

ABSTRACT

BALRAJ, BHAVANA. Multilabel Active Learning for User Context Recognition In-the-Wild. (Under the direction of Dr. Munindar P. Singh).

User context recognition refers to the automatic inference of user context (comprising a user's activities, social circle, and location) from sensory data. Context recognition poses several challenges due to high variability in behavioral patterns across users and contexts.

Context recognition can be viewed as a multilabel classification problem in which each data instance is simultaneously associated with multiple contexts. Multilabel learning requires enough labelled data to train high quality classification models. Whereas collection of data in uncontrolled environments has been made easier with the advancement in the sensing capabilities of smartphones, the annotation of large amounts of multilabel data for training supervised models is a tedious process. Active learning, a technique that reduces annotation cost by selecting most useful training instances for label acquisition, is highly suitable for such problems. Research in multilabel active learning is at a nascent stage. This thesis investigates how multilabel active learning applies to user context recognition. Its specific goal is to construct an active learning pipeline that will identify informative queries for annotation.

This thesis compares the performance of two multilabel classification models combined with active learning – problem transformation, in which a multilabel problem is decomposed into several single label problems, with algorithm adaptation, in which single label algorithms are extended to handle multilabel problems. The classifiers are evaluated with balanced accuracy, a metric that accounts for both positive and negative outcomes equally. Results are demonstrated on the publicly available ExtraSensory dataset.

This thesis shows that the problem transformation model is better than an algorithm adapted model in terms of overall balanced accuracy. These results also indicate that active learning brings about a marginal reduction in the total number of queries using algorithm adaptation in comparison with problem transformation. We investigate the potential causes for this behavior and explore methods to mitigate their effects.

© Copyright 2021 by Bhavana Balraj

All Rights Reserved

Multilabel Active Learning for User Context Recognition In-the-Wild

by
Bhavana Balraj

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Computer Science

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Xipeng Shen

Dr. Ruozhou Yu

Dr. Pradeep K. Murukannaiah
External Member

Dr. Munindar P. Singh
Chair of Advisory Committee

DEDICATION

To my parents, grandparents, brother, and sister-in-law.

BIOGRAPHY

Bhavana Balraj comes from Chennai, India. She received her Bachelor of Engineering degree from College of Engineering, Guindy - Anna University, India in 2017. She worked as a Software Developer at Juniper Networks, Bengaluru India for two years following which she started pursuing her master's degree in Computer Science at North Carolina State University. She hopes to continue contributing to research in the field of machine learning and data science.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor Professor Dr. Munindar P. Singh for his unwavering guidance and support throughout my master's journey. I would like to thank him for his involvement in this project and for inspiring me with his dedication and passion for research. I would also like to thank my co-advisor Dr. Pradeep K. Murukannaiah for his valuable insights and directions for the completion of this project.

I sincerely thank my advisory committee, Dr. Xipeng Shen and Dr. Ruozhou Yu for their invaluable feedback. I would also like to thank the Centre for Hybrid Multicore Productivity Research (CHMPR) and National Science Foundation (NSF under grant IIS-1908374) for the support extended.

Special thanks to my friends Arvind Sastha Kumar, Rajshree Jain, Prithvi Patel, Karthik Sridhar and Sekkappan Chockalingam for their constant encouragement. Finally, I would like to thank my family for their emotional support.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Our Solution	3
1.4 Key Challenges	4
1.5 Thesis Organization	4
Chapter 2 Related Work	5
2.1 Context Recognition	5
2.1.1 Context Recognition In-the-Wild	6
2.1.2 Active Learning for Context Recognition	7
2.2 Multilabel Active Learning	7
Chapter 3 Dataset	9
3.1 ExtraSensory Dataset	9
Chapter 4 Active Learning Overview	12
4.1 Framework	13
4.2 Algorithm	13
4.3 Query Mechanism	14
4.4 Baseline	14
Chapter 5 Multilabel Active Learning	16
5.1 Multilabel Classification	16
5.1.1 Problem Transformation	16
5.1.2 Algorithm Adaptation	17
5.2 Sampling Strategies	19
5.2.1 Issues	20
5.2.2 Uncertainty Sampling	21
5.3 Balanced Sampling	22
5.3.1 Over Sampling	22
5.3.2 Hyperparameter for balancing	23
5.3.3 Informativeness	23
Chapter 6 Evaluation Metrics	25
6.1 Balanced Accuracy	25
6.2 Macro Balanced Accuracy	26

Chapter 7 Experiments and Results	27
7.1 Active Learning Evaluation	27
7.1.1 Fully Supervised Models	27
7.1.2 Active Learning	28
7.2 Sampling strategies	31
7.2.1 Balanced Sampling	31
7.2.2 MLSMOTE	33
7.2.3 Informativeness	33
7.2.4 Warm Seed	36
Chapter 8 Conclusion	38
8.1 Summary	38
8.2 Future Work	38
References	40
APPENDIX	48
Appendix A Acronyms	49

LIST OF TABLES

Table 3.1	Dataset description	11
Table 7.1	Metric comparison for fully supervised method	28
Table 7.2	Metric comparison for active learning	29
Table 7.3	Metric comparison with balanced sampling	34
Table 7.4	Metric comparison for informativeness variants	35
Table 7.5	Metric comparison for varying percentages of warm seed	36
Table A.1	Acronyms	49

LIST OF FIGURES

Figure 3.1	Class distribution - ExtraSensory	10
Figure 3.2	Class distribution - sampled ExtraSensory	11
Figure 4.1	Active learning framework	13
Figure 7.1	Plots of balanced accuracy for all strategies	30
Figure 7.2	Plots of balanced accuracy for algorithm adaptation	32
Figure 7.3	Class distribution with balanced sampling	32
Figure 7.4	Class distribution with MLSMOTE	33
Figure 7.5	Plot of balanced accuracy with MLSMOTE	34
Figure 7.6	Plots of balanced accuracy for informativeness variants	35
Figure 7.7	Plots of balanced accuracy for different warm seeds	37
Figure 7.8	Plots of macro balanced accuracy for different warm seeds	37

CHAPTER

1

INTRODUCTION

1.1 Background

User context is a broad term defined by several aspects of daily life surrounding humans. It is highly subjective and typically comprises of user activities, location, social circle, time, emotions, and so on [Shen et al., 2020]. Recognition of behavioral context enables the development of intelligent context aware applications ranging from monitoring and providing health care assistance to elderly patients [Byrne et al., 2018], [Lee and Dey, 2015], quantifying behavioral changes during COVID-19 [Beukenhorst et al., 2021] to context aware privacy management for mobile devices [Chitkara et al., 2017]. These adaptable applications act as agents that analyze patterns in user behavior as observed by sensors and try to alter their services based on their knowledge about user's context. Context recognition is at the core of enhancing user experience of context aware applications.

In the real world, contexts are correlated. Multiple unrelated contexts can also occur together. Context recognition agents need to holistically deal with multiple contexts to provide relevant services. For example, one could either listen to music or watch TV while running on a treadmill. The agent needs to infer the current context with the help of its inputs and understand that

playing music while the TV is already switched on may not be desirable. We can train supervised machine learning models to detect such contexts of the user. Context recognition can be viewed as a multilabel classification problem where every data instance is associated with multiple context labels simultaneously.

Context recognition models that perform well in the lab may not be robust enough for real world settings [Natarajan et al., 2016]. Traditional models rely on data collected by sensors (GPS, accelerometer, camera, audio) mounted uncomfortably on users when they perform a predefined sequence of actions. The simulated environment and unnatural sensor device placement fail to reflect randomness inherent to human behavior. Data collected in the wild (a) captures the variations in behavioral patterns across users in a natural environment, (b) is obtained from naturally used devices without any constraints on device placement. Also, contexts are complex as different people interact with their context in different ways. They have different patterns of walking, eating, sleeping, and working. There is no one size fit all model that can learn these differences in contexts across users. To better respond to user behavior, solutions to context aware applications need to generalize and handle in-the-wild data reliably well.

Smartphones are ubiquitous and have become the natural choice for in-the-wild data collection. They are cheap, enable autonomous, less intrusive sensing, and collection of in-the-wild user data. Smartphones have higher processing capabilities and are equipped with multimodal sensors such as GPS, accelerometer, barometer, temperature and light sensors, microphones, and proximity detectors. They are easier to carry compared to their wearable counterparts, can collect, and store large amounts of unlabeled sensor data in the background while the user continues to execute day to day activities. We can extract implicit user context information from the environment and usage data of smartphones [Otebolaku and Andrade, 2016].

Context aware services should not be tightly coupled with the availability of all sensors. In-the-wild situations can have missing and irregular sensor data when the sensors are offline (privacy concerns or during phone calls). Combining the right sensor information can be sufficient for identifying certain contexts. For example, watch accelerometer and audio phones data together can help infer the context of washing dishes even when the GPS data is unavailable. The ability to automatically learn from the right sensor features is highly desirable. Also, every participant contributes different proportions of data to the set of available contexts. The label “At school” may not be applicable to an adult. This introduces incomplete labelling and imbalance in the data. Class imbalance can occur (a) between labels (sitting is more common than elevator) (b) within labels (more positive examples for sitting than negative examples). Applications relying on smartphone sensor data need to deal with missing sensor information and class imbalance issues. Context Recognition in-the-wild presents the challenge of learning multilabel,

multimodal classification from imbalanced data.

1.2 Motivation

Developing context aware applications require the user to annotate the collected sensor readings with corresponding context labels. The process of labelling inputs for generating training sets already constitutes a major issue in conventional single label classification. Recording all relevant contexts as ground truth labels becomes a substantial matter of relevance in the more complex multilabel setting. Training fully supervised multilabel classification models require high quality and consistent annotation of large amounts of sensor data. Major issues with labelling sensor data are:

1. Some users may find it hard to accurately recollect the context when annotation is done much later than data collection. Annotation can be biased and subject to memory recollection errors.
2. Contexts may vary across users. When the context label space is large, the human annotator must scrutinize the relevance of every possible context label to the data instance which leads to higher labelling cost compared to single label learning [Cook et al., 2013].

Due to these reasons, mobile centric automatic context detection and inference from smartphone sensor data is on the rise. A specific goal is to minimize both the labelling and computational cost involved in the process.

1.3 Our Solution

Active learning [Settles, 2012] is a predominantly used machine learning paradigm in literature to reduce labelling effort. An active learning algorithm constructs an effective training set for annotation by the user from a large pool of unlabeled data. It minimizes the annotation cost of unlabeled smartphone sensor data by selecting the most informative sensor readings for labelling using sampling strategies. An instance is considered informative if the classifier finds it hard to determine the context categories. These instances once labelled give greater improvement to the performance metric at the same cost or the same improvement at a lower labelling cost.

1.4 Key Challenges

To summarize, we are given a dataset X with N examples where each example x_i is a d -dimensional feature vector representing the multimodal sensor readings. The target labels y_i for every x_i are a subset of all the context labels Y . Fully supervised classifiers can be trained to learn the complex mapping from a sensor feature space X to a context label space Y . Using this setting, we address the following research questions:

RQ1: How can we reduce the labelling effort for multilabel context recognition in-the-wild?

RQ2: What are the methods to improve the performance of active learning for a unified model designed to handle multilabel context data in-the-wild (with missing values and multimodal sensor features)?

In this work, we compare active learning for two traditional multilabel classification solutions - problem transformation and algorithm adaptation, to quantify the improvement brought about by active learning and discuss trade offs between the two solutions. Further, we seek to improve the performance of algorithm adapted model by exploring various active learning strategies. To our knowledge, this is the first evaluation of effectiveness of active learning for multilabel context recognition in the wild. We believe that the preliminary results using ExtraSensory dataset and identification of open questions will help further research in this domain.

1.5 Thesis Organization

Chapter 2 describes existing work for context recognition in-the-wild and multilabel active learning. Chapter 3 summarizes the multilabel context recognition in-the-wild dataset. Chapter 4 presents an overview of active learning for multilabel data. In Chapter 5, we explain in detail two multilabel classification solutions, sampling strategies used, and investigation of other areas to improve the performance of one of our multilabel approaches. Chapter 6 describes the metrics used to evaluate active learning using our models. Chapter 7 describes experiments designed to answer research questions and explains results. Finally, Chapter 8 discusses the limitations and open questions based on our work.

CHAPTER

2

RELATED WORK

In this section, we identify existing approaches for multilabel context recognition and concurrent multilabel active learning strategies in other domains.

2.1 Context Recognition

Radu et al. [2018] focus on maximizing intermodality for wearable sensors using fully connected DNNs and CNNs for activity and context recognition. Multimodal neural networks are shown to be superior in comparison to hand crafted feature extraction, classifier construction, and task specific modelling pipelines. Rault et al. [2017] perform a comparative study about existing works for energy efficient (power on time reduction, communication reduction) health related human context recognition. A complete analysis on the impact of orientation dependent and independent features on HAR with smartphone sensors is presented in [Sousa et al., 2017]. Cao et al. [2018] design a two-level hierarchical group-based scheme that utilizes context awareness in HAR on smartphones. The inter group classification assigns the input feature to an activity group through clustering. The inner group recognition predicts the concrete activity within that group. Hamad et al. [2020] propose a joint diverse temporal learning framework

using multimodal sensors, LSTM, and 1D CNNs to address the issue of class imbalance for less represented activities in HAR. Hossen et al. [2019] utilize smartphone integrated sensors to recognize context of drivers to facilitate outdoor parking. Acharjee et al. [2017] segregate the task of human context recognition into five levels based on the combinations of users, locality, and activities. They describe computational methods using HMMs and KL Divergence to recognize human contexts for single user, single location, and multiple activity scenario. Training classifiers such as KNN, Naive Bayes, and decision trees by combining data from both wrist worn and smartphone sensors outperforms wrist position alone [Shoaib et al., 2016]. Han et al. [2012] propose a position free context recognition system that infers the human activities regardless of the position of the smartphone. The system conserves power by optimally choosing either GMMs for accelerometer data classification or HMMs for audio classification. Pei et al. [2013] simplify human behavior modelling using Location-Motion-Context model that infers context from location information and motion states using Bayesian reasoning. Kurz and Ferscha [2010] propose opportunistic context recognition middleware system where there is no predefined sensor infrastructure, and the system makes use of the available sensors during runtime.

2.1.1 Context Recognition In-the-Wild

End to end trained DNNs can recognize diverse set of contexts in real world setting by jointly learning representations from multi modal sensors [Saeed et al., 2019]. In another work, Saeed et al. [2018] they empirically demonstrate adversarial autoencoder for handling missing sensory features in the dataset. Li et al. [2019] exploit the correlations among labels to mitigate the problem of incomplete and missing labels. A label similarity regularization on the embedding vectors for labels is obtained such that labels that frequently co-occur together will be closer in the embedded label space. Ehatisham-ul Haq et al. [2020] demonstrate a two-level scheme with multiple classifiers to recognize primitive contexts – lying, sitting, standing, and walking. The first level identifies physical activity using smartphone accelerometer data. The second level combines the activity label with 60 other raw sensor data to recognize context. Using an MLP for handcrafted sensor features and a CNN BiLSTM architecture component to exploit temporal correlations in raw input stream, Ge and Agu [2020] propose context recognition under label uncertainty for ExtraSensory dataset. Multi Layer Perceptron combined with nontraditional instance weighting can generalize well to the variations of uncontrolled in the wild user behavioral patterns of ExtraSensory dataset [Vaizman et al., 2018]. The hidden layers of MLP introduce nonlinearity and dimensionality reduction to handle the incomplete, unbalanced, and

missing sensor data. In our work, we aim to improve the performance of this MLP architecture with Active Learning.

2.1.2 Active Learning for Context Recognition

Cruciani et al. [2018] combine a heuristic function with an online training architecture to enable automatic annotation of HAR data. They investigate the performance of supervised classification approaches for noisy data in free living using smartphone. Cruciani et al. [2019] also provide description of a dataset for free living conditions. A balanced batch learning algorithm for CNN is described as an alternative to SMOTE for handling the class skew in such datasets. Pool based (uncertainty sampling) and Stream based (logistic margin sampling) AL approaches for multiclass classification on ExtraSensory Dataset outperform supervised methods by using 10% of training data. Conditional mutual information of the unlabeled data pool is used as a stopping criterion for active learning [Adaimi and Thomaz, 2019]. Oversampling the selected informative instances in every active learning iteration with Border Limited Link SMOTE technique can tackle imbalance and skew of in-the-wild datasets with the same labelling cost [Nguyen et al., 2018]. A dynamic active learning method based on novelty check and affinity propagation incorporating three dimensions of informativeness measure – uncertainty, diversity, and representativeness can help discover unseen new patterns and activities which appear in real world scenarios [Bi et al., 2021]. Utilizing contextual information during query selection combined with entropy and mutual information of unlabeled data pool can also significantly reduce the annotation cost for event recognition [Hasan et al., 2018]. Bettini et al. [2020] propose that knowledge-based reasoning of context data triggers a smaller number of queries for active learning compared to using context data as additional features for learning. Murukannaiah and Singh [2015] demonstrate the effectiveness of using active and semi supervised approaches with uncertainty sampling and logistic regression to learn models of places specific to a user. Fujinami et al. [2019] propose a framework employing active learning to customize context recognition models on mobile devices by training a general classifier with additional yet minimal user specific data. Bagaveyev and Cook [2014] design and implement several AL strategies for HAR.

2.2 Multilabel Active Learning

For a multilabel problem, it is difficult to determine the classifier’s confidence by comparing the response distribution. Wu et al. [2020] extensively summarize and categorize the various multi

label active learning algorithms for image classification based on the sampling and annotation strategies. Daniels and Metaxas [2017] account for cost sensitivity and introduce new objective function based on Hellinger distance to determine oblique splits to handle imbalance in multi label data. Sozykin et al. [2018] implement a 3D CNN based deep HAR system for multilabel class imbalanced action recognition from videos. Brinker [2006] suggest binary version space minimization that queries the closest instance to the boundary for one of the classifiers used in binary relevance. In Brinker [2003], they explain incorporating diversity for active learning using SVM. Messaoud et al. [2019] demonstrate the effectiveness of MLALAL using the disagreement between the major logistic regression learner and auxiliary SVM learner as a query criterion on mobile user app review classification. Reyes et al. [2018] present a new uncertainty measure combining rank aggregation (Condorcet's Borda count) and inconsistency in label space. Huang et al. [2014] connect margin-based sampling to min max formulation of active learning to measure informativeness and representativeness. Wang and Hua [2011] introduce semi supervised and multilabel active learning models for five sample selection strategies. Vasisht et al. [2014] propose a Bayesian based alternate AL scheme to deal with positive label scarcity. Li et al. [2018] define a new instance selection criterion by combining the Label Cardinality Inconsistency (LCI) with query diversity. Cost effective MAL proposed in [Yu et al., 2020] works by selecting the most informative example/label pairs by leveraging uncertainty, label correlation, and label sparsity. Additionally, greedily querying only probable positive sub examples saves query cost. Esuli and Sebastiani [2009] discuss several variants of active learning strategies for multilabel text classification.

CHAPTER

3

DATASET

3.1 ExtraSensory Dataset

We use the ExtraSensory dataset (<http://extrasensory.ucsd.edu/>) to analyze how active learning can help reduce the labelling effort for context recognition in-the-wild. It is a publicly available dataset. The multimodal sensor data from smartphones was collected using the ExtraSensory mobile application (running in the background). The dataset comprises of measurements from five high frequency smartphone sensors: (a) Accelerometer, (b) Gyroscope, (c) Location, (d) Phone State, and (e) Audio along with accelerometer readings from a smart watch. The measurements were sampled for a window of 20 secs every minute totaling around 300k examples. The dataset also captures the missing sensors. Approximately only half the examples have all the six sensors.

Participants were given a huge vocabulary of contexts to choose from. 60 users recorded their contexts while they were engaged in their authentic behavior without conforming to any specific order of activities. The label cardinality is three for this dataset. Every example was annotated with a combination of at least three relevant labels describing detailed aspects of their context – what the participants were doing, the position of their phone, who they were

with, and so on. Users were allowed to label a main activity (set of seven mutually exclusive activities such as sitting, lying down, walking, running) and a secondary activity (with family, indoors, sleeping, eating). The labels provided by the participants were cleaned and processed for missing label information. This resulted in a total of 51 context labels. Participants recorded labels only for relevant contexts resulting in imbalanced data with missing labels. Unlabeled entries were treated as missing or negative appropriately using common sense rules. This step makes labelling reliable for minority labels.

Figure 3.1 describes the data distribution for ExtraSensory dataset. For our experiments, we sample ten labels out of the 51 labels. Labels with different skews are chosen to better represent our imbalanced multilabel data (Table 3.1). These labels are not mutually exclusive and include both main and secondary activities.

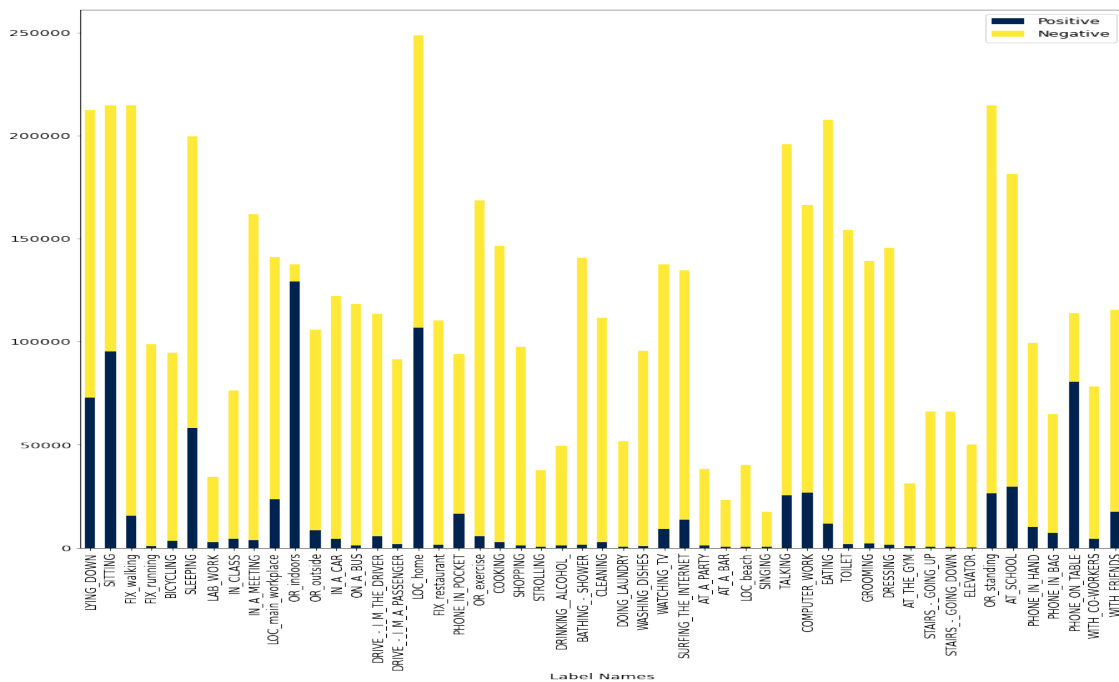


Figure 3.1: Illustration of class distribution for ExtraSensory dataset without missing labels.

Table 3.1: Dataset description.

Labels	Positive Examples
Or_indoors	184692
Loc_home	152892
Sitting	136356
Fix_walking	22136
Watching_tv	13311
Or_outside	12114
Bicycling	5020
Doing_laundry	556
At_a_bar	551
Elevator	200

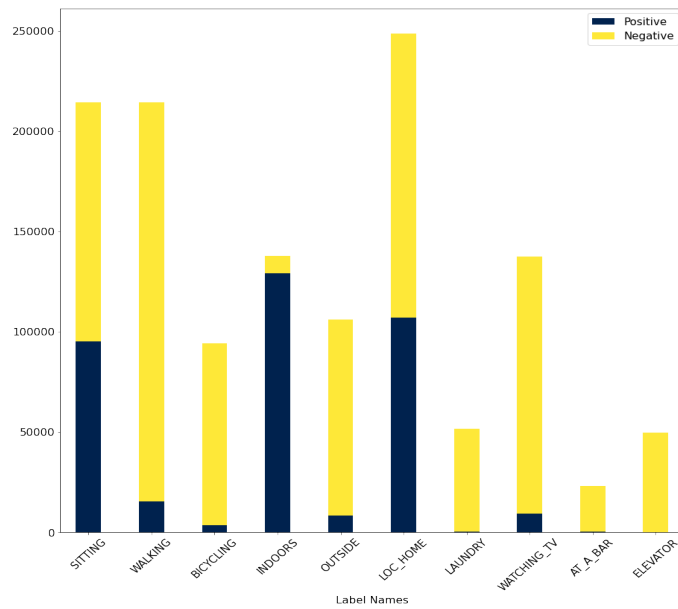


Figure 3.2: Illustration of class distribution for sampled ten label ExtraSensory dataset without missing labels.

CHAPTER

4

ACTIVE LEARNING OVERVIEW

User context recognition is a multilabel and not a multiclass problem. In a **binary classification** problem, the classifier tries to assign either true or false value to every data instance.

Definition 4.1.1 Multiclass classification (MCC) allows data instances to be associated with only one of the several target labels. The target classes are mutually exclusive. The classifier tries to learn a probability distribution over all context labels and outputs the label with highest probability. A softmax activation is used to obtain a distribution such that the sum of probabilities for all the labels is one.

Definition 4.1.2 Multilabel classification (MLC) generalizes the multiclass problem by allowing each instance to be associated with multiple target labels simultaneously. Although there are multiple labels, they do not contradict each other. A sigmoid activation is used to get a binary output value for each label. Here, we want the classifier to learn a probabilistic distribution over the power set of labels where each element of the power set represents a unique combination of context labels. The sum of the probabilities for all the labels need not be one as there can be more than one correct answer.

4.1 Framework

Label acquisition for single label classification is much easier as we need the ground truth for only one category to train our models. For multiclass and multilabel problems, the target space is exponentially large, and the annotator must consider the relevance of an instance to multiple categories even if the positive labels are sparse. Also, repeatedly querying for same/similar data instances may increase the chances of incomplete and incorrect labelling of context categories. Active learning is highly suitable for such problems as it can minimize the queries by eliminating redundancy in annotation. Figure 4.1 shows the general active learning framework.

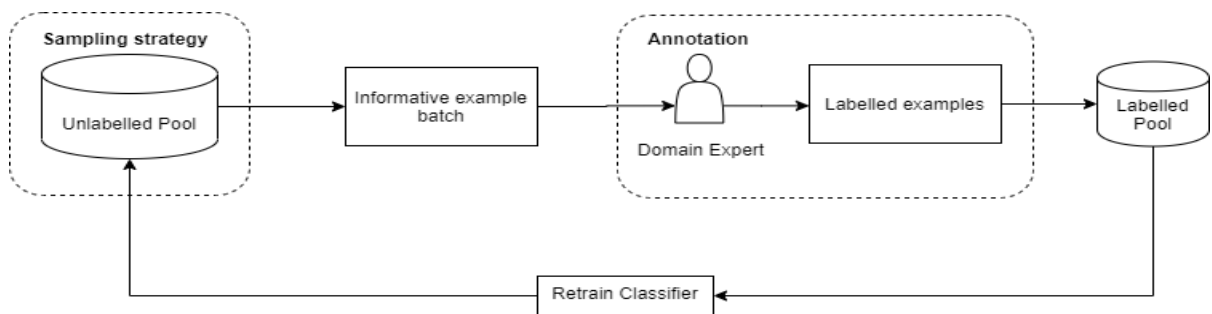


Figure 4.1: Active learning framework.

4.2 Algorithm

Algorithm 1 summarizes the steps in pool based active learning for multilabel problems. There is an unlabeled sensor data pool U and a small set of labelled sensor data L to begin with. An initial multilabel context recognition model θ is trained using labelled data L . In every AL iteration, the algorithm selects a batch of N instances for annotation by the user using an informativeness measure. Firstly, the model θ is used to predict the context labels for all the instances in U . The classifier can predict more than one context label to be true simultaneously. The model predicts the context labels and provides the prediction probabilities for all the labels. The probability for any label can be considered as the measure of how confident the classifier is about its prediction for that label. These probabilities are used to compute the informativeness of the sample. The unlabeled instances in U are now sorted high to low based on their informativeness measure. Top N informative samples (set S) as indicated by their confidence estimates are selected. These samples are expected to bring about the most change to the classifier. We use greedy batch

model sampling where the annotation is done in batches. The context labels for S are provided by a human annotator, the labelled instances are removed from unlabeled pool U , and added to labelled pool L . The model is retrained with augmented L and the process is repeated until the stopping criterion such as number of iterations or performance of the classifier is met.

4.3 Query Mechanism

Active learning algorithms fall under three major categories based on the querying technique (a) membership query synthesis, (b) pool based, (c) stream-based. In membership query synthesis, the active learner synthesizes a query based on the distribution of the input space. The drawback of this technique is that the data point is not sampled from the real world leading which may lead to the inability of the annotator to label that instance. Pool based and stream-based techniques use an informativeness measure to select instances for labelling. Pool based method greedily selects a pool of data by evaluating and ranking all the instances in the unlabeled pool. In stream-based methods, every unlabeled instance is presented sequentially to the classifier. It examines and immediately decides whether to issue a query for that instance before the next instance is presented. Pool based technique is more effective compared to stream-based methods and has attracted a lot of research interest. We will focus on pool-based strategies.

4.4 Baseline

Random sampling strategy in which the instances are selected at random for labelling is chosen as baseline. It gives the average number of labelled examples required to achieve balanced accuracy of fully supervised models. Random strategy simulates random annotation by humans. The active learning sampling strategies are expected to use fewer queries compared to the random strategy.

Algorithm 1: Pool Based Active Learning Algorithm

Input: U : Unlabeled data pool;

L : Labelled training set;

θ : Initial model trained on L ;

N : Batch size;

Output: θ_f : Final trained model;

while *stopping criteria not met* **do**

for every x_i **in** U **do**

pred_probs_i = probability predictions for context labels of X_i ;

Info_i = Compute informativeness using pred_probs_i ;

end

$S \leftarrow$ Sort U based on informativeness measure and select top N examples;

 Acquire labels for instances in S ;

$L \leftarrow L \cup S$;

$U \leftarrow U \setminus S$;

$\theta_t \leftarrow$ Retrain the model with L

end

CHAPTER

5

MULTILABEL ACTIVE LEARNING

5.1 Multilabel Classification

In every active learning iteration, we need to train a base multilabel classifier. We present a short overview of two multilabel classification solutions. We implement, test, and compare active learning for (a) an ensemble of ten binary classifiers, (b) a single multilabel classifier with ten outputs for the multilabel context recognition problem (Source code available at: <https://github.com/bhavanabalraj/Active-Learning-for-Context-Recognition-in-the-wild>). We also discuss the implications of choosing one classifier over the other for active learning.

5.1.1 Problem Transformation

Problem Transformation method breaks a K label classification problem into K single label classification problems. Simplest problem transformation technique is Binary Relevance (BR) where K single label classifiers are trained for each of the output labels. The final multilabel output is obtained by combining the outputs of the K classifiers. Classifier chains overcome this issue by training a chain of Binary Relevance classifiers. The first classifier is trained on the input data. The following classifiers in the chain are trained on the input augmented with the

output from the previous classifiers. The learning from the previous classifiers is propagated through the chain. A multilabel problem can also be converted into a multiclass problem by considering each unique combination of the output labels as a target class. With K labels, there will be 2^K possible combinations of the output labels. In this label power set method, the classifier directly learns the label dependencies and outputs a label combination. Any single label learning algorithm can be used to generate classifiers for problem transformation method.

Binary Relevance

BR methods are conceptually simple and can easily be adapted to handle missing labels [Zhang et al., 2018]. We first design and evaluate a computational model on a label-by-label basis to study the effectiveness of binary relevance for active learning. The binary classifier built for a specific label is expected to achieve a high classification accuracy for that label. Each classifier predicts whether the label is relevant to the instance. These binary predictions are then directly combined to obtain the multilabel output of an example. The Multi Layer Perceptron described in Section 5.1.2 is modified to output a single binary value.

For active learning, every binary classifier computes the informativeness of unlabeled instances based on their prediction probabilities. There are K such classifiers and every classifier will rank the unlabeled pool according to the informativeness measure assigned by it. Combining the informativeness for all the labels to get a single unified value may be detrimental to overall balanced accuracy. The instance useful for one label may not bring about a greater change for another label. We therefore allow each classifier to pick its own next pool of instances for labelling based on its ranking. The average performance across all labels is considered as the performance of binary relevance model in one active learning iteration.

Major drawbacks of BR methods: (a) Training several classifiers incurs a very high computational cost limiting its applicability to multilabel classification problems with fewer target labels (b) This method does not take the label structure information into consideration and trains K disjoint classifiers. To evaluate this classifier's performance, we compare it with a single MLP designed to handle the multilabel ExtraSensory data.

5.1.2 Algorithm Adaptation

Algorithm Adaptation methods extend single label classification algorithms to directly handle multilabel classification, for example, modifying the optimization/cost function. Multilabel decision tree is an adaptation of the C4.5 algorithm to construct a decision tree with multiple labels [Clare and King, 2001]. The entropy calculation for single label is extended to include the

entropies of all the individual labels. Zhang and Zhou [2007] describe a lazy learning approach that utilizes the maximum a posteriori (MAP) probability estimation to determine the target labels. A survey of MLC methods is presented in [Tsoumakas and Katakis, 2007]. Algorithms adapted to solve a specific problem are expected to perform better than problem transformation methods and are computationally less expensive. The MLP described in [Vaizman et al., 2018] is a unified model for multimodal multilabel context recognition in the wild. It optimizes an instance weighted version of MAP estimation using priors. Apart from that, active learning using MLP can also be considered a form of algorithm adaptation (described in Section 5.2) as it adapts traditional active learning strategies to handle multilabel active learning.

Multi Layer Perceptron

We first highlight the implementation details of the best Multilayer Perceptron described in [Vaizman et al., 2017] designed using algorithm adaptation. 175 statistical features of the six core sensors are selected from a total of 225 features for early fusion. Early fusion is a technique of combining sensor specific features into a single d-dimensional vector before training. Combining information from multimodal sensors can reduce misclassification by filling in feature information from other sensors during sensor dropouts. The features are standardized with their mean and standard deviation. We replicate the baseline architecture. The input layer has 175 neurons (one for each feature), and the output layer has ten neurons (one for each context). The input layer is followed by two hidden layers with 16 neurons each. A leaky ReLU activation is used for these layers except for the last output layer. The MLP outputs probability estimates for each of the context labels. We use a sigmoid activation on the output of the last layer with a cut off 0.5 to get the actual output for the context. The MLP finally outputs a binary value for each of the context labels. We train the classifier with a mini batch size of 300 for 40 epochs using SGD optimization with back propagation. The learning rate decreases linearly from 0.1 to 0.01 in every epoch.

Instance Weighting

Incorporating missing instances as negative examples in the training data impacts the classifier’s ability to learn the correct label correlations. When the data is skewed, the model trivially outputs 0 for all classes. To reduce the impact of missing labels on active learning, we retain the nontraditional instance weighting component from baseline implementation. It is computed for every instance label pair using the label wise class proportions and missing matrix of the data. It compensates for the distribution shifts and adapts the model to the target domain. If

the number of positive examples is less for a label, then misclassification of positive prediction is penalized more than the negative prediction. The modified optimization function also ignores the contribution of the missing example label pair in the cost computation.

$$\min_{\theta} \left(\frac{i}{NL} \left(\sum_{i=1}^N \sum_{l=1}^L \Psi c(f(X)_{i,l}, Y_{i,l}) \right) + \lambda \phi(\theta) \right) \quad (5.1)$$

where $c(y, \hat{y})$ is the traditional cross entropy loss and $\phi(\theta)$ is the Frobenius norm of the weighted matrix. The choice of sampling strategy affects the proportion of instances chosen for each label. Unlike fully supervised learning, the proportion of missing labels keeps changing with every new pool of training data added. The instance weights must be recomputed for every AL iteration.

We choose this MLP architecture as one of the base classifiers to be evaluated as (a) it is compact and suitable for deployment to mobile environments, (b) the hidden layers are shown to be effective to share the learned representations among labels through transfer learning, (c) can handle multimodal and missing sensor information with sensor fusion, (d) deals with imbalance using instance weighting, and (e) computationally less expensive compared to problem transformation methods as we train only one classifier.

5.2 Sampling Strategies

Key component of any active learning algorithm is the sampling strategy. Sampling strategies are device agnostic and work similarly on all devices. Sampling strategies fall under two categories: (a) Uncertainty Sampling (b) Query by Committee.

Uncertainty Sampling

Uncertainty sampling strategy selects points that the classifier is most uncertain about. The objective of any classification task is to identify the hyperplanes that separate different classes. The labels that the classifier can predict with high confidence are probably already correct. However, data instances near the class boundaries are usually confusing for classifiers. The classifiers are uncertain about the correct target class for those points. By selecting only these confusing instances and querying the domain expert for correct feedback, we can effectively reduce the size of the training data required.

Query by Committee

Query by Committee strategy [Melville and Mooney, 2004] selects examples by measuring the disagreement between an ensemble of classifiers. In each active learning iteration, a committee of classifiers is generated based on the current classifier. The classifiers predict the output labels for all the instances. Labels for examples about which these classifiers have most disagreement are acquired and added to the training set.

5.2.1 Issues

Traditional sampling strategies may not directly work for multilabel problems. The uncertainty of an instance cannot be determined based on how well the classifier performs on a single label. The MLP predicts the output probabilities for all the labels of unlabeled instances at once. To rank these instances, we need a different notion of uncertainty that integrates all the class membership probabilities. Consider an MLC problem of classifying two data points x_1 and x_2 with four possible output classes. The Least Confident Prediction (LCP) sampling strategy depends on the probability of the most probable class. Suppose the probability distributions are given by $x_1 - [0.25, 0.1, 0.2, 0.8]$ and $x_2 - [0.1, 0.3, 0.9, 0.6]$. In single label case, using least confident prediction, the classifier will have lesser confidence for the most probable class (4) of data point x_1 compared to the most probable class (3) for data point x_2 . Using this strategy, x_1 will be selected for labelling. However, in a multi label case, x_2 has two correct labels (3 & 4), is much harder to predict correctly and helps the classifier learn a new label combination. Here x_2 should be considered as more valuable compared to x_1 .

Margin based methods that compute the distance from the separating hyperplanes are also not applicable as there are multiple hyperplanes in multilabel classification. Directly extending these techniques from the single label AL literature may lead the classifier to incorrectly choose lesser valuable instances. Thus, the most important challenge of multilabel active learning is to develop an efficient strategy to measure total informativeness considering all the labels of an unlabeled instance. In our work, three uncertainty sampling strategies are explored to compute the informativeness for every label of an unlabeled instance. Additionally, we explore three alternatives of these strategies based on the label dimension.

5.2.2 Uncertainty Sampling

Least Confident Prediction

Definition 4.4.1 (LCP). Difference between 100% and the probability of the most confident prediction.

$$Unc_{LCP}(x_i, y_{i,j}) = 2 * (\operatorname{argmin}(P(y_{i,j} = 1|x_i), P(y_{i,j} = 0|x_i))) \quad (5.2)$$

The classifier is confident about its prediction for a label if the probability is 1. If the probability for a label is lesser than 1, it indicates that the classifier is less confident about its prediction for that label by an amount equal to the difference between 1 and this probability. In MLC, each label has two possible classes 0 or 1. LCP for a label is obtained subtracting the probability of the class with the most confident prediction (0 or 1) from 1. It is normalized to a 0–1 range by multiplying with a factor proportional to the number of classes (two in this case).

Normalized Entropy

Definition 4.4.2 (Entropy). A measure of the change that will be brought about by the unlabeled instance to the current state of the classifier.

$$Unc_{ENT}(x_i, y_{i,j}) = - \sum_{m \in \{0,1\}} P(y_{i,j} = m|x_i) \log(P(y_{i,j} = m|x_i)) \quad (5.3)$$

where Unc_{ENT} is the normalized entropy of the unlabeled instance, $P(y_{i,j})$ is the class membership probability estimate provided by the MLC. Normalized entropy can be used to gauge the classifier's confidence about its prediction for multi label classification [Park and Simoff, 2015]. Normalized entropy is used as an approximation of the label power set distribution.

Margin

Definition 4.4.3 (Margin). Margin sampling is an extension of Best-versus-Second-Best strategy. In BvSB, the difference between the probabilities of top two probable classes in MCC is used as the uncertainty for predicting that instance.

$$Unc(x_i) = (P(y_1|x) - P(y_2|x)) \quad (5.4)$$

where y_1 is the best and y_2 is the second-best class. In our case, there are only two classes – positive and negative for each label and the difference between the two is the uncertainty for

that label.

$$Unc_{MAR}(x_i, y_{i,j}) = (P(y_{i,j} = 1|x) - P(y_{i,j} = 0|x)) \quad (5.5)$$

5.3 Balanced Sampling

We analyze the data distribution of the various sampling strategies across active learning iterations. Figure 3.2 shows that certain context labels such as *laundry*, *elevator*, *at a bar* have fewer positive examples than negative examples. These labels constitute the minority labels of the dataset. Charte et al. [2015a] propose a metric to measure the imbalance ratio in a multilabel dataset (MLD). Although this metric is highly useful to measure imbalance in multiclass datasets, it can provide insights into the distribution of positive examples across classes in multilabel datasets.

Definition 5.3.1 (Mean Imbalance Ratio). Imbalance ratio per label is the ratio between majority label and any label y . The mean imbalance ratio offers an average imbalance level in the multilabel dataset.

$$IRLbl(y) = \frac{\operatorname{argmax} \left(\sum_{i=1}^{|D|} h(l', Y_i) \right)}{\sum_{i=1}^{|D|} h(l, Y_i)}, h(l, Y_i) = \begin{cases} 1 & l \in Y_i, \\ 0 & l \notin Y_i \end{cases} \quad (5.6)$$

$$MeanIR = \frac{1}{|L|} \sum_{l=L_1}^{L_{|L|}} (IRLbl(l)) \quad (5.7)$$

Higher MeanIR indicates higher imbalance across labels. MeanIR for the ExtraSensory Dataset is **92.74**. On an average, the label with the maximum number of positive examples is 92 times greater than other labels. We try two techniques to maintain (a) equal number of positive and negative examples, (b) equal number of examples for all classes, (c) minimum number of positive examples in each class.

5.3.1 Over Sampling

It is possible that the random strategy selects enough positive examples for minority labels while the example-based sampling strategies miss them. With more data about the minority labels, the random strategy might be able to learn well about those labels and have more equal number of samples.

MLSMOTE is an oversampling technique used for multiclass and multilabel data. In MLSMOTE, minority labels that have $IRLbl$ greater than MeanIR are first identified. The nearest neighbors of the positive points of the minority labels are generated and the labels are copied over from the available labelled training data. Synthetic generation of instances for oversampling using MLSMOTE [Charte et al., 2015b] cannot be directly applied to this dataset as we are dealing with missing values. MLSMOTE is modified to oversample new instances along with missing labels. The missing values are also copied over to the synthetic instances. While we rely on instance weighting and oversampling to handle the skew within a label, we need to reduce the effect of imbalance across labels. We explain a method to achieve this in the following section.

5.3.2 Hyperparameter for balancing

A rudimentary balancing technique is explored by maintaining a constant proportion of positive examples for all context labels in each round of active learning. This technique is adopted to ensuring that all labels have a minimum number of positive samples to learn from.

The threshold parameter, t is tunable, and the effect of tuning t is discussed in Chapter 7. The minimum threshold is increased to see if active learning performs well with enough positive samples for each label. The ExtraSensory Dataset has characteristics that does not allow the example-based sampling strategies including diversity sampling to choose effective instances. Steps to eliminate skew are counter effective possibly due to the label correlations.

5.3.3 Informativeness

A simple unification technique is to use the mean of informativeness of all the labels as the total informativeness of the instance. In our experiments we compute Mean LCP, Mean Entropy and Mean Margin for every unlabeled instance.

$$Unc(x_i) = \frac{1}{L} \left(\sum_{y_j \in Y} Unc(x_i, y_{i,j}) \right) \quad (5.8)$$

In addition to mean, we can also consider the maximum or minimum value of informativeness among all the labels as the informativeness of an instance. Intuitively, if one of the labels of an instance has a higher entropy or lcp value, it implies that the instance contributes more to that label. Rank order of this measure is the quantification of the maximum change that will be brought by labelling this instance to any label.

This is slightly different for BvSB. A larger margin implies the classifier is more confident about its prediction. We are interested in the use of minimum margin as it indicates higher uncertainty. Thus, the instances are ranked in the increasing order of minimum margin. Two variants along the label dimension for each of the three strategies gives a total of six sampling strategies.

Ramirez-Loaiza et al. [2017] observe through their empirical study that the best active learning algorithm varies based on the dataset and the performance measure of interest. We investigate another metric Macro Balanced Accuracy (described in Chapter 6) to get a complete understanding of why active learning performs only at par with random sampling with algorithm adapted model.

CHAPTER

6

EVALUATION METRICS

Overall performance is usually presented by the standard evaluation metrics such as precision, recall, F1 scores, and ROC area. For an MLC problem, predictions can be fully correct, partially correct, or fully incorrect. This makes evaluation challenging than SLC. Confusion matrices show the interclass variability in performance and the detailed distribution of true and false positive rates. As our dataset is highly skewed both across and within labels, we can expect our test set to have skewed ratios of representative examples both within a single class and across classes.

6.1 Balanced Accuracy

Naive classification accuracy applied to this test set will yield misleading optimistic estimates. For labels with very few positive examples, a trivial classifier that always declares false will have very high overall accuracy. Although the instance weighting element adjusts the cost of misclassification and is expected to reduce this bias, we need a fair metric incorporating both positive and negative classes across all context labels. Pereira et al. [2018] provide a thorough analysis of evaluation measures and give concrete suggestions for choosing evaluation measures

for MLCs. Vaizman et al. [2018] use a generalized version of accuracy called the balanced accuracy (BA) given by Equation 6.1 for measuring their MLP performance. BA is not very sensitive to class skew and suppresses inconsistent trends by equally considering the contribution of positive/negative classes.

$$BA = \frac{\left(\frac{TP}{P} + \frac{TN}{N}\right)}{2} \quad (6.1)$$

6.2 Macro Balanced Accuracy

Although BA is a comparatively fair version of naive accuracy, it only measures the overall performance of the model across all labels in MLC. This metric does not treat all labels equally and is greatly influenced by the labels whose correct predictions dominate the test set. To compare the performance of BR and MLP, we need a metric that is applicable to both methods. In BR method, we average the BA across all labels. We can use a similar metric called the Macro BA for MLP. Macro BA metric (Equation 6.2) is the average of label wise BA and gives equal weightages for all labels. We analyze how AL affects the classifier’s Macro BA for every label in addition to the overall BA.

$$MacroBA = \frac{1}{L} \sum_{i \in L} BA_i \quad (6.2)$$

CHAPTER

7

EXPERIMENTS AND RESULTS

7.1 Active Learning Evaluation

In this section, we investigate RQ1: *How can we reduce the labelling effort for multi label context recognition in the wild?* We primarily focus on the base classifier selection for active learning pipeline.

7.1.1 Fully Supervised Models

As a first step, we set up experiments to discuss the tradeoffs between two multilabel classifiers. We train a fully supervised single MLP classifier with ten output neurons and ten binary classifiers with one output neuron. A 70–30 split of train and test data is used. Experiments are repeated for 10 runs and the average balanced accuracy is reported. In binary relevance, both balanced accuracy and macro balanced accuracy are the same as there is only one output label. It is compared with balanced accuracy and macro balanced accuracy of MLP. Macro BA for MLP is computed as the average of label wise BA where true positives, true negatives, false positives, and false negatives for each label are obtained from the multilabel output. Both BA and Macro BA turn out to be the same for MLP (0.85). Fully supervised binary classifiers trained for a

specific label outperform the fully supervised MLP by **4%** (Table 7.1).

Table 7.1: Balanced accuracy for problem transformation (label wise, average) and algorithm adaptation with 10 labels.

Labels	PT	AA
Or_indoors	0.92	0.89
Loc_home	0.88	0.81
Sitting	0.81	0.76
Fix_walking	0.84	0.81
Watching_tv	0.87	0.77
Or_outside	0.93	0.89
Bicycling	0.92	0.85
Doing_laundry	0.91	0.81
At_a_bar	0.97	0.90
Elevator	0.89	0.80
Average BA	0.89	0.85

7.1.2 Active Learning

We expect any model trained with active learning to achieve the balanced accuracy of fully supervised models with lesser number of labelled examples. We set up an experiment with a 70–30 test train split on the dataset. Both binary relevance and MLP are trained initially with 2% of total training data. The effect of varying warm seed size is discussed in 7.2.4. In each AL iteration, a pool of 1000 labelled instances are added to the labelled data and both classifiers are retrained.

Balanced Accuracy

We summarize the total number of labelled examples required by both models to achieve best performance using different sampling strategies in Table 7.2. LCP strategy is effective in reducing the number of queries with both multilabel models. One observation is that the MLP reaches its full BA with lesser number of queries compared to BR model using LCP (**11%** lesser queries).

Unlike margin and LCP sampling strategies, we do not observe a significant reduction in the

Table 7.2: Percentage of labelled data with random, mean margin, entropy, and lcp strategies for problem transformation and algorithm adaptation.

Labels	Random	Margin	Entropy	LCP
Or_indoors	0.46	0.27	0.46	0.25
Loc_home	0.17	0.10	0.15	0.9
Sitting	0.19	0.9	0.17	0.11
Fix_walking	0.50	0.33	0.51	0.29
Watching_tv	0.51	0.29	0.68	0.28
Or_outside	0.68	0.37	0.56	0.32
Bicycling	0.75	0.30	0.63	0.28
Doing_laundry	0.95	1.00	1.00	0.64
At_a_bar	0.49	0.34	0.42	0.37
Elevator	0.57	1.00	1.00	0.50
PT (0.89)	0.56	0.33	0.65	0.30
AA – BA (0.85)	0.29	0.30	0.83	0.19
AA – MBA (0.85)	0.53	0.52	0.79	0.50

number of labelled examples using entropy strategy for both models as shown in Figure 7.1. Although the difference in actual percentages of the queries across strategies is significant, there is minimal gap in the performance (balanced accuracy) of random and LCP strategies from 20% to 30% of the queries as shown in Figure 7.1 which is further investigated in the following section.

Our findings from binary relevance data show that for labels *doing laundry*, *at a bar*, and *elevator*, almost all the data is used up by the sampling strategies to reach the full BA with binary relevance. This indicates a possible correlation between the proportion of positive labels to the total number of labelled examples required.

Macro Balanced Accuracy

More labelled examples are required to achieve the Macro BA of the fully supervised models with MLP compared to problem transformation method. With either metric, LCP strategy outperforms other strategies. Macro BA is higher than average BA for binary relevance as it considers all labels simultaneously and does not have the flexibility to choose instances useful for each label separately.

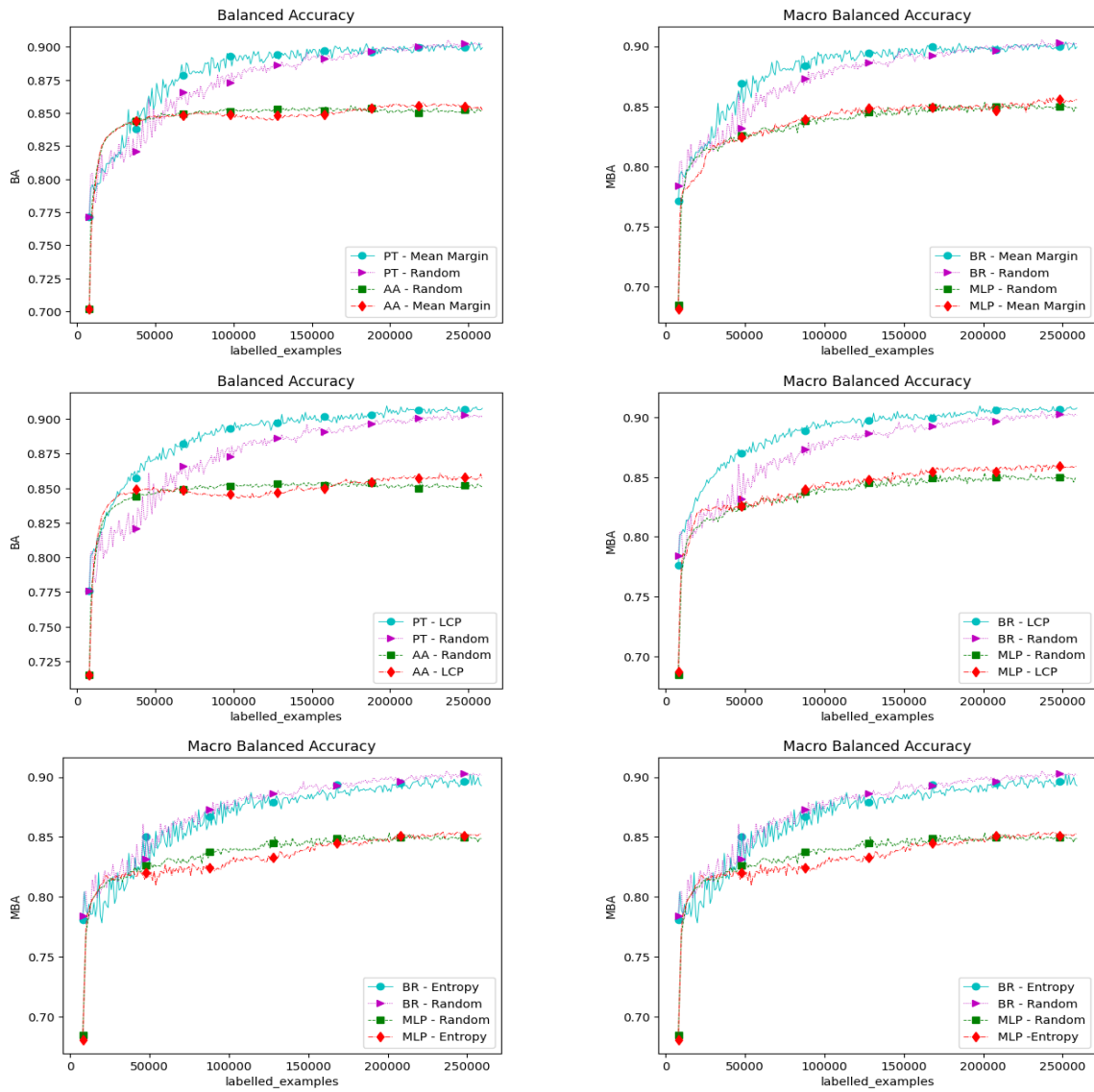


Figure 7.1: Balanced accuracy and macro balanced accuracy for PT and AA with all strategies.

Trade-off

BR methods give higher overall performance at the cost of training 10 different classifiers (one for each label). The MLP is a single classifier that can simultaneously predict the output labels for all contexts. With lesser computational cost, it performs comparably well on the ExtraSensory Dataset (we see only a 4% difference in their fully supervised performances). For context sensitive applications, BR models with active learning can work well. If resources are constrained, we could use the less accurate yet less compute intensive MLP.

7.2 Sampling strategies

In this section we investigate RQ2: *What are the methods to improve the performance of active learning for a unified model designed to handle multi label context recognition in the wild (with missing values and multimodal sensor features) ?*

Figure 7.2 indicates that the random sampling strategy performs comparably well with other sampling strategies in terms of BA and Macro BA. This is unlike the single label case where the strategies use up fewer queries compared to random sampling to reach their best accuracy. We hypothesize two possible reasons:

1. characteristics of in-the-wild dataset such as skew could be hindering the active learning algorithm from learning the right contexts.
2. example-based sampling strategies are unable to capture the informativeness of the instances fully.

To verify and eliminate these reasons, we try balanced sampling and evaluate other variants of the traditional sampling strategies.

7.2.1 Balanced Sampling

We first set up an experiment to balance out the proportion of positive examples for each label in every iteration. We compare the performance of the sampling strategies when the minimum number of positive examples per label, t is set to 10, 20, 30 and 50. For random sampling, t positive examples for each label are chosen at random. For the other three strategies, the unlabeled instances are ranked based on their informativeness and the top t instances for each label are chosen. Increasing the threshold value decreases the total number of labelled examples required to reach full BA until $t = 30$. This balancing proves to be detrimental beyond $t = 50$.

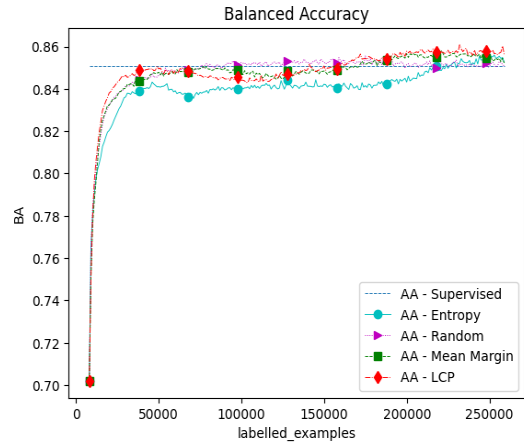


Figure 7.2: Balanced accuracy for algorithm adaptation with various sampling strategies.

At this stage, the less useful instances start getting selected that cause more damage than the benefits accrued using the sampling strategies.

In this experiment, it is guaranteed that all strategies learn from a minimum positive proportion of data (Figure 7.3). Even when maintaining a minimum number, the addition of positive samples for one context label inadvertently increases the positive samples for other relevant labels affecting the correlation among different labels. This calls for better insights into the actual examples that are selected by the various strategies.

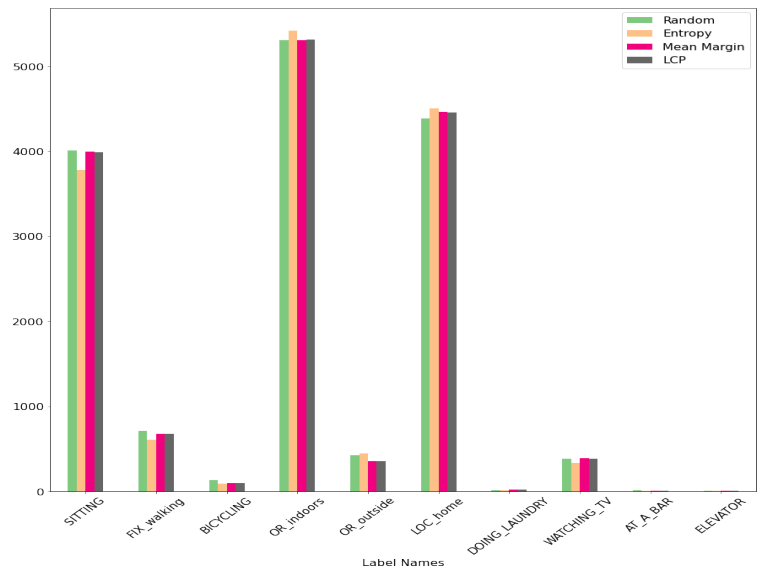


Figure 7.3: Proportion of positive examples per label with balanced sampling.

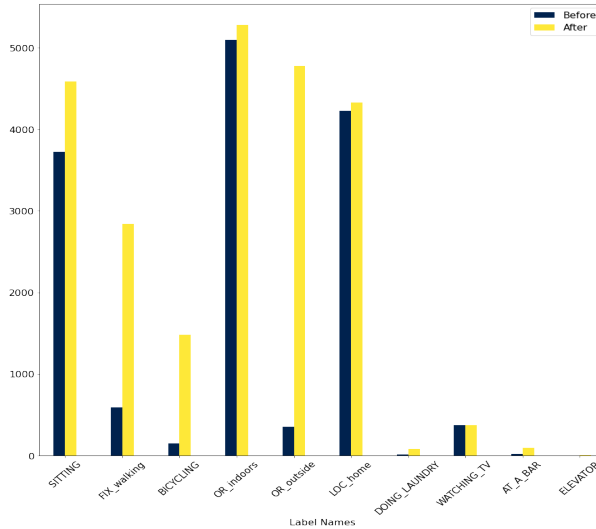


Figure 7.4: Proportion of positive examples before and after over sampling using MLSMOTE.

7.2.2 MLSMOTE

MLSMOTE algorithm works inefficiently for the sampled dataset. The minority labels for this dataset are bicycling, outside, laundry, at a bar, and elevator. After oversampling, the number of positive instances has increased for the labels *bicycling*, *walking*, and *outside* as shown in Figure 7.4. Increase in the instances of walking was unintended. The labels with less than 1000 positive examples do not get oversampled properly. The skew across labels is never eliminated even after oversampling the minority labels. We conclude that the highly skewed distribution of data in the wild does not allow our models to completely leverage the benefits of active learning algorithms.

7.2.3 Informativeness

Table 7.4 summarizes the total number of labelled examples required by the three sampling strategies and their variants described in Section 5.2. Min Margin strategy requires lesser proportion of labelled examples compared to other strategies to achieve full BA. The samples are ranked based on the minimum margin between the positive and negative classes for all the labels. Eliminating the uncertainty for just one of the labels helps the classifier learn better about the instance compared to mean margin. Similarly, Max LCP is better compared to all the other strategies in achieving full Macro BA. These results indicate that the classifier is probably only

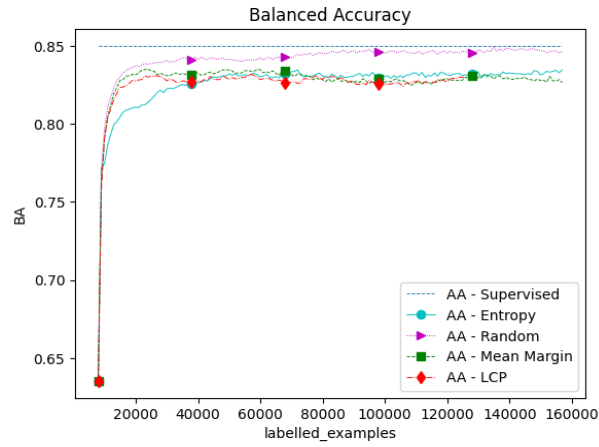


Figure 7.5: Balanced accuracy for various sampling strategies with MLSMOTE.

Table 7.3: Proportion of labelled examples with $t = 10, 20, 30$ and 50 .

Strategy	BA	Macro BA
Random + 10	0.17	0.64
Random + 20	0.10	0.44
Random + 30	0.9	0.38
Random + 50	0.9	0.48
Mean Margin + 10	0.69	0.74
Mean Margin + 20	0.14	0.73
Mean Margin + 30	0.10	0.70
Mean Margin + 50	0.35	1.0
Mean Entropy + 10	0.29	0.83
Mean Entropy + 20	0.66	0.82
Mean Entropy + 30	0.35	0.60
Mean Entropy + 50	0.76	1.0
Mean LCP + 10	0.25	0.84
Mean LCP + 20	0.19	0.97
Mean LCP + 30	0.20	0.59
Mean LCP + 50	0.35	0.95

uncertain about one or a few of its labels. The useful uncertainty information of one of the labels is probably lost in the mean strategy.

Table 7.4: Proportion of labelled data for six variants of informativeness.

Strategy	BA	Macro BA
Random	0.29	0.50
Mean Margin	0.30	0.52
Min Margin	0.17	0.52
Mean Entropy	0.82	0.79
Max Entropy	0.80	0.71
Mean LCP	0.19	0.53
Max LCP	0.20	0.31

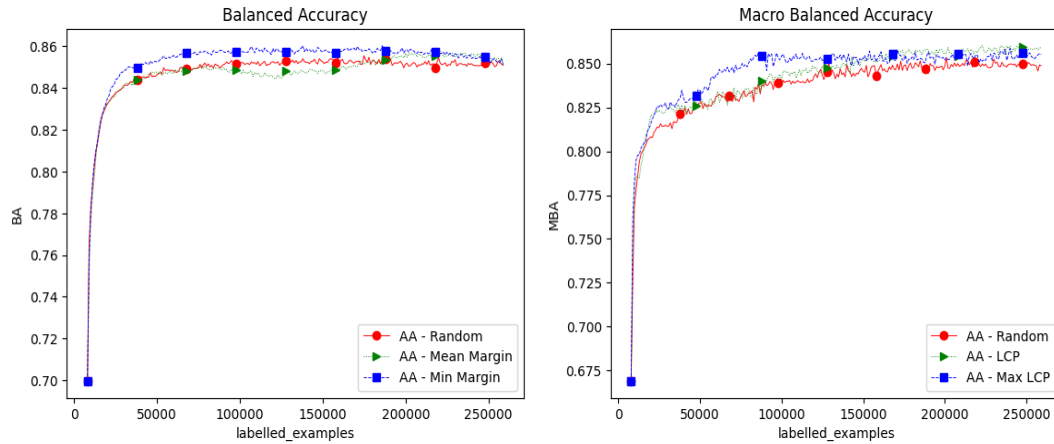


Figure 7.6: Balanced accuracy with min margin and macro balanced accuracy with max lcp for AA.

7.2.4 Warm Seed

We additionally conduct experiments to observe the effect of varying the initial warm seed size. Training is repeated with 5% and 10% of initial warm seed data for five runs. With 5% of initial training, all strategies including random strategy help the classifiers achieve full BA much faster than 2% initial training. With more data, the classifiers tend to have a better mapping of the underlying data distribution. At 10% initial training, all strategies require lesser labelled examples compared to random strategies. However, we observe an increase in the total labelled examples in comparison with 2% and 5% case. As random annotation increases, strategies are shown to require more examples to come to effect.

Table 7.5: Proportion of labelled data for different sampling strategies with 5% (top four rows) and 10% (bottom four rows) of initial warm seed training with algorithm adaptation model.

Strategy	BA	Macro BA
Random	0.21	0.45
Mean Margin	0.18	0.65
Entropy	0.15	0.51
LCP	0.16	0.63
Random	0.52	0.49
Mean Margin	0.24	0.60
Entropy	0.23	0.66
LCP	0.24	0.54

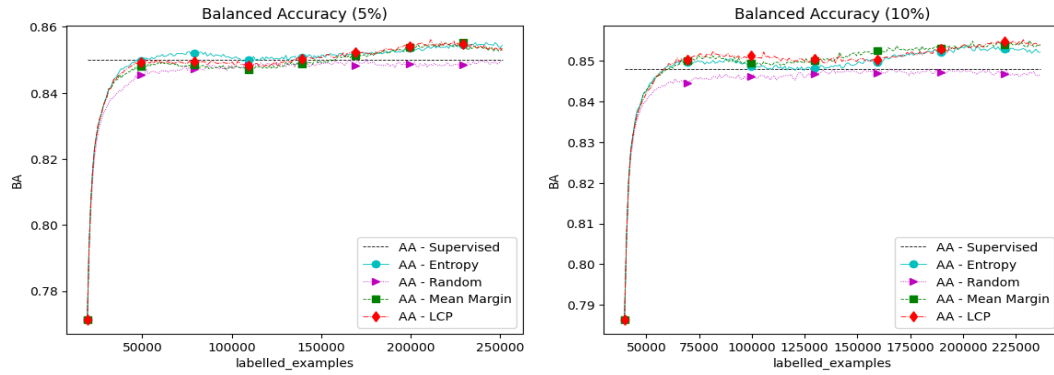


Figure 7.7: Balanced accuracy for algorithm adaptation model with different sampling strategies and 5% and 10% initial warm seed training.

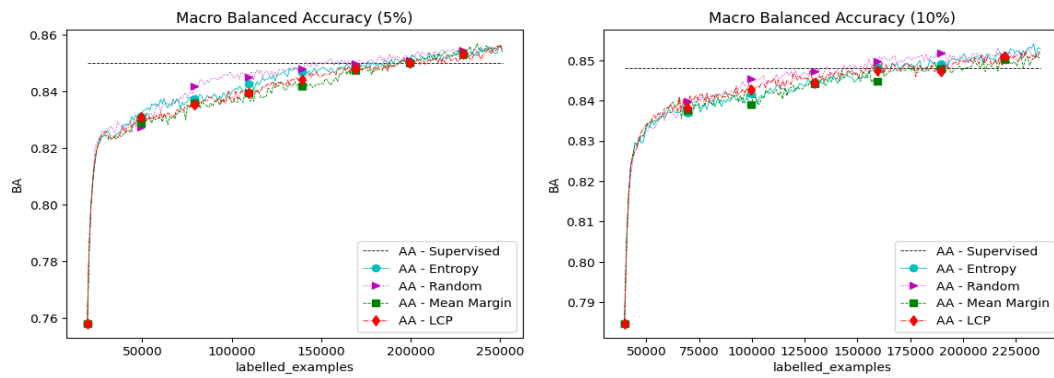


Figure 7.8: Macro balanced accuracy for algorithm adaptation model with different sampling strategies and 5% and 10% initial warm seed training.

CHAPTER

8

CONCLUSION

8.1 Summary

In our work, we present a detailed analysis of active learning for context recognition in the wild with problem transformation and algorithm adaptation base classifiers. We show that MLP is computationally less expensive and performs comparably well with active learning. We then compare the performance of six different sampling strategies and show that mean lcp and min margin strategies achieve full BA with lesser number of labelled examples. Finally, we analyze reasons as to why random sampling performs equally well as other strategies by exploring another metric called Macro BA and summarize effects of few balancing techniques. We believe our work will enable further research in the application of active learning to context recognition in-the-wild.

8.2 Future Work

As the ExtraSensory dataset is very large, we demonstrate our results for 10 labels with different skews. The primary limitation of our work is straitened exploration of potential solutions to

enlisted problems based on the 10 labels. In addition to conducting a thorough analysis using all 51 labels, we plan to perform a comparative study of other example based sampling strategies on both PT and AA models to conclusively comment on the performance of traditional AL strategies for multi label context recognition in the wild. We also plan to:

1. Study the effect of missing annotations in active learning.
2. Exploit correlations along label dimension to design an effective sampling strategy for natural settings.
3. Combine semi supervised learning with active learning to further reduce labelling cost.

REFERENCES

- Dulal Acharjee, SP Maity, and Amitava Mukherjee. Hidden markov model a tool for recognition of human contexts using sensors of smart mobile phone. *Microsystem Technologies*, 23(3): 571–582, 2017.
- Rebecca Adaimi and Edison Thomaz. Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–23, 2019.
- Salikh Bagaveyev and Diane J. Cook. Designing and evaluating active learning methods for activity recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, page 469–478, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330473. doi: 10.1145/2638728.2641674. URL <https://doi.org/10.1145/2638728.2641674>.
- Claudio Bettini, Gabriele Civitarese, and Riccardo Presotto. Caviar: Context-driven active and incremental activity recognition. *Knowledge-Based Systems*, 196:105816, 2020.
- Anna L Beukenhorst, Ella Collins, Katherine M Burke, Syed Minhajur Rahman, Margaret Clapp, Sai Charan Konanki, Sabrina Paganoni, Timothy M Miller, James Chan, Jukka-Pekka Onnela, et al. Smartphone data during the covid-19 pandemic can quantify behavioral changes in people with als. *Muscle & nerve*, 63(2):258–262, 2021.
- Haixia Bi, Miquel Perello-Nieto, Raul Santos-Rodriguez, and Peter Flach. Human activity recognition based on dynamic active learning. *IEEE Journal of Biomedical and Health Informatics*, 25(4):922–934, 2021. doi: 10.1109/JBHI.2020.3013403.
- Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 59–66, Washington, DC, USA, 2003. AAAI Press. ISBN 1577351894.
- Klaus Brinker. On active learning in multi-label classification. In Myra Spiliopoulou, Rudolf Kruse, Christian Borgelt, Andreas Nürnberger, and Wolfgang Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages 206–213, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-31314-4.

- Caroline A Byrne, Rem Collier, and Gregory MP O’Hare. A review and classification of assisted living systems. *Information*, 9(7):182, 2018.
- Liang Cao, Yufeng Wang, Bo Zhang, Qun Jin, and Athanasios V Vasilakos. Gchar: An efficient group-based context—aware human activity recognition on smartphone. *Journal of Parallel and Distributed Computing*, 118:67–80, 2018.
- Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neuro-computing*, 163:3–16, 2015a.
- Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015b.
- Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I Hong, and Yuvraj Agarwal. Does this app really need my location? context-aware privacy management for smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3): 1–22, 2017.
- Amanda Clare and Ross D. King. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD ’01, page 42–53, Berlin, Heidelberg, 2001. Springer-Verlag. ISBN 3540425349.
- Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. Transfer learning for activity recognition: A survey. *Knowledge and information systems*, 36(3):537–556, 2013.
- Federico Cruciani, Ian Cleland, Chris Nugent, Paul McCullagh, Kåre Synnes, and Josef Hallberg. Automatic annotation for human activity recognition in free living using a smartphone. *Sensors*, 18(7):2203, 2018.
- Federico Cruciani, Chen Sun, Shuai Zhang, Chris Nugent, Chunping Li, Shaoxu Song, Cheng Cheng, Ian Cleland, and Paul Mccullagh. A public domain dataset for human activity recognition in free-living conditions. In *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 166–171, Leicester, UK, Aug 2019. IEEE. doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00071.

- Zachary A. Daniels and Dimitris N. Metaxas. Addressing imbalance in multi-label classification using structured hellinger forests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'17, page 1826–1832. AAAI Press, 2017.
- Muhammad Ehatisham-ul Haq, Muhammad Awais Azam, Yusra Asim, Yasar Amin, Usman Naeem, and Asra Khalid. Using smartphone accelerometer for human physical activity and context recognition in-the-wild. *Procedia Computer Science*, 177:24–31, 2020.
- Andrea Esuli and Fabrizio Sebastiani. Active learning strategies for multi-label text classification. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, pages 102–113, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-00958-7.
- Kaori Fujinami, Trang Thuy Vu, and Koji Sato. A framework for human-centric personalization of context recognition models on mobile devices. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCOM/CyberSciTech)*, pages 885–888, Fukuoka, Japan, 2019. IEEE, IEEE Computer Society.
- Wen Ge and Emmanuel Agu. Cruft: Context recognition under uncertainty using fusion and temporal learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 747–752, Miami, FL, USA, 2020. IEEE, IEEE.
- Rebeen Ali Hamad, Longzhi Yang, Wai Lok Woo, and Bo Wei. Joint learning of temporal models to handle imbalanced data for human activity recognition. *Applied Sciences*, 10(15): 5293, 2020.
- Manhyung Han, La T. Vinh, Young-Koo Lee, and Sungyoung Lee. Comprehensive context recognizer based on multimodal sensors in a smartphone. *Sensors*, 12(9):12588–12605, 2012. URL <https://proxying.lib.ncsu.edu/index.php/login?url=https://www-proquest-com.prox.lib.ncsu.edu/scholarly-journals/comprehensive-context-recognizer-based-on/docview/1537596852/se-2?accountid=12725>. Copyright - Copyright MDPI AG 2012; Last updated - 2018-10-06.
- Mahmudul Hasan, Sujoy Paul, Anastasios I Mourikis, and Amit K Roy-Chowdhury. Context-

- aware query selection for active learning in event recognition. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):554–567, 2018.
- Md Ismail Hossen, Goh Kah Ong Michael, Tee Connie, Siong Hoe Lau, and Ferdous Hossain. Smartphone-based context flow recognition for outdoor parking system with machine learning approaches. *Electronics*, 8(7):784, 2019.
- Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, 2014.
- Marc Kurz and Alois Ferscha. Sensor abstractions for opportunistic activity and context recognition systems. In Paul Lukowicz, Kai Kunze, and Gerd Kortuem, editors, *Smart Sensing and Context*, pages 135–148, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16982-3.
- Matthew L Lee and Anind K Dey. Sensor-based observations of daily living for aging in place. *Personal and Ubiquitous Computing*, 19(1):27–43, 2015.
- Lu Li, Yang Li, Xiangxiang Xu, and Lin Zhang. A maximal correlation embedding method for multilabel human context recognition: Poster abstract. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks, IPSN '19*, page 305–306, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362849. doi: 10.1145/3302506.3312601. URL <https://doi-org.prox.lib.ncsu.edu/10.1145/3302506.3312601>.
- Shao-Yuan Li, Yuan Jiang, Nitesh V Chawla, and Zhi-Hua Zhou. Multi-label learning from crowds. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1369–1382, 2018.
- Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.
- Montassar Ben Messaoud, Ilyes Jenhani, Nermine Ben Jemaa, and Mohamed Wiem Mkaouer. A multi-label active learning approach for mobile app user review classification. In Christos Douligeris, Dimitris Karagiannis, and Dimitris Apostolou, editors, *Knowledge Science, Engineering and Management*, pages 805–816, Cham, 2019. Springer International Publishing. ISBN 978-3-030-29551-6.

- Pradeep K Murukannaiah and Munindar P Singh. Platys: An active learning framework for place-aware application development and its evaluation. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 24(3):1–32, 2015.
- Annamalai Natarajan, Gustavo Angarita, Edward Gaiser, Robert Malison, Deepak Ganesan, and Benjamin M. Marlin. Domain adaptation methods for improving lab-to-field generalization of cocaine detection using wearable ecg. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '16*, page 875–885, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616. doi: 10.1145/2971648.2971666. URL <https://doi.org/10.1145/2971648.2971666>.
- Ky Trung Nguyen, François Portet, and Catherine Garbay. Dealing with Imbalanced data sets for Human Activity Recognition using Mobile Phone sensors. In *3rd International Workshop on Smart Sensing Systems*, Rome, Italy, June 2018. URL <https://hal.archives-ouvertes.fr/hal-01950472>.
- Abayomi Moradeyo Otebolaku and Maria Teresa Andrade. User context recognition using smartphone sensors and classification models. *Journal of Network and computer applications*, 66:33–51, 2016.
- Laurence A. F. Park and Simeon Simoff. Using entropy as a measure of acceptance for multi-label classification. In Elisa Fromont, Tijn De Bie, and Matthijs van Leeuwen, editors, *Advances in Intelligent Data Analysis XIV*, pages 217–228, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24465-5.
- Ling Pei, Robert Guinness, Ruizhi Chen, Jingbin Liu, Heidi Kuusniemi, Yuwei Chen, Liang Chen, and Jyrki Kaistinen. Human behavior cognition using smartphone sensors. *Sensors*, 13(2):1402–1424, 2013. URL <https://proxying.lib.ncsu.edu/index.php/login?url=https://www-proquest-com.prox.lib.ncsu.edu/scholarly-journals/human-behavior-cognition-using-smartphone-sensors/docview/1537480895/se-2?accountid=12725>. Copyright - Copyright MDPI AG 2013; Last updated - 2018-10-05.
- Rafael B Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz HC Merschmann. Correlation analysis of performance measures for multi-label classification. *Information Processing & Management*, 54(3):359–369, 2018.

- Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–27, 2018.
- Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data mining and knowledge discovery*, 31(2):287–313, 2017.
- Tifenn Rault, Abdelmadjid Bouabdallah, Yacine Challal, and Frédéric Marin. A survey of energy-efficient context recognition systems using wearable sensors for healthcare applications. *Pervasive and Mobile Computing*, 37:23–44, 2017.
- Oscar Reyes, Carlos Morell, and Sebastián Ventura. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273:494–508, 2018.
- Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Synthesizing and reconstructing missing sensory modalities in behavioral context recognition. *Sensors*, 18(9):2967, 2018.
- Aaqib Saeed, Tanir Ozcelebi, Stojan Trajanovski, and Johan J. Lukkien. End-to-end multi-modal behavioral context recognition in a real-life setting. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8, Ottawa, ON, Canada, July 2019. IEEE.
- Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.
- Qiang Shen, Stefano Teso, Wanyi Zhang, Hao Xu, and Fausto Giunchiglia. Multi-modal subjective context modelling and recognition, 2020.
- Muhammad Shoaib, Stephan Bosch, Ozlem D. Incel, Hans Scholten, and Paul J. M. Havinga. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, 16(4):426, 2016. URL <https://proxying.lib.ncsu.edu/index.php/login?url=https://www-proquest-com.prox.lib.ncsu.edu/scholarly-journals/complex-human-activity-recognition-using/docview/1780818465/se-2?accountid=12725>. Copyright - Copyright MDPI AG 2016; Last updated - 2018-10-05.
- Wesllen Sousa, Eduardo Souto, Jonatas Rodrigues, Pedro Sadarc, Roozbeh Jalali, and Khalil El-Khatib. A comparative analysis of the impact of features on human activity recognition

- with smartphone sensors. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web, WebMedia '17*, page 397–404, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350969. doi: 10.1145/3126858.3126859. URL <https://doi-org.prox.lib.ncsu.edu/10.1145/3126858.3126859>.
- Konstantin Sozykin, Stanislav Protasov, Adil Khan, Rasheed Hussain, and Jooyoung Lee. Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 146–151, Busan, Korea, 2018. IEEE, IEEE Computer Society.
- Grigorios Tsoumakos and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
- Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing*, 16(4):62–74, 2017.
- Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018.
- Deepak Vasisht, Andreas Damianou, Manik Varma, and Ashish Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, page 472–481, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623759. URL <https://doi-org.prox.lib.ncsu.edu/10.1145/2623330.2623759>.
- Meng Wang and Xian-Sheng Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):1–21, 2011.
- Jian Wu, Victor S. Sheng, Jing Zhang, Hua Li, Tetiana Dadakova, Christine Leon Swisher, Zhiming Cui, and Pengpeng Zhao. Multi-label active learning algorithms for image classification: Overview and future promise. *ACM Comput. Surv.*, 53(2), March 2020. ISSN 0360-0300. doi: 10.1145/3379504. URL <https://doi-org.prox.lib.ncsu.edu/10.1145/3379504>.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xiangliang Zhang. Cmal: Cost-effective multi-label active learning by querying subexamples. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020. ISSN 1558-2191. doi: 10.1109/TKDE.2020.3003899.

Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202, 2018.

APPENDIX

APPENDIX

A

ACRONYMS

A summary of all acronyms is documented in Table A.1.

Table A.1: A summary of acronyms used in alphabetical order.

Acronym	Abbreviation
Active Learning	AL
Binary Relevance	BR
Balanced Accuracy	BA
Exact Match Ratio	EMR
Macro Balanced Accuracy	Macro BA (or) MBA
Multi Class Classification	MCC
Multi Label Active Learning	MLAL
Multi Label Classification	MLC
Multi Label Dataset	MLD
Multi Layer Perceptron	MLP
Problem Transformation	PT
Single Label Classification	SLC