_{nr}} \leftarrow Pop(D_{nr}, t_{start}, t_{end})$ ▷ Data slicing on $D_{nr}$
  $R(Sliced_{D_{nr}}, Clus_R, Q_{utlty}, Rules_{P_n})$ ▷ Apply the summarisation function to data slice
  $D_{nr} \leftarrow Sliced_{D_{nr}} + D_{nr}$ ▷ Add data slice into $D_{nr}$
  $count_{prov} \leftarrow Count(D_{nr})$ ▷ Count the instances in $D_{nr}$ after data summarisation
  $SS = \frac{count_{ori} - count_{prov}}{count_{ori}}$ ▷ Calculate the ratio of space saving
  $UpdateDataBank(D_{nr}, SS)$
  **return** $D_{nr}$
**end procedure**

The input for data summarisation algorithm consists of three main parameters: $D_{nr}$, $t_{start}$ and $t_{end}$. These parameters are required as the data summarisation can be performed for a given data slice of $D_{nr}$ within the range of $t_{start}$ and $t_{end}$. Essentially, the input parameters are similar to the Section 6.4.1.

In fact, the estimation of QoD for the given data slice is an inclusive process in the early steps of the summarisation procedure (denoted as $DS$ procedure in Algorithm 2). For a given data slice of $D_{nr}$ denoted as $Sliced_{D_{nr}}$ (a portion that is taken out through the $Pop(...)$ operation), the data slice summarisation process referred as $R(...)$ requires the following items:

1. **Data slice** (denoted as $\boldsymbol{Sliced}_{D_{nr}}$) within the range of $t_{start}$ and $t_{end}$ from $D_{nr}$.
2. **Quality of data** (denoted as $\boldsymbol{Q}_{utlty}$), regarding the utility from data consumers' viewpoint.
3. **Rules** (denoted as $\boldsymbol{Rules}_{P_n}$) for data summarisation, pre-defined by the provider $P_n$ that is identified from $D_{nr}$.
4. **List of clusters** (denoted as $\boldsymbol{Clus}_R$), produced by density-based clustering process.

The procedure of data summarisation requires the initial computation of $\boldsymbol{Q}_{utlty}$, which refers to previous quality estimation in Equation 6.1. Subsequently, it is followed by the provider identification for the dataset and its $\boldsymbol{Rules}_{P_n}$, which are essentially pre-defined by the corresponding provider. In QDaS, the providers have to be aware that their data is subject to summarisation. Hence, the summarisation functions (rules) must be defined prior to publishing their data through QDaS's platform (i.e., a data broker).

Before slicing the portion of data out of $D_{nr}$ for data summarisation, clustering must be performed. In the QDaS framework, density-based clustering (e.g., DBSCAN) is recommended. Essentially, the clustering is required as it plays a significant role in transforming the data slice $\boldsymbol{Sliced}_{D_{nr}}$ into a new set of data representation. In this chapter, the transformation of these instances refers to data summarisation process according to the rules given by the dataset's provider. Hence, to perform the transformation of data representation, both quality (inferred from its usage) and natural grouping of the given data portion are required in our proposed framework.

The following Algorithm 3 shows the sequence of data slice summarisation process (referred as $R(...)$ function), which requires the set of parameters $Sliced_{D_{nr}}$, $Clus_R$, $Q_{utlty}$, and $Rules_{P_n}$.

**Algorithm 3** Summarisation of a Data Slice

```
1:  procedure R(Sliced_Dnr, Clus_R, Q_utlty, Rules_Pn)
2:      coin_test ← MonteCarloTest(Q_utlty)
3:      if coin_test is success then
4:          cluster ← None
5:          buf ← [ ]                                   ▷ Initialise empty time-interval buffer
6:          summarisedData ← [ ]
7:          for each instance in Sliced_Dnr do
8:              prevCluster ← cluster
9:              cluster ← IdentifyClusterOfInstance(instance, Clus_R)
10:             if cluster ≠ prevCluster and prevCluster is not None then
11:                 newInstance ← Summarise(buf, Rules_Pn)
12:                 summarisedData.append(newInstance)
13:                 buf ← [ ]
14:             end if
15:             previousCluster ← cluster
16:             buf.append(instance)
17:         end for
18:         if length(buf) > 0 then
19:             newInstance ← Summarise(buf, Rules_Pn)
20:             summarisedData.append(newInstance)
21:             buf ← [ ]
22:         end if
23:         if NeedsArchiving(Sliced_Dnr) then
24:             ExecuteArchivingProcess(Sliced_Dnr)
25:         end if
26:         Sliced_Dnr ← summarisedData
27:     end if
28: end procedure
```

Before justifying whether to produce a new data representation, a coin toss test (based on probability derived from Monte Carlo simulation) needs to be performed. Consequently, the data summarisation process will continue if the coin toss test is successful based on the intuition that low quality of data will have a high chance for data summarisation (refer to the approach in Section 6.4.2.2).

For every instance in $Sliced_{D_{nr}}$, its cluster needs to be identified. Hence, a buffer $buf$ is used to hold consecutive instances that belong to the same cluster. When the identification of cluster changes, the data summarisation is applied to all instances in the buffer according to the provider's rules $Rules_{P_n}$. In other words, the cluster change detected in Algorithm 3 determines the portion of data (indexed by timestamp $t$) in $Sliced_{D_{nr}}$ to be summarised since the last cluster change.

In a real world scenario, the raw data could still need to be stored for certain purposes (e.g., provenance). Therefore, we also introduce the notion of data archiving in QDaS, which is enabled after summarising $Sliced_{D_{nr}}$ (line 23–25 in Algorithm 3). Ultimately, this procedure will replace the original $Sliced_{D_{nr}}$ with a new data representation, which will be integrated back to $D_{nr}$.

#### 6.4.3.1 Provisioned Value for Data Providers

Since the number of instances for a dataset $D_{nr}$ can be reduced through the smart data summarisation process in QDaS, the ratio of space saving (referred as $SS$ in Algorithm 2) can be used as the indicative measure to motivate the data providers to review and improve the quality of their data. Essentially, $SS$ of a given dataset $D_{nr}$ can be computed via:

$$SS_{D_{nr}} = \frac{C(D_{nr}) - C(DS(D_{nr}, t_{start}, t_{end}))}{C(D_{nr})} \tag{6.5}$$

where $C(DS(D_{nr}, t_{start}, t_{end}))$ is the total count of instances in $D_{nr}$ after data summarisation and $C(D_{nr})$ is the total count of original instances in $D_{nr}$ (i.e., before data summarisation). In this case, greater $SS$ indicates worse result as the size of dataset $D_{nr}$ for a provider $P_n$ is greatly reduced. Hence, high $SS$ values should motivate data providers to review their datasets carefully before publishing them to data consumers.

#### 6.4.3.2 Reputation of Providers

Moreover, our framework offers the capability to estimate the quality of providers based on the space saving ratio of their data, as the immediate result of QDaS's data summarisation. The indicative measurement to be given to data consumers is referred as reputation. Hence, the reputation of a data provider $Reputation.P_n$ can be computed as the following:

$$Reputation.P_n = \frac{SS.P_n}{\sum_{i=1}^{b} SS.P_i} \tag{6.6}$$

for $b$ is the count of all providers $P$, $P_n \in P$ and $SS.P_n$ is the product of an aggregate function for all datasets of $P_n$. Essentially, $SS.P_n$ can be obtained by computing mean space saving ratio for $D_{nr}$ as shown in the following equation.

$$SS.P_n = \frac{\sum_{r=1}^{c} SS_{D_{nr}}}{c} \tag{6.7}$$

where $c$ is the total count of heterogeneous datasets provided by $P_n$ and $0 < SS_{D_{nr}} < 1$. It should be noted that $SS_{D_{nr}}$ should only be included in Equation 6.7 if it is non-zero. Consequently, a smaller $Reputation.P_n$ value indicates a better reputation for a data provider $P_n$. Therefore, the ranking of data providers should be performed based on the inversed order of the computed reputation. It should be noted that the value $SS.P_n$ corresponds to the reputation within the time range from $t_{start}$ to $t_{end}$ for data summarisation. Thus, the overall reputation of a provider within the restricted time period ($t_{min} \leq t \leq t_{max}$) would depend on the function provided in QDaS. For example, a data broker may define summation or moving average as the function to produce the measurement of providers' reputation (in terms of space saving ratio of data summarisation) for the data consumers.

## 6.5 IoT Data Usage Generation Technique Employed in QDaS

The prerequisite for the QDaS framework is to have data usages available before quality estimation and summarisation. However, the availability of such data usage is not common due to the fact that most current applications in the literature do not consider a data-as-a-service model. Hence, in order to validate the effectiveness of the proposed QDaS framework, in this section we present an IoT data usage estimation technique that is built on data generation approaches available in the literature. The proposed technique first computes the agreement (referred to as the inter-rater agreement in Section 6.5.1) between data consumers and data providers through an analytical mechanism, which can then be used for generating IoT data usages (Section 6.5.2).

### 6.5.1 Inter-rater Agreement for Data Usage Generation

In order to simulate the behaviour of data consumers for their data access, the QDaS framework allows the mechanism to generate data usages by leveraging the analytical models that can be derived from the providers' datasets. Given a fixed number of usages $N_u$, the task is to generate the usages for each dataset supplied by the providers. It should be noted that the data usage generation process corresponds to the act of QDaS performing smart data summarisation task within the time range of $t_{start}$ and $t_{end}$, $\forall\ D_{nr}$. The generation of data usage is based on the average inter-rater agreement that can be obtained through the following steps:

1. Generate $N_c$ numbers of data consumers $c_z$ for a given $D_{nr}$, where $N_c$ can be generated from $Poisson(\gamma)$ distribution.
2. Pick a portion of random data points from the $D_{nr}$ for each data consumer. For each portion, random noise (e.g., Gaussian noise) can be embedded. Consequently, this step allows each data consumer to build a consumer model (denoted as $Model_c$).
3. Prepare training $PTrain$ and testing $PTest$ portions from the $D_{nr}$.
4. Build a provider model (denoted as $Model_p$) from $PTrain$.
5. Perform test on $PTest$ for the $Model_p$ and all $Model_c$.
6. Compute inter-rater agreement between each $Model_c$ and $Model_p$.
7. Compute the average of inter-rater agreement values from the previous step.

The above steps are executed in every cycle of storage management. It should be noted that the terms "iteration" and "cycle" are being used interchangeably in this chapter. Every cycle of storage management corresponds to the non-overlapping sliding window for a fixed time interval of $t_{start}$ and $t_{end}$ (within the temporal domain). Consequently, all generated models $Model_c$ in the first cycle can be reused subsequent cycles of storage management. Hence, the following approaches can be used to calculate inter-rater agreement between $Model_p$ and each $Model_c$.

1. **Correctness based generation approach**

   In 1960, Cohen [Kohen, 1960] introduced a method for measuring inter-rater agreement to test the reliability of data. With this measurement, insights about data reliability can be attained to the extent that the representation of variables correctly meets expectations [McHugh, 2012]. In many studies, the kappa statistic has been used globally for reliability measurement.

   The standard Cohen's kappa coefficient can be attained through the following formula:

   $$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6.8}$$

   where $P_o$ is the relative observed agreement amongst raters(i.e., probability of observation) and $P_e$ is the probability of expected agreement.

   In QDaS, this approach is typically used for supervised learning. Inherently, the inter-rater agreement $IRA_\kappa$ between $Model_p$ and $Model_c$ can be calculated through the following:

   $$IRA_\kappa = \frac{Eval_o - Eval_e}{1 - Eval_e} \tag{6.9}$$

   where $Eval_o$ corresponds to the observed evaluation metric for an analytical model $Model_c$, $Eval_e$ is the expected evaluation metric produced by the testing phase of $Model_p$. Both $Eval_o$ and $Eval_e$ should be normalized to a ratio value $r$ where $0 \leq r \leq 1$.

   This measurement can be used for an application that requires a categorical vector to be predicted, such as classification. In this case, the applicable correctness based quality metrics for both $Eval_o$ and $Eval_e$ are: accuracy, precision, recall, F1 score, etc.

   In this case, the boundary of inter-rater agreement ranges between the worst value $IRA_{\kappa\_worst}$ where $Eval_o = 0$ and optimal value $IRA_{\kappa\_best}$ where $Eval_o = 1$. Hence, the quantification of inter-rater agreement $IRA_c$ can be finally computed as follows:

   $$IRA_c = \frac{IRA_\kappa - IRA_{\kappa\_worst}}{IRA_{\kappa\_best} - IRA_{\kappa\_worst}} \tag{6.10}$$

where $IRA_c$ is within a value range of $[0,1]$. Ultimately, the final inter-rater agreement $IRA_{D_{nr}}$ (for a given dataset $D_{nr}$) can be inferred by averaging all $IRA_c$ values. In real world scenario, there is a condition where $Eval_e = 1$ as the testing result of $Model_p$. Therefore, the result of $IRA_\kappa$ would be undefined. In this case, the $IRA_c$ can be derived from the value of $Eval_o$ as it fits the assumption of how agreeable the estimation of $Eval_o$ towards $Eval_e = 1$. In addition, $IRA_c = 1$ is assumed when $Eval_o = Eval_e$.

2. **Deviation based generation approach**

   In several cases, annotations or ground-truth vectors can be in numerical form. In practice, this scenario is often be handled by a model built using regression methods. Deviation based generation approach leverages the intrinsic characteristic of how far a prediction deviates from the given analytical model. Hence, common data mining techniques aim to minimise the deviation, to assess better fit for the analytical model.

   In 1982, Koch [Koch, 1982] introduced a method called intraclass correlation coefficient (ICC) to measure the relative similarity of units that are organized in groups. The standard intraclass coefficient from two subjects can be attained through the following formula:

$$ICC = \frac{SD^2 - sd^2}{SD^2} \tag{6.11}$$

   where SD is the between-subject standard deviation and sd is the within-subject standard deviation (the typical or standard error of measurement). Hence, the inter-rater agreement for deviation based approach is defined as the following:

$$IRA_{icc} = \frac{(Eval_e)^2 - (Eval_o)^2}{(Eval_e)^2} \tag{6.12}$$

   where $Eval_o$ corresponds to the observed evaluation metric (deviation) for $model_c$, $Eval_e$ is the expected evaluation metric (deviation) produced by the testing phase of $model_p$.

   Therefore, the optimal value can be derived when $Eval_o$ is equal to $Eval_e$. For example, root mean squared error (RMSE) and root mean squared (RMS) are commonly used in regression analysis to validate its model. Therefore, the predictability of annotations for given data can be redefined as:

$$IRA_{icc} = \frac{(RMSE_e)^2 - (RMSE_o)^2}{(RMSE_e)^2} \tag{6.13}$$

   where $RMSE_o$ corresponds to testing of $Model_c$ and $RMSE_e$ is the expected deviation derived from testing of $Model_p$. On the contrary, $(RMSE_o)^2$ and $(RMSE_e)^2$ can be substituted with $MSE_o$ and $MSE_e$ for predictability measure of regression with respect to mean squared error.

The boundary of deviation scales from worst value $IRA_{icc\_worst}$ where $Eval_o = 2 * Eval_e$ to optimal value ($IRA_{icc\_best} = 0$) where $Eval_o = Eval_e$. It is assumed that both $Model_p$ and all $Model_c$ have non-zero deviation measures from testing phase. Consequently, the evaluation metrics should follow the rules defined in Equation 6.14.

$$IRA_{icc} = \begin{cases} IRA_{icc\_best}, & \text{if } Eval_o < Eval_e \\ IRA_{icc\_worst}, & \text{if } Eval_o > 2 * Eval_e \\ IRA_{icc} \text{ from Equation 6.12}, & \text{otherwise} \end{cases} \tag{6.14}$$

Hence, the quantification of inter-rater agreement $IRA_c$ can be finally computed as follows:

$$IRA_c = \frac{IRA_{icc} - IRA_{icc\_worst}}{IRA_{icc\_best} - IRA_{icc\_worst}} \tag{6.15}$$

where $IRA_c$ is within a value range of $[0, 1]$. Similar to the correctness based generation approach, the final inter-rater agreement $IRA_{D_{nr}}$ can be inferred by averaging all $IRA_c$ values. Moreover, the deviation based generation metric can be leveraged for unsupervised learning - for instance, the measure of reconstruction error for dimensionality reduction [Saul et al., 2006]. In a real-world scenario, the value of $Eval_e$ produced by $Model_p$ can be zero. Therefore, the result of $IRA_{icc}$ would be undefined. In this case, the $IRA_c$ can be set to zero where it is assumed that the expected model would produce zero deviation (i.e., $Eval_e = 0$).

### 6.5.2 Data Usage Generation Algorithm

Having the values of inter-rater agreement for each dataset, the list of datasets $D$ is constructed from $D_{nr}$ by descending order of $IRA_{D_{nr}}$. In this case, the first dataset will have the highest rank in $D$. Consequently, the first cycle of storage management will only use $IRA_{D_{nr}}$ for ranking. Subsequent cycles will leverage multiple criteria: inter-rater agreement measure $IRA_{D_{nr}}$ (from the current cycle) and the ratio of space saving $SS_{D_{nr}}$ (from the previous cycle of storage management). Essentially, the rank of each dataset can be obtained by following:

$$Rank_{D_{nr}} = \frac{IRA_{D_{nr}} + (1 - SS_{D_{nr}})}{2} \tag{6.16}$$

In order to generate the data usage, CRP is leveraged to reflect the behaviour in consuming data based on the popularity of each dataset. In this case, it is determined by the relevance of analytical models (of data consumers) to the given dataset from the provider and reputation of data summarisation (as a result of previous data summarisation). Consequently, each table in CRP represents a dataset that may be consumed. Therefore, the number of allocated customer in a table (CRP) represents the data access/usage of a dataset in QDaS.

The order of generated tables in CRP is aligned with the order of datasets in $D$. In other words, $D_{nr}$ will be used as the label for the tables according to the rank order. Since the number of tables in CRP is potentially infinite, it is feasible to have unlabelled tables. In this case, these tables can be labelled as $D_{nr}$ with the probability $IRA_{D_{nr}}$ as described in the following Algorithm 4.

**Algorithm 4** Data Usage Generation Algorithm

```
procedure GenerationProcedure
    for each table in unlabelledTables do
        isLabelled ← False
        while isLabelled is False do
            for each dataset in D do
                IRA_{D_nr} ← dataset.getIRA()
                isLabelled ← table.labelAsDataset(dataset, IRA_{D_nr})
                if isLabelled then
                    break;
                end if
            end for
        end while
    end for
end procedure
```

Finally, the usage for a given dataset $D_{nr}$ is aggregated via:

$$Usage_{D_{nr}} = \sum_{i=1}^{m} usage(table_i) \tag{6.17}$$

where $m$ is the number of tables with $D_{nr}$ label and $table_i$ corresponds to the labelled table (as $D_{nr}$) from CRP.

## 6.6 Experimental Evaluation of QDaS Using Datasets from real-world Case Studies

In this section, the evaluation of QDaS framework is presented, using open datasets such as the StudentLife dataset [Wang et al., 2014a] and the air quality dataset [Vito et al., 2008]. We use the correctness based usage generation approach for the StudentLife dataset and deviation based generation approach for air quality dataset [Vito et al., 2008]. In all the experiments, the analytical models are built using scikit-learn [Pedregosa et al., 2011], a well-known machine learning library for Python.

### 6.6.1 Activity Recognition - StudentLife

The StudentLife dataset was collected from students in Dartmouth College over ten weeks term to assess their mental health, academic performance and behavioural trends. However, the experiment for this case study only leverages the automatic inference of physical activity from the accelerometer sensor of the mobile phones.

#### 6.6.1.1 Data Preprocessing

The dataset contains summarised sensor data that are derived from 49 subjects (mobile phone users). The original authors of StudentLife [Wang et al., 2014a], Wang et al. extracted the 25 features from smartphone accelerometer sensor for activity inference. Therefore, these features (such as mean vertical acceleration, mean horizontal acceleration, etc.) are included in our experiments to predict the inferred activities (e.g., stationary, walking, running and unknown).

#### 6.6.1.2 Simulation for QDaS Framework (Student Life Activity Recognition)

In our study, each mobile sensing node (user) acts as a data provider, which delivers a dataset continuously to a data broker. Hence, each dataset could be utilised by a set of data consumers. Figure 6.5 shows a comprehensive and general workflow of our experiment settings and processes of QDaS simulation.

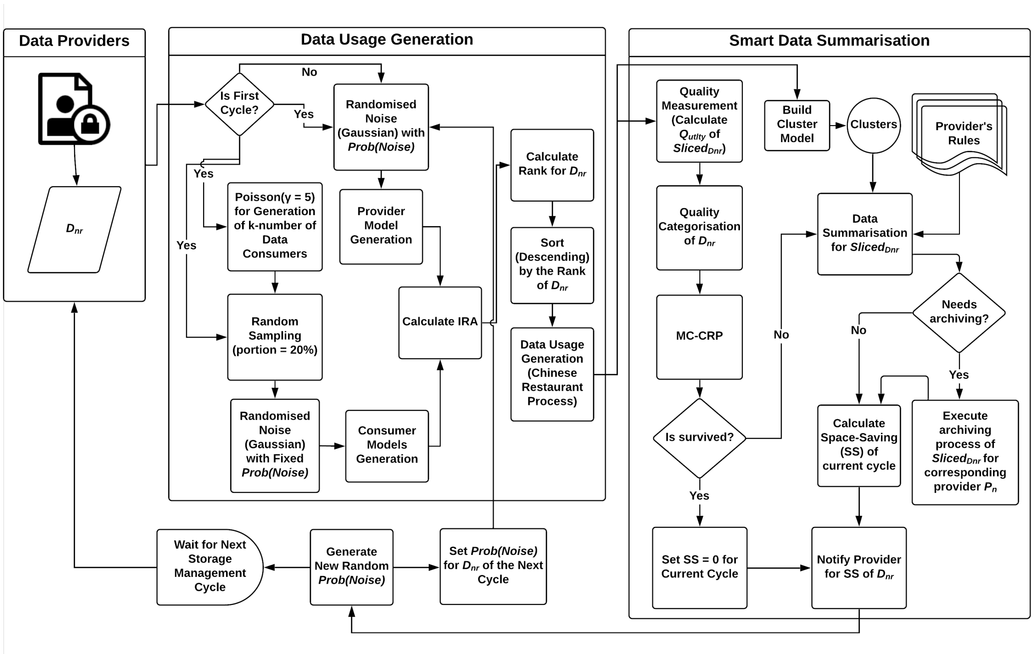



Figure 6.5: Workflow of storage management in the experiments.

In a given cycle, the range of temporal boundary for the data is restricted to 5 days in which the storage management would be performed for the first day. In short, sliding window model of 5 days is used to simulate the data arrival in next cycle and define the portion of data to be managed in a cycle (i.e., the first day of 5 days window). The summarised data from previous cycles are included in the process of provider model generation. In the simulation, 25 data providers are randomly selected from the StudentLife subjects (mobile phone users).

The first cycle of storage management requires the generation of data consumer models which are based on the given dataset $D_{nr}$. Before these models are built, a $k$-number data consumers is determined for a given dataset where $k$ is withdrawn from Poisson($\gamma$) distribution with $\gamma = 5$. For each data consumer, 20% portion of random samples are taken from $D_{nr}$, which features are infused with Gaussian noise given a fixed probability for an instance $Prob(Noise) = 0.3$, noise variance $variance_{noise} = 0.3 * variance_{original}$ where $variance_{original}$ is the variance of a feature.

The type of model for each data consumer is chosen randomly from these classifiers with default parameters: Naive Bayes (**NB**), Decision Tree (**DT**), Random Forests (**RF**), Multilayer Perceptron (**MLP**) and Support Vector Classifiers (**SVC**). On the other hand, the generation of provider model is based on the state-of-art algorithm, **RF** classifier.

The classification model of the provider is built based on train set of $D_{nr}$ in which each instance can be infused with Gaussian noise given an initial probability $0 \leq Prob(Noise) \leq 0.9$ which is defined randomly. In this case, the $variance_{noise}$ is similar to the setup used in the generation of data consumer models. It should be noted that $Prob(Noise)$ can vary in the consecutive cycles, which can affect the provider's decision to increase the quality of data for $D_{nr}$. Moreover, the train and test are randomly sampled applying the 60%-40% rule. Hence, the inter-rater agreement (IRA) measure can be computed from classification accuracy metrics between the provider's and consumer's models on the 40% of $D_{nr}$ (i.e., test set). Consequently, the generated consumer models would be utilised for the next cycle of storage management.

Once IRA is computed for $D_{nr}$, the Chinese Restaurant Process (CRP) is used to generate data usages for all ranked datasets of the data providers with the following parameters: $n$ = 10,000,000 and $\alpha = 1.0$ (refers to Section 6.5.2). Subsequently, the data usages that are generated from the previous step are then used for quality measurement of $Sliced_{D_{nr}}$. In this case, $Sliced_{D_{nr}}$ is extracted within the temporal boundary of the first day from the 5 days window. Hence, the data summarisation may be performed on $Sliced_{D_{nr}}$. The computation of quality measurement ($Q_{utlty}$) for $Sliced_{D_{nr}}$ shall not include the generated data usages that may have been allocated to other empty $Sliced_{D_{nr}}$.

The survivability of $Sliced_{D_{nr}}$ (i.e., not to be summarised) depends on the random chance of simulated MC-CRP process (refers to Section 6.4.2.2) where the number of Monte Carlo trial $n_{trial} = 1000$, and the parameters for CRP in each trial are: $n = 10{,}000$ and $\alpha = 1.0$.

To build the model of density based clustering (used for data summarisation), we leveraged DBSCAN algorithm with $Eps = 0.3$ and $min_{pts} = ln(n)$ (heuristic used in [Birant and Kut, 2007]) where $n$ is the number of instances in a given $D_{nr}$. It should be noted that the clustering process is only performed on the feature vectors. In our evaluation, we use k-means clustering (with default parameters) as the non-density based algorithm for the performance comparison. Consequently, the instances in $Sliced_{D_{nr}}$ can be associated with the generated cluster labels, which would be used for the purpose of data summarisation process. In terms of the provider rules, data summarisation complies with the following rules: latest time is picked, the average function is applied to all feature dimensions, and majority voting for the human activity vector (i.e., the most frequent activity label). After the process of data summarisation of $Sliced_{D_{nr}}$, the ratio of space saving (*SS*) can be computed and be given to the data provider, so that the provider may improve their quality of data (with less noise) in the next cycle. As a result, RR may control the new value $Prob(Noise)$ that is used to generate noise in next cycle (for the purpose of building provider's model). Moreover, we parallelised the processes for data summarisation on all ranked datasets in order to fully utilise CPU resources and speed up the runtime of our experiments. The implementation was programmed in Python, which runs in a platform with 2 CPUs of Intel Xeon 5607 (total 8 cores with frequency 2.27GHz), 48 GB RAM (12x4GB DDR ECC) and CentOS 7 operating system.

#### 6.6.1.3 Evaluation and Discussion

The simulation ranges to the total of 23 iterations (cycles) for QDaS data summarisation (denoted as *DS*). First, the effect of *DS* on the space saving ratio is observed between the usage of density-based clustering (DBSCAN) and non-density clustering (k-means). The evaluation is mainly focused on the summarised datasets which are mainly categorised as $LowQ_{utlty}$ (low quality of data according to the usage measures). As shown in Figure 6.6, the density plot reflects a distinct concentration of high space savings for density-based clustering (on summarised datasets), which leap closer to 100% space savings. On the other hand, non-density based clustering algorithm shows a significant difference in terms of lower space savings, which are condensed between 12% and 37%.

Inherently, higher space savings of the given datasets should stimulate their data providers to improve the quality of data which could lead to increased data usages. As shown in Figure 6.7, inter-rater agreement measures (mean $IRA_c$) between data providers and consumers reach a high concentration

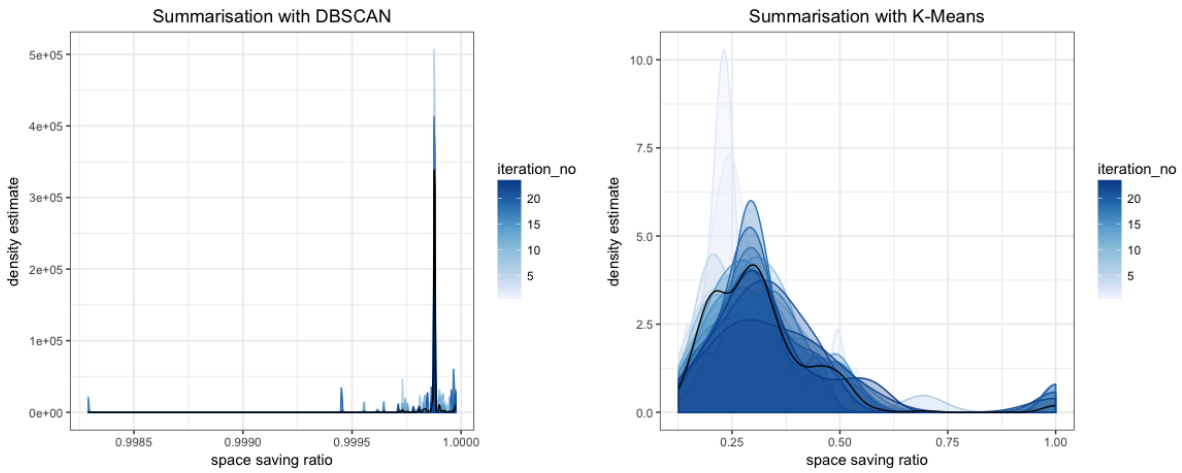



Figure 6.6: Density plot for space saving ratio (refers to Equation 6.5) of summarised $LowQ_{utlty}$ datasets.

between 55% and 75%. Although some datasets may have lower inter-rater agreement measures for density based clustering prior to *DS* process, the repeated measure of ANOVA test ($p = 0.06$) suggests no significant difference between clustering techniques (density and non-density based algorithms) for inter-rater measures. In other words, density based clustering proves to be an effective approach to produce a higher ratio of space saving in a *DS* cycle, and stable inter-rater agreement (derived from analytical models of data providers and consumers) is relatively maintained throughout the simulation cycles.

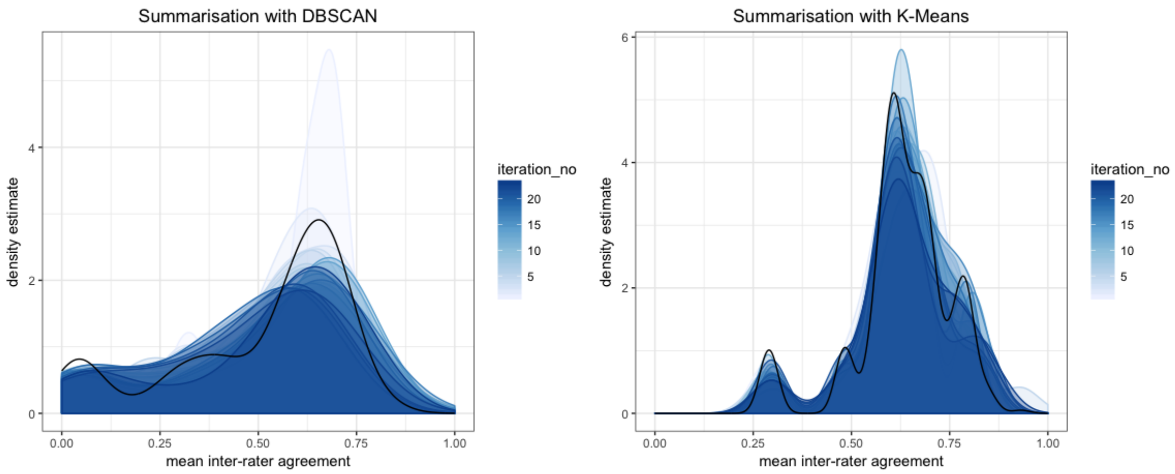



Figure 6.7: Density plot for mean $IRA_c$ (refers to Equation 6.10) of summarised $LowQ_{utlty}$ datasets.

### 6.6.2 Regression Analysis - Urban Air Quality

The air quality dataset was collected in an Italian city where the area is significantly polluted. All corresponding features of sensed data were generated by average responses of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The co-located reference certified analyzer also recorded five ground-truth measurements of air quality in the same area. In our study, the problem is addressed by using IoT devices which are significantly cheaper than deploying a more accurate and expensive static machine. In this case, the objective of a regression model is to predict the numerical values of ground-truth given the features of a ubiquitous sensing device.

#### 6.6.2.1 Data Preprocessing

The air quality dataset contains 9357 instances. It consists of 9 feature vectors (including timestamp) and 5 ground-truth vectors. In our experiment settings, we divided this dataset to 5 different datasets in order to simulate various data providers and data types. For each dataset, all the feature vectors are retained and only one ground truth vector is used for prediction. As a result, we set a fixed number of providers to 10 where each of them owns 5 datasets.

In total, 50 datasets will be used for the QDaS experiment in terms of data usage generation, ranking and data summarisation. In addition, we constructed new feature vectors by calculating the value differences between the current hour and previous hour for each feature dimension (excluding the timestamp). In other words, given feature value $f_i$ and previous feature value $f_{i-1}$, the difference $\Delta f$ is calculated by simple subtraction:

$$\Delta f = f_i - f_{i-1} \tag{6.18}$$

In short, the dimensions expand to the following features: $timestamp$, $\Delta CO$, $CO$, $\Delta NMHC$, $NMHC$, $\Delta NO_x$, $NO_x$, $\Delta NO_2$, $NO_2$, $\Delta O_3$, $O_3$, $\Delta Temperature$, $Temperature$, $\Delta RelativeHumidity$, $RelativeHumidity$, $\Delta AbsoluteHumidity$ and $AbsoluteHumidity$. For each dataset owned by the data provider, the predicted vector corresponds to one of these ground-truth vectors (captured by the reference certified analyzer): $CO$(GT), $NMHC$(GT), $C_6H_6$(GT), $NO_x$(GT) and $NO_2$(GT). Inherently, these experiments would be based on the objective to predict the measurements of air quality captured by the expensive and accurate sensors. Therefore, regression models are used for predicting the numerical values of air quality measurements.

#### 6.6.2.2 Simulation for QDaS Framework (Air Quality Prediction)

In terms of the experiment settings, the parameters used for QDaS simulation are similar to the previous case study (in Section 6.6.1.2). In a given cycle, the range of temporal boundary for the data is restricted to 20 days in which the storage management would be performed for the first 5-days. In short, a sliding window model of 20 days is used to simulate the data arrival in the next cycle and define the portion of data to be managed in a cycle (i.e., the first 5-days of 20 days window). The summarised data from previous cycles are included in the process of provider model generation.

The type of model for each data consumer is chosen randomly from these algorithms with default parameters: Linear Regression (**LN**), Decision Tree (**DT**), Random Forests (**RF**), Multilayer Perceptron (**MLP**) and Support Vector Regression (**SVR**). On the other hand, the generation of the provider model is constructed based on the **RF** algorithm. Moreover, the inter-rater agreement measure is calculated using the deviation based generation approach. In this case, root mean squared error (RMSE) is leveraged as the deviation metric.

#### 6.6.2.3 Evaluation and Discussion

In terms of the evaluation of the ratio of space saving and inter-rater agreement measures between density and non-density based clustering, the output of experimental results is similar to the previous case study. Figure 6.8 shows a significant difference between space saving ratio between DBSCAN and k-means. In other words, a more compact data representation is produced that leads to a better output of storage management for a data broker. Inherently, non-density based clustering leads to lower space saving ratio, with concentration skewed below 55% for later cycles of *DS*.

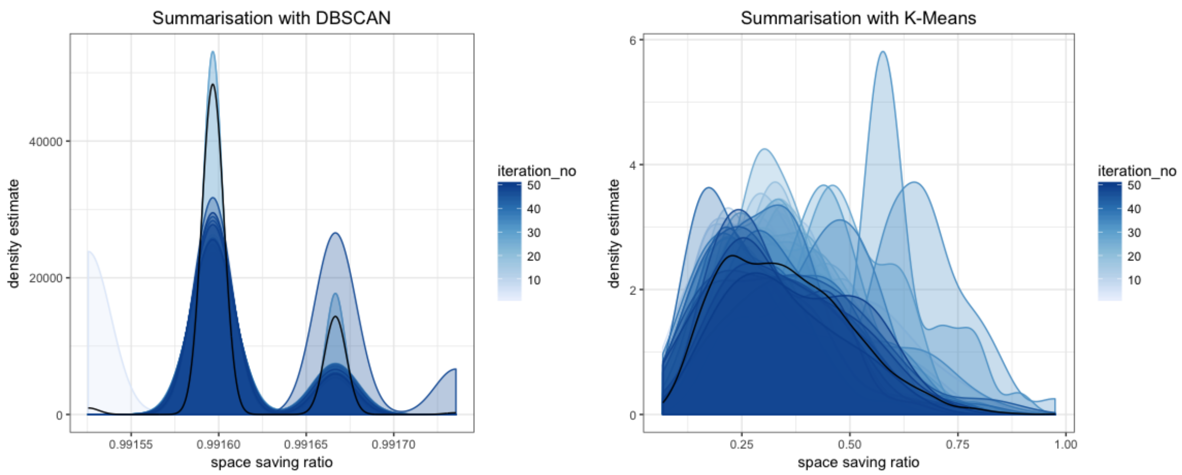



Figure 6.8: Density plot for space saving ratio (refers to Equation 6.5) of summarised $LowQ_{utlty}$ datasets.

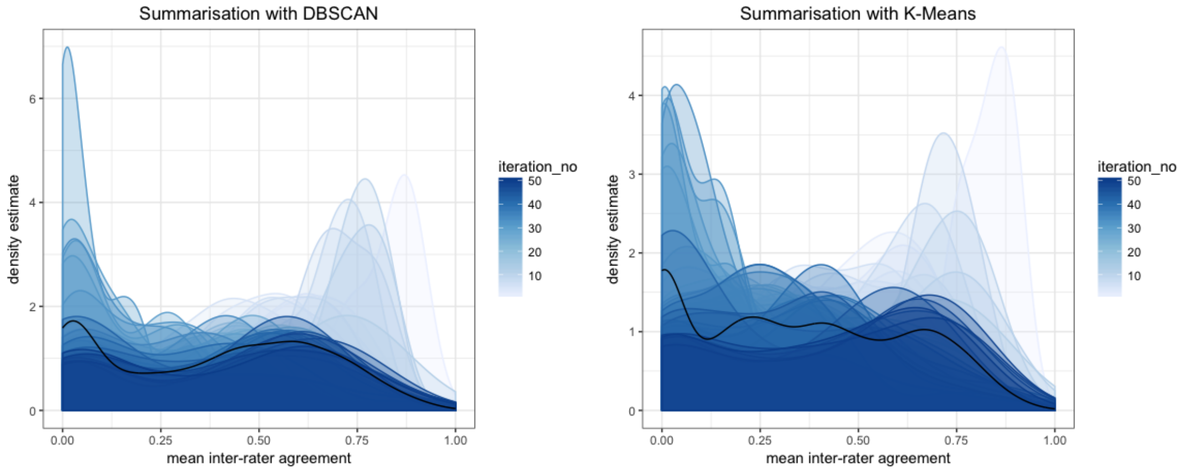



Figure 6.9: Density plot for mean $IRA_c$ (refers to Equation 6.15) of summarised $LowQ_{utlty}$ datasets.

In Figure 6.9, the density plot shows an even concentration of inter-rater agreement in later cycles of DS. This result suggests data providers are behaving in ways to revise and improve their quality of data over the time. Consequently, we applied the repeated measure of ANOVA test ($p = 0.54$) which indicates no significant difference between clustering techniques (density and non-density based algorithms) for inter-rater measures. Hence, the variation of $IRA_c$ distribution for summarised datasets may also be affected by many types of datasets that are owned by each data providers. In this case, the dimension of datasets (total of 50 heterogeneous ranked datasets) is significantly higher than the previous case study.

To conclude, our experiments suggest that the density-based clustering approach of data summarisation proves to be more effective for storage management, by producing high space saving ratio while maintaining stable inter-rater agreement between data providers and consumers.

## 6.7 Conclusion

In this chapter, we presented (QDaS), a novel domain agnostic framework for effective storage and management of IoT data in cloud. The proposed framework incorporated a novel technique to estimate the quality of data driven derived from the utility of the IoT data. Based on the quality estimation technique, QDaS employs a smart summarisation mechanism to manage the storage of IoT data in the cloud.

Extensive experimental evaluations were conducted in order to demonstrate the effectiveness of the proposed QDaS framework using real-world datasets. The results have shown that our framework provides a robust approach in performing efficient storage management, based on the notion of quality of data that is determined by the usage measures. We demonstrated experimentally that the amount of data stored in the cloud could be reduced as a result of the QoD-driven data summarisation mechanism. The output of this research presents promosing result that can be used for scaling mobile sensing in smart environments. In fact, the density-based data summarisation has been utilised to derive a more compact representation of sensor features and accelerated the overall learning process of our previous experiments (refer to Chapter 4).

# Chapter 7

# CONCLUSION

Situation awareness plays a significant role in intelligent systems and therefore has become a critical subject in multiple research disciplines over recent decades. In the realm of pervasive computing, understanding the situation of mobile users will enable the full capacity and capabilities of their smart devices to be more context-aware for their engaged activities and tasks wherever and whenever.

In this dissertation, we have taken a critical step closer towards situation awareness in ubiquitous computing research by novel contextual modelling approaches to mobile users based on their pervasive signals from mobile sensing. To conclude, current research on situation inference and context recognition from mobile sensing is growing. With the rapid development of smart devices in terms of their processing power and resources, it is believed that the future assistive technologies should enable more ubiquitous applications to be more aware of user situations and their contextual dependencies to offer proactive support in recommending resources or actions for enhanced cognitive augmentation (i.e., improved productivity and decision making).

The aim of the research was to build systematic frameworks that have intelligent predictive capabilities in situation inference, not only from the first-person view for multi-context recognition but also expanding our study towards a larger scale of mobile sensing experiments from the crowd. Hence, this result is achieved through incremental research efforts from contextual modelling on the collection of multivariate time series data that can be sampled in an irregular manner, with the risks of sensor deactivation during pervasive sensing in daily life. Moreover, the mobile data that are sourced by multiple sources can be easily influenced by the sensing noise within the environments and subjective user annotations under certain circumstances. Nonetheless, the benefit of this study can be reaped directly from our extensive experiments and their novel contributions, ranging from awareness and insights that can be gained by the mobile users on their environments, to better support for activity-based decision making and proactive resource recommendations enabled by the future intelligent assistive technologies. We again believe that this research is truly pivotal for more innovative development of ubiquitous applications given the exponential proliferation of internet-connected smart devices.

## 7.1 Research Questions and Answers

The core chapters in this thesis addressed the cumulative research challenges facing intelligent situation inference from in-the-wild mobile sensing. The issues that we have discussed so far are circulated around building intelligent assistants that can be situation-aware, given the traits of human activity, behaviours and mobility from pervasive signals sensed in daily life. To understand the key factors associated with situation inference and context recognition from mobile sensing, four distinct research questions were developed, as outlined in Section 1.3. In each core chapter, we validated our proposed frameworks, modelling approaches and novel algorithms by experimenting on real-world datasets and mobile sensing data collection. Therefore, each research question and its solution is summarised below.

- ***RQ-1.*** *How to find an optimal window size in processing multivariate sensor data for multi-context recognition?*

  The first research question was addressed in Chapter 2. We presented OPTWIN, a novel technique to provide a recommendation for window size to perform effective extraction of time-interval based features in the temporal segmentation process of continuous multivariate sensor data streaming. This technique leverages the proposed multi-objective function, which minimises the window impurity and maximises the measure of class separability between the window instances. For the measure of class separability, we use the metric computation of Kullback-Leibler (KL) divergence scores as mentioned in Section 2.4.2, to measure the difference between two probability distributions $P$ and $Q$ on numeric features (extracted from multivariate sensor data). In other domains, the class separability measure is often used to rank features, mainly for feature selection. On the other hand, we introduced an additional metric in our multi-objective technique, a notion of an impure window where there is at least a mixture of two subsequent contextual labels in a window during the temporal segmentation process. Such phenomenon that occurs in the sliding window model denotes the transition of user contexts in time series data when they are processed in a continuous manner. To improve the performance of context recognition, it was initially assumed that this overlapping issue of contextual labels should be minimised. Consequently, we extended this technique to adapt to the multi-context scenario, which is evaluated with the real-world dataset to perform multi-activity recognition in smart home environments. As a result, this technique improves the confidence in performing multi-context recognition in smart environments, while reducing the general effort in sensitivity analysis and fine-tuning model parameters in the learning phase.

- ***RQ-2.*** *How to perform multi-context recognition from multidimensional sensor data and user annotations?*

  The second research question is constituted by the interrelated issues that we addressed in Chapters 3 and 4. For mobile situation inference, recognition of human activities and contexts is extremely important, since the results of such processes can be used to describe and construct a particular user situation. Therefore, we aim to improve the general accuracy of multi-context recognition by contributing new modelling approaches and machine learning techniques, which require rigorous iterations of testing and validation against the typical ways of performing activity and context recognition.

  To address the second research question (RQ-2), we sub-divided our focus to two distinct issues. Firstly, the issue of simultaneous recognition of multiple mobile user contexts was addressed in Chapter 3. Since the combination of multiple user contexts (e.g., human activities and transportation modes) in Chapter 3 were the result of decomposition of raw user annotations, our research direction was then targeted to improving the predictability of these annotations. More research challenges were imposed particularly for Chapter 4 due to the fact that these raw user annotations could be acquired via the prevalent use of the experience sampling method in mobile sensor data collection. In many situations of each mobile user, these raw annotations could be subjective according to the mental state of the corresponding user.

  In Chapter 3, we developed a new modelling approach to tackle the issue of simultaneous inference of multiple user contexts for in-the-wild mobile sensing. In short, we proposed a new modelling approach (refer to Section 3.4.1) called Context-based Activity Recognition (CBAR). This novel modelling approach assumes that the user contexts can be described in a hierarchical form, and therefore should have tightly coupled relationships (i.e., dependencies) with the low-level labels of human activities. In this case, the typical approach of performing multi-context recognition on multiple prediction targets is to build independent classifiers on various label sets. Our proposed Mobile Context Recognition System (MCRS) utilises both multi-stage and multi-target inference concepts to perform a robust simultaneous recognition of human activity and transportation mode, including their associated environmental contexts accurately. Our results on iterative experiments showed general accuracy improvement, highlighted for Decision Tree based classifier with the most significant gain of 19.3% in terms of exact match on the predicted output of multi-dimensional label sets. In this case, the independent classification of these context labels was performed as the baseline approach for our iterative experiments, yielding the worse performance in general (as shown in Table 3.10).

Since we were able to perform more accurate multi-context recognition previously, our research in Chapter 4 was mainly focused on the issue of improving the predictability of the raw user annotations. Essentially, the multiple contextual labels that we used in the experiments of Chapter 3 are in fact the result of a decomposition process (e.g., through entity recognition) from raw user annotations. Consequently, we leveraged the same real-world mobile sensing dataset (i.e., CrowdSignal dataset [Welbourne and Tapia, 2014]) for the experiment in Chapter 4. We proposed a new problem definition for predicting raw user annotations just in time before the experience sampling of user annotation is triggered (i.e., in-situ survey to ask for recent user activities, contexts or even a situation). Consequently, we proposed a novel framework named as CoAct-nnotate, which comprises semi-supervised learning components of multi-view learning (co-training) and active learning. CoAct-nnotate leverages the multi-view learning concept by first building a classifier per sensor. To perform an annotation prediction, the semi-supervised learning module in CoAct-nnotate requires the input of feature bags (for all sensors that are available at the prediction time). Essentially, a feature bag is a derived representation from the concept of multi-instance learning, where a "bag" should contain multiple instances of extracted features. Inherently, this "bag" should be marked with a raw user annotation/label. Moreover, our framework is robust enough to adapt to user feedback and be able to improve the overall predictive model progressively through the application of active learning. Thus, the empirical evaluation of our repeated experiments yielded the improvement of correctness measure (accuracy over the whole duration of mobile sensing in-the-wild) by 35.94%, outperforming the result of those conventional approaches. In particular, the average progression of correctness for annotation prediction of CoAct-nnotate can also be visibly seen to outperform all baselines, as shown in Figure 4.6 of Section 4.4.3.

- ***RQ-3.*** *How to recognise tasks of a mobile user from continuous contextual signals?*

  The effective acquisition of user annotations was addressed for RQ-2. Hence, the system to log pervasive signals from smart devices and record user annotations was then leveraged for data collection of human tasks in daily life. For the third research question (RQ-3), we addressed an issue of situation inference (i.e., task recognition) in Chapter 5. To answer this research question, we proposed a novel contextual modelling approach over CPS signals that can be derived from daily activities in mobile sensing, and align the extracted features for task recognition. Consequently, we presented a new intelligent task recognition problem, which consists of 1) the definition of cyber, physical and social activities of a mobile user, 2) presence-based task boundary construction, and 3) final recognition phase for user tasks. In order to solve this problem, we proposed a process (refer to Section 5.4.2) to define the boundary of user tasks based on in-situ annotations. Moreover, we included the process of CPS-based task modelling and learning on the feature sets that can be derived from both sensing data and user annotations. To validate the proposed framework, we evaluated the intelligent task recognition models on the four-week dataset that we collected on non-professionals and busy-professionals. We conducted a detailed analysis and showed that we could improve the overall performance of task recognition by leveraging all CPS activities of these mobile users. Based on the feature importance analysis, physical and social activities that are predominantly included in top-10 ranked features for intelligent task recognition. Inherently, cyber activities could be used to enhance the task recognition in certain situations (e.g., for working or leisure purposes).

- ***RQ-4.*** *How to evaluate the quality of crowdsourced data from mobile sensing environments?*

  Given that the data could be streamed from many providers through in-the-wild mobile sensing, the fourth research question (RQ-4) is tailored for the scalability issue of mobile sensing experiment itself. Therefore, we addressed this issue in Chapter 6. In this case, we considered a limitation of cloud computing to store raw sensing data. This limitation is a non-trivial issue which would restrict the innovations in ubiquitous computing research. For scalable context recognition and situation inference via mobile sensing, there is a need to build compact machine learning models based on the most informative sensing data. In order to address this issue, we presented a smart data summarisation technique based on the quantification of quality measure (i.e., data utility) on crowdsourced sensor data. The processes of QoD (Quality of Data) estimation (Section 6.4.1) and data summarisation (Section 6.4.3) are essentially integrated into QDaS, a novel domain agnostic framework for effective storage and management of IoT data in the cloud. The data summarisation process can be initiated on the slice of multivariate sensing data, which are categorised as for example “low-quality data”. To determine if a data slice needs to be summarised, it requires a novel coin toss test process which depends on the outcome of MC-CRP (Section 6.4.2.2).

  Not only can the QoD be quantified, but the reputation of data providers belonging to the mobile sensor data can also be computed. This mechanism can be used to encourage the data providers (e.g., mobile users) to improve the quality of their future data collection for mobile sensing experiments. Based on our empirical experiment and evaluation, the amount of stored data could be significantly reduced by density-based data summarisation process. Nevertheless, our smart data summarisation technique was able to achieve a high space-saving ratio while maintaining stable inter-rater agreement between machine learning models, according to the simulation that we performed on activity recognition and air quality estimation datasets. In fact, the density-based data summarisation component in QDaS was also used in CoAct-nnotate for deriving a compact representation of feature bags to improve the multi-view annotation prediction model progressively in a scalable manner.

## 7.2 Limitations and Future Directions of Research

The contributions in this thesis are built upon progressive research efforts on the systematic processes towards situation inference. In particular, the methods, algorithms and frameworks presented in this research can be utilised, adapted and extended to solve various problems within the mobile sensing domain. It should be noted that we envision situation awareness as the norm in future intelligent systems. Therefore, we consider this research as a pioneering stepping-stone with many possible research directions, given the current on-going challenges of mobile sensing in-the-wild.

Approximating the best window size for temporal segmentation across multiple label sets (context or human activity sets) is beneficial to building an intelligent model. However, this approach could be limited with the emergence of new contexts (e.g., new modes of transportation) in the mobile sensing environments. For instance, the ferry[1] is integrated as another mode of transportation in New South Wales, Australia. Uber recently announced[2] their plan to trial the aerial taxi service in Melbourne, Australia. With the everchanging user activities, behaviours and trends in mobile sensing environments, the future research on adaptable window sizes should be considered. In such cases, the window sizes should grow and shrink according to activities, contexts and situations of the mobile users. Since our windowing technique is robust for multi-activity recognition scenarios, the limitation of modelling multiple contexts is thus addressed in Chapter 3. It was indicated that the overall classification performance can be improved by inducing the notion of dependency between context labels (including human activities) across multidimensional context spaces. However, the on-going challenge of mobile sensing in-the-wild is incorporated with the emergence of a new class label in daily life. Consequently, this research direction could be pursued with the utilisation of non-supervised learning techniques.

[1] https://transportnsw.info/tickets-opal/opal/fares-payments/adult-fares
[2] https://www.abc.net.au/news/2019-06-12/uber-elevate-set-to-take-off-in-australia/11199466

Human activities and their contextual information can be inferred from annotations during the mobile sensing in-the-wild. In Chapter 4, the prediction of mobile user annotations can be improved using progressive learning techniques. Since the acquisition of in-situ user annotations can be performed through ESM, the issues on the subjectivity are still prominent during in-the-wild mobile sensing. Such issues require immediate needs for future work to investigate and explore the actual user burden in a typical diary study on mobile sensing. Moreover, the selection of features and learning parameters can also affect the performance of intelligent models. Hence, progressive techniques to adapt the feature selection and its learning parameters can be pursued by considering evolving user contexts in daily life. It should be noted that the user contexts can evolve based on behaviour changes that are the long-term goals of an individual. Given the evolving user activities and contexts, recognising human tasks in daily life (Chapter 5) would be insufficient, to build a more proactive intelligent assistant. Thus, tracking the progression of tasks in a particular situation should be included for future works, in order for an intelligent system to be helpful in supporting humans in daily life. Since annotations that are related to user situations can be acquired through both experience sampling and daily reconstruction methods, another future work should include reasoning and fusing both types of annotations for situation inference. On the other hand, a compact representation of individual mobile sensing data is crucial for building a robust and scalable intelligent system. Hence, its computation, including the required resources, is often designed in a centralised approach. By offloading such computation (also the model building) in a decentralised manner (e.g., federated learning [Smith et al., 2017] and blockchain technology [Dai et al., 2019]), future research should consider extending the notion of data quality and summarisation.

# Appendix A

# Experience Sampling Method (ESM): Survey

Considering the following user task of **Writing WSDM workshop paper**, the questions in ESM-based survey would be presented as follows (based on user contribution on the hourly survey at 01:35 pm):

**Q1.** Have you engaged in one or more activity between [12:00 PM] and [01:00 PM]?

- Yes, I engaged in one activity. » **Proceeds to Q2.**
- Yes, I engaged in more than one activity. » **Proceeds to Q2.**
- No, I did not engage in any activities. » **Finishes the survey.**

**Q2.** Which activity did you spend most of your time on?

- **Writing WSDM workshop paper.**

**Q3.** What category does this activity belong to?

- Work-related tasks
- Personal organization, reflection or care (includes commuting, cleaning and house improvement)
- Caring (household or non-household members)
- Social, exercise & relaxation (entertainment)
- Civil obligations
- Other: ____________________

**Q4.** To which activity does **"Writing WSDM workshop paper"** belong to?

- Communication
- Documentation
- Planning
- Admin and management
- Education
- IT
- Finance
- Physical
- Problem solving
- Low-level
- Project
- Customer care
- Meals and breaks
- Travel
- Other: _______________

**Q5.** When did you start this activity?

- Between [12:00 PM] and [01:00 PM]
- Between [11:00 AM] and [12:00 PM]
- Before [11:00 AM]

**Q6.** Is this a new activity?

- Yes, this is a new activity
- No, I was working on this activity previously (e.g., yesterday)

**Q7.** How much progress did you make towards completing this activity by [01:00 PM]?

- 0% – 19%
- 20% – 39%
- 40% – 59%
- 60% – 79%
- 80% – 99%
- 80% – 99%
- 100% (complete)

**Q8.** Did you use electronic device(s) in order to complete your task/activity? (such as using a computer, smartphone, or tablet)

- Yes
- No

**Q9.** Did you engage in any physical activities in order to complete your task/activity? (such as moving desks, walking to the shops, or cycling home)

- Yes
- No

**Q10.** Did you engage in any social interactions in order to complete your task/activity? (such as meeting someone in person, calling someone)

- Yes
- No

# Appendix B

# Daily Reconstruction Method (DRM): Survey

Considering the following user task of **Writing WSDM workshop paper**, the questions in DRM-based survey would be presented as follows:

**Q1.** What time did you wake up today? (hh:mm)

**Q2.** How many hours did you spend for sleeping (in total)?

- More than 8 hours
- 8 hours
- 7 hours
- 6 hours
- 5 hours
- Less than 5 hours

**Introduction to rest of survey:** Thinking about today, we'd like you to reconstruct what your day was like, as if you were writing in your diary.

Think of your day as a continuous series of scenes or episodes in a film. Each episode is a task that you have performed or in progress towards the completion.

Each task should at least be performed in one-hour duration. In this study, we aim to understand how an intelligent assistant can help in recognizing and managing your daily tasks, to increase the overall productivity and your quality of life. **Next »**

**Q3.** Have you attempted/progressed on any tasks today?

- Yes » **Proceeds to Q4.**
- No » **Finishes the survey.**

**Q4.** Please enter the description of one task you attempted/progressed on today.

**Q5.** To which category does **Writing WSDM workshop paper** belong to?

- Work-related tasks
- Personal organization, reflection or care (includes commuting, cleaning and house improvement)
- Caring (household or non-household members)
- Social, exercise & relaxation (entertainment)
- Civil obligations
- Other: ____________________

**Q6.** To which of the activity/task-type does **Writing WSDM workshop paper** belong to?

- Communication
- Documentation
- Planning
- Admin and management
- Education
- IT
- Finance
- Physical
- Problem solving
- Low-level
- Project
- Customer care
- Meals and breaks
- Travel
- Other: _______________

**Q7.** What kind of trigger did you initiate **Writing WSDM workshop paper**?

- Deadline
- Reminder/alarm (e.g., through digital notification)
- Ad-hoc/spontaneously
- Needs for resources
- Other: ____________________

**Q8.** What is the approximate time when you started **Writing WSDM workshop paper** (hh:mm format)?

**Q9.** Approximate progress when you started **Writing WSDM workshop paper**:

- 0% – 19%
- 20% – 39%
- 40% – 59%
- 60% – 79%
- 80% – 99%
- 80% – 99%
- 100% (complete)

**Q10.** What is the approximate time when you stopped **Writing WSDM workshop paper** (hh:mm format)?

**Q11.** Approximate progress when you stopped **Writing WSDM workshop paper**:

- 0% – 19%
- 20% – 39%
- 40% – 59%
- 60% – 79%
- 80% – 99%
- 80% – 99%
- 100% (complete)

**Q12.** How satisfied are you with the progress of **Writing WSDM workshop paper**?

- Extremely satisfied
- Somewhat satisfied
- Neither satisfied nor dissatisfied
- Somewhat dissatisfied
- Extremely dissatisfied

**Q13.** Thinking about the urgency of this task, what was your perceived priority of when you started **Writing WSDM workshop paper**:

- High
- Medium
- Low

**Q14.** Who were you directly interacting with in the progression of Writing WSDM workshop paper?

- None
- Spouse/significant other
- Household member(s)
- Friend(s)
- Co-worker(s)
- Boss(es)
- Others: _______________

**Q15.** Describe your activities and contexts involved for the progression of **Writing WSDM workshop paper**.

**Q16.** **Writing WSDM workshop paper**? Recalling today's tasks, is there any more task you attempted to progress on?

- Yes » **Loops back to Q4.**
- No » **Finishes the survey.**

# Appendix C

# Approved Ethics Application

College Human Ethics Advisory Network (CHEAN)
College of Science, Engineering and Health

Email: seh-human-ethics@rmit.edu.au
Phone: [61 3] 9925 4620
Building 91, Level 2, City Campus/Building 215, Level 2, Bundoora West Campus

11 May 2018

Dr Flora Salim
School of Science
RMIT University

Dear Dr Salim

**SEHAPP 09-18 The study of task progression for an intelligent assistant**

Thank you for submitting your amended application for review.

I am pleased to inform you that the CHEAN has approved your application for a period of **14 Months** from the date of this letter to **31 July 2019** and your research may now proceed.

The CHEAN would like to remind you that:

All data should be stored on University Network systems. These systems provide high levels of manageable security and data integrity, can provide secure remote access, are backed up on a regular basis and can provide Disaster Recover processes should a large scale incident occur. The use of portable devices such as CDs and memory sticks is valid for archiving; data transport where necessary and for some works in progress. The authoritative copy of all current data should reside on appropriate network systems; and the Principal Investigator is responsible for the retention and storage of the original data pertaining to the project for a minimum period of five years.

**Please Note:** Annual reports are due on the anniversary of the commencement date for all research projects that have been approved by the CHEAN. Ongoing approval is conditional upon the submission of annual reports failure to provide an annual report may result in Ethics approval being withdrawn.

Final reports are due within six months of the project expiring or as soon as possible after your research project has concluded.

The annual/final reports forms can be found at:
www.rmit.edu.au/staff/research/human-research-ethics

Yours faithfully,

[redacted]

**Associate Professor Barbara Polus**
**Chair, Science Engineering & Health**
**College Human Ethics Advisory Network**

| | | |
|---|---|---|
| Cc | Student Investigator/s: | Jonathan Liono, [redacted], School of Science |
| | | Amin Sadri, [redacted], School of Science |
| | | Johanne Trippas, [redacted], School of Science |
| | Other investigator/s: | Dr Yongli Ren, School of Science |
| | | A/Prof Falk Scholer, School of Science |
| | | Prof Mark Sanderson, Enabling Capability Platforms, RMIT University |
| | | Dr Ryen White, Microsoft Research |
| | | Dr Roy Zimmerman, Microsoft Research |
| | | Dr Paul Bennett, Microsoft Research |
| | | Ahmed Hassan Awadallah, Microsoft Research |

# Bibliography

Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy. Adaptive mobile activity recognition system with evolving data streams. *Neurocomputing*, 150:304–317, 2015. Cited on page 40.

Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy. Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)*, 51(4):71, 2018. Cited on pages 84 and 114.

T. Abeel, Y. V. d. Peer, and Y. Saeys. Java-ml: A machine learning library. *Journal of Machine Learning Research*, 10(Apr):931–934, 2009. Cited on page 24.

M. J. Adams, Y. J. Tenney, and R. W. Pew. Situation awareness and the cognitive management of complex systems. In *Situational Awareness*, pages 43–62. Routledge, 2017. Cited on page 4.

J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011. Cited on page 18.

E. Aïmeur, O. Lawani, and K. Dalkir. When changing the look of privacy policies affects user trust: An experimental study. *Computers in Human Behavior*, 58:368–379, 2016. Cited on page 120.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985. Cited on page 126.

C. Anderson, I. Hübener, A.-K. Seipp, S. Ohly, K. David, and V. Pejovic. A survey of attention management systems in ubiquitous computing environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2):1–27, 2018. ISSN 2474-9567. Cited on page 87.

G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 33–40, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273501. Cited on page 75.

D. Ardagna and B. Pernici. Adaptive service composition in flexible processes. *IEEE Transactions on software engineering*, 33(6), 2007. Cited on page 119.

I. B. Arief-Ang, F. D. Salim, and M. Hamilton. Human occupancy recognition with multivariate ambient sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016. Cited on page 57.

I. B. Arief-Ang, F. D. Salim, and M. Hamilton. DA-HOC: Semi-Supervised Domain Adaptation for Room Occupancy Prediction using CO2 Sensor Data. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, BuildSys '17, pages 1:1–1:10, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5544-5. Cited on page 57.

I. B. Arief-Ang, M. Hamilton, and F. D. Salim. A scalable room occupancy prediction with transferable time series decomposition of co2 sensor data. *ACM Transactions on Sensor Networks*, 2018a. Cited on page 57.

I. B. Arief-Ang, M. Hamilton, and F. D. Salim. Rup: Large room utilisation prediction with carbon dioxide sensor. *Pervasive and Mobile Computing*, 46:49–72, 2018b. Cited on page 57.

M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya. Big data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79:3–15, 2015. Cited on page 118.

U. Atz. Evaluating experience sampling of stress in a single-subject research design. *Personal and ubiquitous computing*, 17(4):639–652, 2013. Cited on page 59.

L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer networks*, 54(15): 2787–2805, 2010. Cited on page 116.

L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive computing*, pages 1–17, 2004. Cited on pages 59, 61, and 62.

C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16, 2009. Cited on pages 117 and 119.

G. Bedny and W. Karwowski. *A systemic-structural theory of activity: Applications to human performance and work design*. CRC press, 2006. Cited on page 84.

G. Bedny and D. Meister. *The Russian theory of activity: Current applications to design and learning*. Psychology Press, 2014. Cited on page 84.

R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961. Cited on page 19.

A. Beloglazov and R. Buyya. Energy efficient resource management in virtualized cloud data centers. In *Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing*, pages 826–831. IEEE Computer Society, 2010. Cited on page 119.

L. Benoit, R. Benoit, É. Belin, R. Vadaine, D. Demilly, F. Chapeau-Blondeau, and D. Rousseau. On the value of the kullback–leibler divergence for cost-effective spectral imaging of plants by optimal selection of wavebands. *Machine Vision and Applications*, pages 1–11, 2015. Cited on page 24.

F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, and D. Lottridge. Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2 (3):91:1–91:24, Sept. 2018. ISSN 2474-9567. doi: 10.1145/3264901. URL http://doi.acm.org/10.1145/3264901. Cited on page 83.

N. v. Berkel, D. Ferreira, and V. Kostakos. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, 50(6):93, 2017. Cited on pages 59 and 91.

D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007. Cited on pages 74 and 138.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998. Cited on page 62.

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*, 1(3):38, 2017. Cited on page 60.

Y.-J. Byon and S. Liang. Real-time transportation mode detection using smartphones and artificial neural networks: Performance comparisons between smartphones and conventional global positioning system sensors. *Journal of Intelligent Transportation Systems*, 18(3):264–272, 2014. Cited on page 51.

L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2015. Cited on pages 117 and 119.

R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*, 41(1):23–50, 2011. Cited on page 120.

J. Cancela, M. Pastorino, E. Moreno, and M. A. Waldmeyer. A mobile monitoring tool for the automatic activity recognition and its application for parkinson's disease rehabilitation. In *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*, pages 1457–1460. Springer, 2015. Cited on page 21.

T.-D. Cao, T.-V. Pham, Q.-H. Vu, H.-L. Truong, D.-H. Le, and S. Dustdar. Marsa: a marketplace for realtime human sensing data. *ACM Transactions on Internet Technology (TOIT)*, 16(3):16, 2016. Cited on page 120.

G. Cardone, A. Corradi, L. Foschini, and R. Ianniello. Participact: A large-scale crowdsensing platform. *IEEE Transactions on Emerging Topics in Computing*, 4(1):21–32, 2016. Cited on page 119.

S. Chaisiri, B.-S. Lee, and D. Niyato. Optimization of resource provisioning cost in cloud computing. *IEEE Transactions on Services Computing*, 5(2):164–177, 2012. Cited on page 120.

C. K. Chang, H.-y. Jiang, H. Ming, and K. Oyama. Situ: A situation-theoretic approach to context-aware service evolution. *IEEE Transactions on Services Computing*, 2(3):261–275, 2009. Cited on page 8.

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*, pages 3551–3554. IEEE, 2006. Cited on page 60.

J. Y. Chen and J. E. Thropp. Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):1063–1076, 2007. Cited on page 86.

L. Chen, J. Hoey, C. Nugent, D. Cook, and Z. Yu. Sensor-based activity recognition. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):790–808, Nov 2012a. ISSN 1094-6977. Cited on pages 19 and 40.

L. Chen, C. D. Nugent, and H. Wang. A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):961–974, 2012b. Cited on page 7.

Y. Cheng, M. S. Iqbal, A. Gupta, and A. R. Butt. Cast: Tiering storage for data analytics in the cloud. In *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*, pages 45–56. ACM, 2015. Cited on page 120.

H. Cho, J. An, I. Hong, and Y. Lee. Automatic sensor data stream segmentation for real-time activity prediction in smart spaces. In *Proceedings of the 2015 Workshop on IoT Challenges in Mobile and Industrial Systems*, IoT-Sys '15, pages 13–18, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3502-7. Cited on page 41.

D. Cook, M. Schmitter-Edgecombe, A. Crandall, C. Sanders, and B. Thomas. Collecting and disseminating smart home sensor data in the casas project. In *Proceedings of the CHI Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research*, pages 1–7, 2009. Cited on page 16.

D. J. Cook and S. K. Das. How smart are our environments? an updated look at the state of the art. *Pervasive and mobile computing*, 3(2):53–73, 2007. Cited on page 7.

D. J. Cook, G. M. Youngblood, E. O. Heierman III, K. Gopalratnam, S. Rao, A. Litvin, and F. Khawaja. Mavhome: An agent-based smart home. In *PerCom*, volume 3, pages 521–524, 2003. Cited on page 16.

M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal*, 17(2):386–403, 2017. Cited on page 60.

M. Csikszentmihalyi and R. Larson. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*, pages 35–54. Springer, 2014. Cited on pages 44, 59, 61, and 89.

M. Csikszentmihalyi, R. Larson, and S. Prescott. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6(3):281–294, Sep 1977. ISSN 1573-6601. Cited on page 89.

Y. Dai, D. Xu, S. Maharjan, Z. Chen, Q. He, and Y. Zhang. Blockchain and deep reinforcement learning empowered intelligent 5g beyond. *IEEE Network*, 33(3):10–17, 2019. Cited on page 150.

B. Das, N. C. Krishnan, and D. J. Cook. Racog and wracog: Two probabilistic oversampling techniques. *IEEE transactions on knowledge and data engineering*, 27(1):222–234, 2015. Cited on page 72.

C. De Maio, G. Fenza, V. Loia, and F. Orciuoli. Making sense of cloud-sensor data streams via fuzzy cognitive maps and temporal fuzzy concept analysis. *Neurocomputing*, 256:35–48, 2017. Cited on page 4.

N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006. Cited on page 118.

A. Even and G. Shankaranarayanan. Utility-driven assessment of data quality. *ACM SIGMIS Database*, 38(2):75–93, 2007. Cited on page 119.

W. Fan. Data quality: From theory to practice. *ACM SIGMOD Record*, 44(3):7–18, 2015. Cited on page 120.

T. Feng and H. J. Timmermans. Transportation mode recognition using {GPS} and accelerometer data. *Transportation Research Part C: Emerging Technologies*, 37:118 – 130, 2013. ISSN 0968-090X. Cited on pages 19 and 40.

Y.-H. Feng, T.-H. Teng, and A.-H. Tan. Modelling situation awareness for context-aware decision support. *Expert Systems with Applications*, 36(1):455 – 463, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.09.061. URL http://www.sciencedirect.com/science/article/pii/S0957417407004769. Cited on page 4.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973. Cited on page 126.

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*, pages 91–100. ACM, 2014. Cited on page 62.

G. L. Foresti, M. Farinosi, and M. Vernier. Situational awareness in smart environments: socio-mobile and sensor data fusion for emergency response to disasters. *Journal of Ambient Intelligence and Humanized Computing*, 6(2):239–257, Apr 2015. ISSN 1868-5145. doi: 10.1007/s12652-014-0227-x. URL https://doi.org/10.1007/s12652-014-0227-x. Cited on page 4.

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*, pages 57–70. ACM, 2007. Cited on page 61.

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*, pages 1043–1052. ACM, 2009. Cited on page 61.

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*, 45:11–19, 2015. Cited on page 59.

R. K. Ganti, F. Ye, and H. Lei. Mobile Crowdsensing : Current State and Future Challenges. *IEEE Communications Magazine*, 49(11):32–39, 2011. ISSN 0163-6804. doi: 10.1109/MCOM.2011.6069707. Cited on pages 36 and 119.

Á. García-Recuero, S. Esteves, and L. Veiga. Quality-of-data for consistency levels in geo-replicated cloud data stores. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, volume 1, pages 164–170. IEEE, 2013. Cited on page 119.

C. Geib, V. Agrawal, G. Sukthankar, L. Shastri, and H. Bui. Architectures for activity recognition and context-aware computing. *AI Magazine*, 36(2):3–9, 2015. Cited on pages 7 and 8.

D. Georgakopoulos and P. P. Jayaraman. Internet of things: from internet scale sensing to smart services. *Computing*, 98(10):1041–1058, Oct 2016. ISSN 1436-5057. Cited on page 116.

J. Gershuny. Costs and benefits of time sampling methodologies. *Social Indicators Research*, 67(1-2): 247–252, 2004. Cited on page 60.

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*, pages 713–722. ACM, 2015. Cited on pages 59 and 62.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008. Cited on page 17.

D. Gordon, J. Czerny, and M. Beigl. Activity recognition for creatures of habit. *Personal and ubiquitous computing*, 18(1):205–221, 2014. Cited on pages 18 and 21.

R. Gouveia and E. Karapanos. Footprint tracker: supporting diary studies with lifelogging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2921–2930. ACM, 2013. Cited on page 59.

D. Graus, P. N. Bennett, R. W. White, and E. Horvitz. Analyzing and predicting task reminders. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP'16)*, pages 7–15. ACM, 2016. Cited on page 87.

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*, pages 469–475. IEEE, 2007. Cited on page 63.

B. Guo, R. I. Damper, S. R. Gunn, and J. D. Nelson. A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. *Pattern Recognition*, 41(5):1653–1662, 2008. Cited on pages 19 and 24.

T. Guo, Z. Yan, and K. Aberer. An adaptive approach for online segmentation of multi-dimensional mobile data. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pages 7–14. ACM, 2012. Cited on page 18.

C. Gurrin, A. F. Smeaton, A. R. Doherty, et al. Lifelogging: Personal big data. *Foundations and Trends® in Information Retrieval*, 8(1):1–125, 2014. Cited on page 59.

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*, 33:50–72, 2016. Cited on page 59.

S. Hachem, A. Pathak, and V. Issarny. Service-oriented middleware for large-scale mobile participatory sensing. *Pervasive and Mobile Computing*, 10:66–82, 2014. Cited on page 119.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009. ISSN 1931-0145. Cited on pages 31 and 54.

S. Hemminki, P. Nurmi, and S. Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 13:1–13:14, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2027-6. Cited on pages 19, 40, and 51.

J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmaki, and H. T. Toivonen. Time series segmentation for context recognition in mobile devices. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 203–210. IEEE, 2001. Cited on page 19.

H. S. Hossain, M. A. A. H. Khan, and N. Roy. Active learning enabled activity recognition. *Pervasive and Mobile Computing*, 38:312–330, 2017. Cited on page 63.

T. Hu, M. Xiao, C. Hu, G. Gao, and B. Wang. A qos-sensitive task assignment algorithm for mobile crowdsensing. *Pervasive and Mobile Computing*, 2017. Cited on page 119.

W. Hua, Y. Song, H. Wang, and X. Zhou. Identifying users' topical tasks in web search. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 93–102. ACM, 2013. Cited on page 84.

S. S. Intille, K. Larson, E. M. Tapia, J. S. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. Using a live-in laboratory for ubiquitous computing research. In *International Conference on Pervasive Computing*, pages 349–365. Springer, 2006. Cited on page 16.

J. Iyilade and J. Vassileva. A framework for privacy-aware user data trading. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 310–317. Springer, 2013. Cited on page 120.

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*, volume 898, 2015. Cited on page 63.

P. P. Jayaraman, C. Perera, D. Georgakopoulos, and A. Zaslavsky. Mosden: a scalable mobile collaborative platform for opportunistic sensing applications. *arXiv preprint arXiv:1405.5867*, 2014. Cited on page 119.

R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM, 2008. Cited on page 84.

H. Junker, O. Amft, P. Lukowicz, and G. Tröster. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*, 41(6):2010–2024, 2008. Cited on page 60.

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*, 306(5702):1776–1780, 2004. Cited on page 61.

R. Kawajiri, M. Shimosaka, and H. Kashima. Steered crowdsensing: incentive design towards quality-oriented place-centric crowdsensing. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 691–701. ACM, 2014. Cited on page 119.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural computation*, 13(3):637–649, 2001. Cited on page 54.

M. Khabsa, A. El Kholy, A. H. Awadallah, I. Zitouni, and M. Shokouhi. Identifying task boundaries in digital assistants. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 107–108, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3186952. URL https://doi.org/10.1145/3184558.3186952. Cited on page 86.

S. Khalifa, M. Hassan, and A. Seneviratne. Pervasive self-powered human activity recognition without the accelerometer. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 79–86. IEEE, 2015. Cited on page 84.

C. D. Kidd, R. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. MacIntyre, E. Mynatt, T. E. Starner, and W. Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *International Workshop on Cooperative Buildings*, pages 191–198. Springer, 1999. Cited on page 16.

Y. Kim, F. Pereira, F. Zhao, A. Ghorpade, P. Zegras, and M. Ben-Akiva. Activity recognition for a smartphone based travel survey based on cross-user history data. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 432–437, Aug 2014. Cited on pages 19 and 40.

G. G. Koch. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*, 1982. Cited on page 133.

J. Kohen. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46, 1960. Cited on page 132.

N. C. Krishnan and D. J. Cook. Activity recognition on streaming sensor data. *Pervasive and Mobile Computing*, 10, Part B:138 – 154, 2014. ISSN 1574-1192. Cited on page 19.

B. Kröse, T. Kasteren, C. Gibson, T. Dool, et al. Care: Context awareness in residences for elderly. In *Sixth International Conference of the International Society for Gerontechnology*, 2008. Cited on page 16.

A. B. Krueger and D. A. Schkade. The reliability of subjective well-being measures. *Journal of public economics*, 92(8):1833–1845, 2008. Cited on page 61.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951. Cited on page 24.

G. Laatikainen, O. Mazhelis, and P. Tyrväinen. Role of acquisition intervals in private and public cloud storage costs. *Decision Support Systems*, 57:320–330, 2014. Cited on page 120.

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*, pages 11–16. ACM, 2008. Cited on page 62.

N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, Sept 2010. ISSN 0163-6804. Cited on pages 40 and 119.

M. Last. Multi-target classification: Methodology and practical case studies. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 280–283. Springer, 2016. Cited on page 46.

J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012. Cited on page 118.

Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: a methodology for information quality assessment. *Information & management*, 40(2):133–146, 2002. Cited on pages 117 and 119.

A. N. Leont'ev. *Activity, Consciousness, and Personality*. Prentice-Hall Englewood Cliffs, Nj, 1978. Cited on page 84.

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*, pages 557–566. ACM, 2010. Cited on page 59.

L. Li, H. Deng, A. Dong, Y. Chang, and H. Zha. Identifying and labeling search tasks via query-based hawkes processes. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–740. ACM, 2014. Cited on page 84.

L. Li, H. Deng, Y. He, A. Dong, Y. Chang, and H. Zha. Behavior driven topic transition for search task identification. In *Proceedings of the 25th International Conference on World Wide Web*, pages 555–565. International World Wide Web Conferences Steering Committee, 2016. Cited on page 84.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. Cited on page 24.

W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011. Cited on page 86.

J. Liono, A. K. Qin, and F. D. Salim. Optimal time window for temporal segmentation of sensor streams in multi-activity recognition. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS 2016, pages 10–19, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4750-1. doi: 10.1145/2994374.2994388. URL http://doi.acm.org/10.1145/2994374.2994388. Cited on pages vi, 14, 41, 84, and 114.

J. Liono, Z. S. Abdallah, A. K. Qin, and F. D. Salim. Inferring transportation mode and human activity from mobile sensing in daily life. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous '18, pages 342–351, New York, NY, USA, 2018a. ACM. ISBN 978-1-4503-6093-7. doi: 10.1145/3286978.3287006. URL http://doi.acm.org/10.1145/3286978.3287006. Cited on pages vi, 14, and 115.

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*, 2018b. ISSN 0743-7315. doi: https://doi.org/10.1016/j.jpdc.2018.03.013. URL http://www.sciencedirect.com/science/article/pii/S074373151830220X. Cited on pages vii, 15, 57, 70, 74, and 81.

J. Liono, F. D. Salim, N. van Berkel, V. Kostakos, and A. K. Qin. Improving Experience Sampling with Multi-view User-driven Annotation Prediction. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2019a. Cited on pages vi and 14.

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)*, 2019b. Cited on pages vi, 15, 82, and 96.

A. Liu, J. Ghosh, and C. E. Martin. Generative oversampling for mining imbalanced datasets. In *DMIN*, pages 66–72, 2007. Cited on page 72.

D. Liu, D.-M. Sun, and Z.-D. Qiu. Feature selection for fusion of speaker verification via maximum kullback-leibler distance. In *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, pages 565–568. IEEE, 2010. Cited on page 24.

H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded*

*Networked Sensor Systems*, SenSys '10, pages 71–84, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0344-6. Cited on page 41.

S. Maharjan, Y. Zhang, and S. Gjessing. Optimal incentive design for cloud-enabled multimedia crowdsourcing. *IEEE Transactions on Multimedia*, 18(12):2470–2481, 2016. Cited on page 119.

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. Cited on page 132.

A. Mehrotra, R. Hendley, and M. Musolesi. Notifymehere: Intelligent notification delivery in multi-device environments. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR'19)*, pages 103–111. ACM, 2019. Cited on page 87.

R. B. Miller. The human task as reference for system interface design. In *Proceedings of the ACM/SIGGRAPH Workshop on User-oriented Design of Interactive Graphics Systems*, pages 97–100. ACM, 1976. Cited on page 86.

B. Minor and D. J. Cook. Forecasting occurrences of activities. *Pervasive and mobile computing*, 38: 77–91, 2017. Cited on page 62.

R. Nathuji, A. Kansal, and A. Ghaffarkhah. Q-clouds: managing performance interference effects for qos-aware clouds. In *Proceedings of the 5th European conference on Computer systems*, pages 237–250. ACM, 2010. Cited on page 119.

A. G. Neiat, A. Bouguettaya, T. Sellis, and Z. Ye. Spatio-temporal composition of sensor cloud services. In *Web Services (ICWS), 2014 IEEE International Conference on*, pages 241–248. IEEE, 2014. Cited on page 119.

T. Nguyen, V. Nguyen, F. D. Salim, and D. Phung. Secc: Simultaneous extraction of context and community from pervasive signals. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9, March 2016. doi: 10.1109/PERCOM.2016.7456501. Cited on page 86.

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*, pages 1602–1607. ACM, 2016. Cited on page 62.

I.-S. Oh, J.-S. Lee, and C. Y. Suen. Analysis of class separation and combination of class-dependent features for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1089–1094, 1999. Cited on page 24.

G. Okeyo, L. Chen, H. Wang, and R. Sterritt. Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. *Pervasive and Mobile Computing*, 10, Part B:155 – 172, 2014. ISSN 1574-1192. Cited on page 41.

M. Panteli and D. S. Kirschen. Situation awareness in power systems: Theory, challenges and applications. *Electric Power Systems Research*, 122:140–151, 2015. Cited on page 4.

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*, 12: 2825–2830, 2011. Cited on pages 75, 103, and 135.

V. Pejovic and M. Musolesi. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Comput. Surv.*, 47(3):1–29, 2015. ISSN 0360-0300. Cited on page 87.

H. Peng, Z. Ruan, F. Long, J. H. Simpson, and E. W. Myers. V3d enables real-time 3d visualization and quantitative analysis of large-scale biological image data sets. *Nature biotechnology*, 28(4):348–353, 2010. Cited on page 86.

C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Sensing as a service model for smart cities supported by internet of things. *Transactions on Emerging Telecommunications Technologies*, 25(1): 81–93, 2014. Cited on page 116.

J. D. Pfautz. Depth perception in computer graphics. Technical report, University of Cambridge, Computer Laboratory, 2002. Cited on page 86.

R. W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59 (1-2):55–64, 2003. Cited on page 5.

S. Pirttikangas, K. Fujinami, and T. Nakajima. Feature selection and activity recognition from wearable sensors. In *International Symposium on Ubiquitious Computing Systems*, pages 516–527. Springer, 2006. Cited on pages 18 and 21.

J. S. Preden, K. Tammemäe, A. Jantsch, M. Leier, A. Riid, and E. Calis. The benefits of self-awareness and attention in fog and mist computing. *Computer*, 48(7):37–45, 2015. Cited on page 4.

K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010. Cited on page 17.

M. S. Rahaman, M. Hamilton, and F. D. Salim. Predicting imbalanced taxi and passenger queue contexts in airport. In *Proc. of the Pacific Asia Conf. on Info. Systems (PACIS)*, 2017a. Cited on page 57.

M. S. Rahaman, Y. Mei, M. Hamilton, and F. D. Salim. Capra: A contour-based accessible path routing algorithm. *Information Sciences*, 385-386:157 – 173, 2017b. ISSN 0020-0255. Cited on page 57.

M. S. Rahaman, M. Hamilton, and F. D. Salim. Using big spatial data for planning user mobility. In S. Sakr and A. Zomaya, editors, *Encyclopedia of Big Data Technologies*, pages 1–6. Springer, 2018. ISBN 978-3-319-63962-8. Cited on pages 57 and 85.

J. Read. MEKA: a multi-label extension to WEKA. http://meka.sourceforge.net/, 2012–2015. [Online; accessed 02-December-2015]. Cited on page 54.

J. Read, I. Žliobaitė, and J. Hollmén. Labeling sensing data for mobility modeling. *Information Systems*, 57:207–222, 2016. Cited on page 20.

S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010. Cited on page 51.

Y. Ren, M. Tomko, F. D. Salim, K. Ong, and M. Sanderson. Analyzing web behavior in indoor retail spaces. *Journal of the Association for Information Science and Technology*, 68(1):62–76, 2017. Cited on page 85.

Y. Ren, M. Tomko, F. D. Salim, J. Chan, C. L. A. Clarke, and M. Sanderson. A location-query-browse graph for contextual recommendation. *IEEE TKDE*, 30(2):204–218, Feb 2018a. ISSN 1041-4347. Cited on page 85.

Y. Ren, M. Tomko, F. D. Salim, J. Chan, and M. Sanderson. Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces. *EPJ Data Science*, 7(1): 1, Jan 2018b. ISSN 2193-1127. Cited on pages 85 and 87.

D. Riboni and C. Bettini. Cosar: hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing*, 15(3):271–289, 2011. ISSN 1617-4909. Cited on pages 19 and 40.

A. Sadri, Y. Ren, and F. D. Salim. Information gain-based metric for recognizing transitions in human activities. *Pervasive and Mobile Computing*, 38:92 – 109, 2017. ISSN 1574-1192. Cited on page 57.

H. Sagha, S. T. Digumarti, J. d. R. Millán, R. Chavarriaga, A. Calatroni, D. Roggen, and G. Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 36–40. IEEE, 2011. Cited on page 29.

D. Sánchez, M. Tentori, and J. Favela. Activity recognition for the smart hospital. *IEEE intelligent systems*, 23(2):50–57, 2008. Cited on page 17.

I. H. Sarker and F. D. Salim. Mining user behavioral rules from smartphone data through association analysis. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 450–461. Springer International Publishing, 2018. Cited on page 86.

L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. *Semisupervised learning*, pages 293–308, 2006. Cited on page 134.

I. Schweizer and B. Schmidt. Kraken.me: multi-device user tracking suite. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 853–862. ACM, 2014. Cited on page 44.

R. Sellami, S. Bhiri, and B. Defude. Supporting multi data stores applications in cloud environments. *IEEE Transactions on Services Computing*, 9(1):59–71, 2016. Cited on page 121.

W. Shao, F. D. Salim, A. Song, and A. Bouguettaya. Clustering big spatiotemporal-interval data. *IEEE Transactions on Big Data*, 2(3):190–203, 2016. Cited on page 74.

W. Shao, F. D. Salim, T. Gu, N.-T. Dinh, and J. Chan. Traveling officer problem: Managing car parking violations efficiently using sensor data. *IEEE Internet of Things Journal*, 5(2):802–810, April 2018. ISSN 2327-4662. Cited on page 57.

W. Sherchan, P. P. Jayaraman, S. Krishnaswamy, A. Zaslavsky, S. Loke, and A. Sinha. Using on-the-move mining for mobile crowdsensing. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 115–124, July 2012. Cited on page 119.

S. Sigg, U. Blanke, and G. Troster. The telepathic phone: Frictionless activity recognition from wifi-rssi. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 148–155. IEEE, 2014. Cited on page 84.

V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017. Cited on page 150.

T. Sohn, A. Varshavsky, A. LaMarca, M. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. Griswold, and E. de Lara. Mobility detection using everyday gsm traces. In P. Dourish and A. Friday, editors, *UbiComp 2006: Ubiquitous Computing*, volume 4206 of *Lecture Notes in Computer Science*, pages 212–224. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-39634-5. Cited on page 41.

H. Song, A. K. Qin, and F. D. Salim. Multi-resolution selective ensemble extreme learning machine for electricity consumption prediction. In D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, editors, *Neural Information Processing*, pages 600–609, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70139-4. Cited on page 57.

L. Spirkovska and S. K. Lodha. Context-aware intelligent assistant approach to improving pilot's situational awareness. 2004. Cited on page 4.

N. Stanton, P. Salmon, G. Walker, E. Salas, and P. Hancock. State-of-science: situation awareness in individuals, teams and systems. *Ergonomics*, 60(4):449–466, 2017. Cited on page 4.

A. Stisen, H. Blunck, M. B. Kjærgaard, T. S. Prentow, A. Mathisen, S. Sørensen, and K. Grønbæk. Task phase recognition and task progress estimation for highly mobile workers in large building complexes. *Pervasive and Mobile Computing*, 38:418–429, 2017. Cited on page 86.

S. Stumpf, X. Bao, A. Dragunov, T. G. Dietterich, J. Herlocker, K. Johnsrude, L. Li, and J. Shen. Predicting user tasks: I know what you're doing. In *20th National conference on artificial intelligence (AAAI-05), workshop on human comprehensible machine learning*, 2005. Cited on page 87.

X. Su, H. Tong, and P. Ji. Activity recognition with smartphone sensors. *Tsinghua Science and Technology*, 19(3):235–249, June 2014. Cited on pages 18 and 40.

Y. Sun, N. J. Yuan, Y. Wang, X. Xie, K. McDonald, and R. Zhang. Contextual intent tracking for personal assistants. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pages 273–282. ACM, 2016. Cited on page 87.

M. Susi, V. Renaudin, and G. Lachapelle. Motion mode recognition and step detection algorithms for mobile phone users. *Sensors*, 13(2):1539–1562, 2013. Cited on page 51.

D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and C. Haring. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In *2008 Second International*

*Conference on Pervasive Computing Technologies for Healthcare*, pages 100–102. IEEE, 2008. Cited on pages 7 and 17.

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*, 7(4): 282–293, 2014. Cited on page 63.

R. L. S. Torres, D. C. Ranasinghe, and Q. Shi. Evaluation of wearable sensor tag data segmentation approaches for real time activity classification in elderly. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pages 384–395. Springer, 2013. Cited on pages 7 and 18.

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, and M. Sanderson. Learning About Work Tasks to Inform Intelligent Assistant Design. In *Proceedings of the 2019 Conference on Human Information Interaction & Retrieval*, CHIIR '19, 2019. Cited on pages 82, 85, 89, 93, and 98.

H.-L. Truong and S. Dustdar. On evaluating and publishing data concerns for data as a service. In *Services Computing Conference (APSCC), 2010 IEEE Asia-Pacific*, pages 363–370. IEEE, 2010. Cited on page 119.

H.-L. Truong, S. Dustdar, J. Gotze, T. Fleuren, P. Muller, S.-E. Tbahriti, M. Mrissa, and C. Ghedira. Exchanging data agreements in the daas model. In *Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific*, pages 153–160. IEEE, 2011. Cited on page 121.

G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006. Cited on page 19.

P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. Cited on page 18.

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*, page 554. ACM, 2018a. Cited on page 59.

Y. Vaizman, N. Weibel, and G. Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):168:1–168:22, Jan. 2018b. ISSN 2474-9567. Cited on pages 41 and 59.

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*, CHI '16, pages 4711–4721, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858348. URL http://doi.acm.org/10.1145/2858036.2858348. Cited on page 59.

S. D. Vito, E. Massera, M. Piga, L. Martinotto, and G. D. Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750 – 757, 2008. ISSN 0925-4005. Cited on page 135.

D. T. Wagner, A. Rice, and A. R. Beresford. Device analyzer: Large-scale mobile data collection. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):53–56, 2014. Cited on page 118.

J. Wan, M. J. O'grady, and G. M. O'hare. Dynamic sensor event segmentation for real-time activity recognition in a smart home context. *Personal and Ubiquitous Computing*, 19(2):287–301, 2015. ISSN 1617-4909. Cited on page 41.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020581. URL http://doi.acm.org/10.1145/2020408.2020581. Cited on page 40.

L. Wang, T. Gu, X. Tao, and J. Lu. Sensor-based human activity recognition in a multi-user scenario. In *European Conference on Ambient Intelligence*, pages 78–87. Springer, 2009. Cited on pages 18 and 21.

R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proc. of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014a. Cited on pages 118, 135, and 136.

R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996. Cited on pages 117 and 119.

R. Y. Wang, H. B. Kon, and S. E. Madnick. Data quality requirements analysis and modeling. In *Data Engineering, 1993. Proceedings. Ninth International Conference on*, pages 670–677. IEEE, 1993. Cited on pages 117 and 119.

S. Wang, Z. Liu, Q. Sun, H. Zou, and F. Yang. Towards an accurate evaluation of quality of cloud service in service-oriented cloud computing. *Journal of Intelligent Manufacturing*, 25(2):283–291, 2014b. Cited on page 119.

Z. Wang, M. Jiang, Y. Hu, and H. Li. An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors. *Information Technology in Biomedicine, IEEE Transactions on*, 16(4):691–699, July 2012. ISSN 1089-7771. Cited on page 40.

W. Wayne. Data quality and the bottom line: achieving business success through a commitment to high quality data. *TDWI report*, 2004. Cited on page 117.

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*, UbiComp '14 Adjunct, pages 873–877, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3047-3. doi: 10.1145/2638728.2641309. URL http://doi.acm.org/10.1145/2638728.2641309. Cited on pages 47, 59, 73, and 147.

R. W. White. Opportunities and challenges in search interaction. *Commun. ACM*, 61(12):36–38, Nov. 2018. ISSN 0001-0782. doi: 10.1145/3195180. URL http://doi.acm.org/10.1145/3195180. Cited on page 84.

R. W. White and A. Hassan Awadallah. Task duration estimation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM'19)*, pages 636–644. ACM, 2019. Cited on page 87.

G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996. ISSN 0885-6125. Cited on page 40.

Wikibon. A comprehensive list of big data statistics, 2012. URL http://wikibon.org/blog/big-data-statistics/. Cited on page 117.

D. Wu, M. J. Greer, D. W. Rosen, and D. Schaefer. Cloud manufacturing: Strategic vision and state-of-the-art. *Journal of Manufacturing Systems*, 32(4):564–579, 2013. Cited on pages 119 and 121.

H. Ye, T. Gu, X. Tao, and J. Lu. Crowdsourced smartphone sensing for localization in metro trains. In *A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on*, pages 1–9. IEEE, 2014. Cited on page 118.

J. Ye, S. Dobson, and S. McKeever. Situation identification techniques in pervasive computing: A review. *Pervasive and Mobile Computing*, 8(1):36 – 66, 2012. ISSN 1574-1192. doi: https://doi.org/10.1016/j.pmcj.2011.01.004. URL http://www.sciencedirect.com/science/article/pii/S1574119211000253. Cited on pages 5 and 7.

J. Yin, S. Karimi, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. Cited on page 4.

S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao. Bayesian co-training. *Journal of Machine Learning Research*, 12(Sep):2649–2680, 2011. Cited on page 69.

A. Yurtman and B. Barshan. Activity recognition invariant to sensor orientation with wearable motion sensors. *Sensors*, 17(8):1838, 2017. Cited on page 51.

Ö. Yürür, C. H. Liu, Z. Sheng, V. C. Leung, W. Moreno, and K. K. Leung. Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys & Tutorials*, 18(1):68–93, 2016. Cited on page 7.

A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016. Cited on pages 117 and 119.

Y. Zhang, S. Li, T. Wang, and Z. Zhang. Divergence-based feature selection for separate classes. *Neurocomputing*, 101:32–42, 2013. Cited on page 24.

Y. Zhou, U. Porwal, C. Zhang, H. Q. Ngo, L. Nguyen, C. Ré, and V. Govindaraju. Parallel feature selection inspired by group testing. In *Advances in Neural Information Processing Systems*, pages 3554–3562, 2014. Cited on page 24.

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*, pages 1167–1174. ACM, 2007. Cited on page 62.

C. Zhu and W. Sheng. Human daily activity recognition in robot-assisted living using multi-sensor fusion. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 2154–2159. IEEE, 2009. Cited on pages 18 and 21.

PERITA MANUS MENS EXCULTA