JADAVPUR UNIVERSITY

MASTER DEGREE THESIS

Wi-Fi based Indoor Localization Subject to Varying Granularity and Device Heterogeneity

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Engineering

in

Software Engineering
Jadavpur University

By
Mausam Kundu

Class Roll No.: 001711002018

Examination Roll No.: M4SWE19019

University Registration No.: 140974 of 2017-18

Department of Information Technology

Under the Guidance of
Dr. Chandreyee Chowdhury

Department of Computer Science and Engineering
Faculty of Engineering and Technology,
Jadavpur University, Kolkata

May 2019

To whom it may concern

This is to certify that the work in this thesis entitled ***"Wi-Fi based Indoor Localization Subject to Varying Granularity and Device Heterogeneity"*** has been satisfactorily completed by ***Mausam Kundu*** (Examination Roll No. M4SWE19019, University Registration No. 140974 of 2017-18). It is a bonafide piece of work carried out under my supervision at *Jadavpur University, Kolkata*, for partial fulfillment of the requirements for awarding of the ***Master of Engineering*** in ***Software Engineering*** degree of the ***Department of Information Technology, Faculty of Engineering and Technology, Jadavpur University*** during the academic session 2017-2019.

Dr. Chandreyee Chowdhury
Assistant Professor,
Department of Computer Science and Engineering,
Jadavpur University,
Kolkata.

Forwarded By:

Dr. Bhaskar Sardar
Head of the Department,
Department of Information Technology,
Jadavpur University,
Kolkata.

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology,
Jadavpur University,
Kolkata.

Certificate of Approval

This is to certify that the thesis entitled ***"Wi-Fi based Indoor Localization Subject to Varying Granularity and Device Heterogeneity"*** is a bona-fide record of work carried out by ***Mausam Kundu*** in partial fulfillment of the requirements for the award of the degree of ***Master of Engineering in Software Engineering*** in the ***Department of Information Technology, Jadavpur University*** during the period August 2017 to May 2019. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Examiners:

____

(Signature of Examiner)
Date:

____

(Signature of Examiner)
Date:

Declaration of Originality and Compliance of Academic Ethics

I, Mausam Kundu, declare that this thesis titled, “Wi-Fi based Indoor Localization Subject to Varying Granularity and Device Heterogeneity” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:
____

Date:
____

Abstract

Nowadays, most of the people in outdoor use Global Positioning System (GPS) for navigating to a particular place, finding any location and so on. However, in indoor GPS cannot be used due to its high error range. Researches are going on in the field of Indoor localization as it seems an alternative to GPS in indoor and shows a lot of promise.

Various technologies are available in Indoor Localization such as Ultra Wide Band (UWB), Bluetooth, Wireless-Fidelity (Wi-Fi), but Wi-Fi is preferred among these technologies due to its availability and low cost. Many applications based on Indoor localization system (ILS) are already in use, but precise indoor positioning subject to varying ambient conditions is still out of reach. In fact, precision is found to play a key role in selecting important APs also. Thus, in this thesis our contribution is twofold. First of all, a stable AP selection algorithm is proposed subject to different positioning granularity. The second contribution is to design an ensemble of conditional classifiers for indoor localization that addresses the problem of device and context heterogeneity. The experiments are conducted based on data collected from the Computer Science and Engineering Department, Jadavpur University. The performance of different machine learning algorithms under different indoor conditions are illustrated.

Keywords: Indoor Localization, Machine Learning, Fingerprinting, RSSI, Stable AP, Granularity

Acknowledgements

I take this opportunity to express my deepest gratitude and appreciation to all those people whose guidance and encouragement have helped me towards the successful completion of this thesis.

I would like to express my sincere gratitude to my respected guide Dr. Chandreyee Chowdhury, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, for her unfailing guidance, prolific encouragement, constructive suggestions and continuous involvement during each and every phase of this research work. I feel deeply honored that I got the opportunity to work under her guidance.

I would like to express my sincere thanks to Prof. Samiran Chattopadhyay, Professor, Department of Information Technology, Jadavpur University, for his guidance and courage.

I would like to express my heartfelt gratitude to Mrs. Priya Roy, Research Fellow, Jadavpur University, Kolkata, for her suggestions and unwavering support.

I would also wish to thank Dr. Bhaskar Sardar, Head of the Department of Information Technology, Jadavpur University, Prof. Mahantapas Kundu, Head of the Department of Computer Science and Engineering, Jadavpur University, and Prof. Chiranjib Bhattacharjee, Dean, Faculty of Engineering and Technology, Jadavpur University for providing me all the facilities and for their support to the activities of this research.

During the last one year I had the pleasure to work in our laboratories. I am grateful to all the members of this laboratory for their kind co-operation and help.

I would like to express my gratitude and indebtedness to my parents and all my family members for their unbreakable belief, constant encouragement, moral support and guidance.

Last, but not the least, I would like to thank all my classmates of Master of Engineering in Software Engineering batch of 2017-2019, for their co-operation and support. Their wealth of experience has been a source of strength for me throughout the duration of my work.

Signed:

Mausam Kundu
Examination Roll No.: M4SWE19019
University Registration No.: 140974 of 2017-18
Department of Information Technology
Jadavpur University

Contents

Declaration of Originality and Compliance of Academic Ethics **iii**

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**
1.1 What is Indoor Localization System 2
1.2 Applications of Indoor Localization System 3
1.3 Overall Framework 4
1.4 Motivation 5
1.5 Contribution 6
1.6 Organization of the thesis 7

2 Related Work **8**
2.1 Overview 9
2.2 Machine learning techniques 9
2.2.1 Supervised learning 9
2.2.2 Un-supervised learning 10
2.2.3 Semi-supervised learning 12
2.3 Wi-Fi based Indoor localization 13
2.4 Fusion of Wi-Fi with other technologies 14
2.5 Summary 15

3 Varying Granularity **16**
3.1 Overview 17
3.2 Challenge 17
3.2.1 Grid-cell size selection 17
3.2.2 Important APs selection 18
3.3 Mechanism 19
3.3.1 Summary 21

4 Handling Device Heterogeneity **22**
4.1 Overview 23

4.2 Challenge 23
4.3 Conditional Ensemble 23
4.4 Mechanism 25
4.5 Summary 26

5 Experimental Setup 27
5.1 Overview 28
5.2 Data collection 28
5.3 Data pre-processing 29
5.4 Data analysis 30
5.5 Architecture 31
5.6 Summary 32

6 Result and Discussion 33
6.1 Overview 34
6.2 Result of Granularity and Stable AP selection 34
6.2.1 Cross-Validation 35
6.2.2 Unknown Test Set 35
6.2.3 Stable APs selection 36
6.3 Result of Device heterogeneity 37
6.4 Summary 39

7 Conclusion and Future Work 40

Bibliography 42

List of Figures

1.1 Framework of ILS 5
1.2 Change in signal strength due to different surrounding . . . 5
1.3 Change in signal strength due to different device 6
1.4 Change in signal strength due to different time of data collection 6

3.1 Proposed framework dealing change of context and grid sizes 18

4.1 Ensemble dealing device heterogeneity 24

5.1 Experimental region 28
5.2 Collected raw data 29
5.3 Filling the missing entries 30
5.4 Removal of unstable AP 30
5.5 Execution of kNN algorithm in WEKA 30
5.6 Execution of kNN algorithm in WEKA 31
5.7 Process flow 31

6.1 Different grid sizes in experimental region 34
6.2 Stable APs when grid size considered is 1x1 sq.m. 36
6.3 Average error distance (in meter) for different individual classifiers and Majority voting 38

List of Tables

6.1 Result of 10 fold cross-validation (%) 35
6.2 Result of unknown test set (%) 35
6.3 Result of 10 fold cross validation (% ± Standard deviation) . 37
6.4 Result of unknown test set (%) 37

List of Abbreviations

AOA	**A**ngle **O**f **A**rrival
AP	**A**ccess **P**oint
ARFF	**A**ttribute-**R**elation **F**ile **F**ormat
BayesNet	**B**ayesian **N**etwork
BLE	**B**luetooth **L**ow **E**nergy
BSSID	**B**asic **S**ervice **S**et **Id**entifier
ELM	**E**xtreme **L**earning **M**achine
FT	**F**unction **T**rees
GPS	**G**lobal **P**ositioning **S**ystem
GSM	**G**lobal **S**ystem for **M**obile
ILS	**I**ndoor **L**ocalization **S**ystem
J48	**J**ava **C**4.8
JSON	**J**avaScript **O**bject **N**otation
kNN	**k**-**N**earest-**N**eighbour
RFID	**R**adio **F**requency **Id**entification
RSS	**R**eceived **S**ignal **S**trength
RSSI	**R**eceived **S**ignal **S**trength **I**ndication
SSID	**S**ervice **S**et **Id**entifier
SVM	**S**upport **V**ector **M**achine
TDOA	**T**ime **D**ifference **O**f **A**rrival
TOA	**T**ime **O**f **A**rrival
ToF	**T**ime **o**f **F**light
UHF	**U**ltra **H**igh **F**requency
UWB	**U**ltra-**W**ide**B**and
VLC	**V**isible **L**ight **C**ommunication
WEKA	**W**aikato **E**nvironment for **K**nowledge **A**nalysis
Wi-Fi	**Wi**reless **Fi**delity
WLAN	**W**ireless **L**ocal **A**rea **N**etwork

List of Symbols

C	Set of grid-cells i.e C = $\{c_1$,c_2,...,$c_n\}$
R	Preprocessed RSS from AP
AP	Set of all APs
AP_{stable}	Set of stable APs
cs	cell size having maximum E_{acc} and minimum E_{meter}
T_r	Train set
T_e	Test set
E_{acc}	Localization accuracy
E_{meter}	Localization error in meter
μ	Mean
σ	Standard deviation
CF	Set of classifiers i.e CF = $\{cf_1$,cf_2,...,$cf_n\}$
CN	Set of conditions i.e CN = $\{cn_1$,cn_2,...,$cn_m\}$
cf_i	i^{th} classifier in CF
cf	An Individual classifier
cn_j	j^{th} condition in CN
AC	Accuracy of all classifiers in CF i.e AC = $\{ac_1$,ac_2,...,$ac_n\}$
l_{pred}	Predicted location
l_{act}	Actual location
AC^{mv}	Accuracy of majority voting in changing conditions

Chapter 1

Introduction

1.1 What is Indoor Localization System

Indoor Localization System (ILS) is a popular approach for providing location based services indoors. It is a topic of interest among researchers since it engages the readily available infrastructure for providing the services.

The need for such a system arose due to the poor performance of the Global Positioning System (GPS) in close vicinity such as shopping malls, parking areas. GPS is unable to predict location in indoor environments correctly as the signal from the satellites gets attenuated by roofs, walls of buildings. The error rate of GPS is sometimes larger than the indoor space itself.

Various technologies are available for ILS such as Visible Light Communication (VLC), Ultra Wide Band (UWB), Bluetooth Low Energy (BLE). Various research works [1] [2] [3] are done using these technologies. However, it is found that using these technologies may not be feasible as the infrastructure is not readily available and also not cost effective. Also, there are other factors like the range of BLE is too short, VLC needs light round the clock, high interference in UWB due to metallic and liquid materials and hence, there is a need of intensive research using these technologies to overcome its shortcomings.

Wireless Fidelity (Wi-Fi) is another technology using which ILS is implemented. The major advantage of using Wi-Fi is the availability of Wi-Fi APs in our homes, shopping malls, airports, and other indoor spaces. Thus, the infrastructure for implementing ILS is readily available to be used and also cost-effective. Also, Wi-Fi offers a fair range of connectivity and has no major disadvantage as such. This makes Wi-Fi the most popular choice among all the technologies available.

There are various proposed methods in indoor localization using Wi-Fi. However, no method can completely eliminate shortcomings such as multipath effect. There are two methods, triangulation and scene analysis. In the triangulation method, the basic properties of the triangle are applied to determine the position. The range and direction are both taken into account for prediction. Time of arrival (TOA) and time difference of arrival (TDOA) are range measurement techniques whereas angle of arrival (AOA) is a direction based measurement technique.

Scene analysis involves the RSSI from APs received during the test phase to be compared with the already stored RSSI in the training phase. There are two type of approaches in scene analysis, statistical approach and machine learning approach. The major works using statistical approach involves RADAR [4], HORUS [5]. High accuracy and low computational costs are achieved with the help of HORUS. However, it was effective only in a static environment i.e if there is any change in the environment of testing such as a change in the number of obstacles present, adding new devices results in a significant drop of accuracy. Whereas, machine learning approach uses the popular machine learning algorithms such as Bayesian classifiers [6], SVM [7], kNN [8], Neural Networks [9] etc. to determine the localization accuracy.

The scope of indoor localization is immense. Several researches are done by researchers in this field to improve upon the current disadvantages of the system. Applications based on this technology can provide a huge benefit to the users. Few examples include navigating in case of an emergency, location tracking and navigation in indoors, finding different locations/items in places having large indoor space as discussed in the next section.

1.2 Applications of Indoor Localization System

Wi-Fi access points are readily available in big cities, whether it's train stations, airports, offices, educational institutes, and even in residences. The advantages of GPS in outdoor conditions can be easily observed, whether it's navigation, finding a place globally or locally, even cab booking, all in our fingertips. The scope of GPS is immense, however ILS has the capability to provide same services as GPS in indoor. The indoor and outdoor environments greatly vary with each other. The few challenges of indoor localization are as following:

- Attenuation of signals due to walls, rooftops and other objects in indoor space
- Requires clear line of sight (LOS)
- Higher accuracy required as compared to GPS
- Robustness towards changing indoor conditions

A few applications of Indoor localization which makes an impact in our day to day lives are:

- *Indoor location tracking and navigation* : The preliminary job of indoor localization is to locate objects in indoors. The technology has been used in a shield tunnel construction site in Guangzhou, China [10]. The tracking of location of labors, machinery inside the tunnel is done. A Wi-Fi based positioning system along with active RFID tags and APs are used. The RFID tags are attached to workers and machinery, from there the collected signal and ID's of APs are sent to the outside server.

In hospitals, tracking of patients can be done [11]. Evacuation is possible by locating nearest exit points in case of an emergency.

Amazon [12] use the technology in their warehouses. The warehouses are quite big in area, hence the shifting of objects is done with the help of Kiva robots. It saves time as well as space.

- *Proximity marketing* : The potential of this technology is commercially huge. In shopping malls, not only it can be used to locate shops and navigate. But, it can be used for location based marketing. A shopkeeper can easily push notify the nearby people present in the mall about a new offer/discounts to attract the customers. Location based advertising for different products can also be done. An application developed by a company named Intu [13] allows a retailer having low sales on a particular day to notify the application users walking by the store. Thus, it will change the retail environment and will help small retailers to a great extent.
- *Rescue operations* : People can be benefited with the help of ILS. In case of emergencies, such as fire outbreak or terrorists attack, the people trapped inside the building can be alerted and navigated towards the nearest exit point. Also, the personnel such as firefighters, police can be tracked during evacuation and can be alerted on a timely basis from outside the building.

1.3 Overall Framework

The framework of ILS involves the collection of data, followed by preprocessing of the raw data. The important features of the collected data are identified and machine learning algorithms are applied for classification of the data. Finally, location can be predicted with the help of classified data. The framework is discussed in details with the help of Figure 1.1.

1. *Data collection* : In this phase, data is collected with the help of devices across the whole area of experiment. The data collected during various time of a day due to change in surroundings in indoor space for a whole day. Various different types of devices having different configuration are used in this phase.
2. *Data preprocessing* : This phase mainly removes the noise in data collected in the previous phase. The data is collected during various times of a day over a few days. So, signals may not be received properly from some of the APs. So, machine learning models are bound to fail if unprocessed data are fed as input.

3. *Feature extraction* : After cleaning of the data, it is important to identify the features of the data correctly. The feature extraction ensures that the essence of data is retained, but at the same time those attributes having low contribution is removed. Thus, the data is reduced yet the characteristics of the data are retained.

4. *Learning and classification* : In this phase, various machine learning algorithms are used for training of the model based on the extracted features in feature extraction phase. A new unknown instance is classified with the help of the trained model.

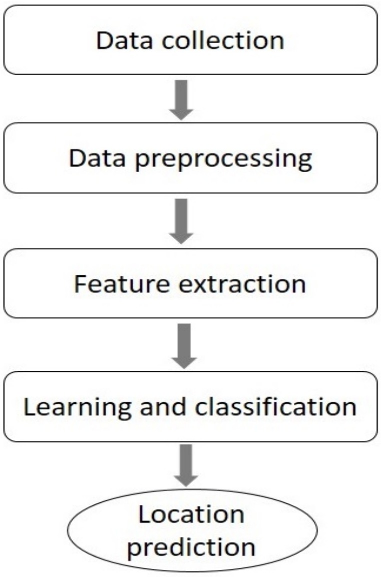
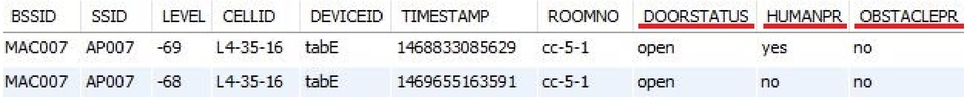


FIGURE 1.1: Framework of ILS

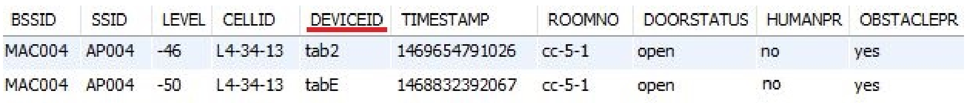
1.4 Motivation

ILS deals with a dynamic indoor environment during prediction. The objects, human keep on moving from one place to another. Depending on that the RSSI data received from APs also changes its characteristics. So, various perspectives of the collected data such as context, device, temporal are shown in the following figures:



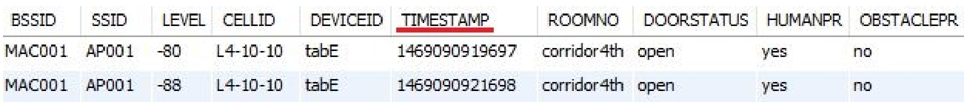
BSSID	SSID	LEVEL	CELLID	DEVICEID	TIMESTAMP	ROOMNO	DOORSTATUS	HUMANPR	OBSTACLEPR
MAC007	AP007	-69	L4-35-16	tabE	1468833085629	cc-5-1	open	yes	no
MAC007	AP007	-68	L4-35-16	tabE	1469655163591	cc-5-1	open	no	no

FIGURE 1.2: Change in signal strength due to different surrounding



BSSID	SSID	LEVEL	CELLID	DEVICEID	TIMESTAMP	ROOMNO	DOORSTATUS	HUMANPR	OBSTACLEPR
MAC004	AP004	-46	L4-34-13	tab2	1469654791026	cc-5-1	open	no	yes
MAC004	AP004	-50	L4-34-13	tabE	1468832392067	cc-5-1	open	no	yes

FIGURE 1.3: Change in signal strength due to different device



BSSID	SSID	LEVEL	CELLID	DEVICEID	TIMESTAMP	ROOMNO	DOORSTATUS	HUMANPR	OBSTACLEPR
MAC001	AP001	-80	L4-10-10	tabE	1469090919697	corridor4th	open	yes	no
MAC001	AP001	-88	L4-10-10	tabE	1469090921698	corridor4th	open	yes	no

FIGURE 1.4: Change in signal strength due to different time of data collection

In figure 1.2, the change in surroundings such as, whether doors are open or closed, human is present or not, any type of obstacle is present or not affect the signal received from the same AP to the same device. Collection of data using different devices under the same condition also affect the strength of signal from APs due to different configuration of different devices as shown in figure 1.3. Figure 1.4 illustrates the collection of data during different times of a day that also results in receiving different signal strength.

The above mentioned perspectives affect the localization performance and hence needs to be addressed so as to obtain a predictable localization accuracy.

1.5 Contribution

Determining location in indoor is a challenging task due to the dynamic nature of the environment. So, to address this change, data is collected keeping in mind the context, temporal and device heterogeneity.

The capability of ILS decreases due to a change in device as it is not possible to train a system with all the devices available. This degradation is due to changing signal strengths obtained from devices having a different configuration. So, to address this the individual classifiers are trained in various conditions and tested in an unknown condition. We proposed an ensemble of condition based classifiers which provide result based on the classification results obtained by individual classifiers in different conditions.

We have also discussed how important is the selection of optimal grid size and identifying stable APs in that particular granular level. We have shown how the increase in size of the grid cells will affect the overall classification accuracy of the system. We proposed an algorithm to find the optimal grid size. We also tried to find stable APs for each granular level. Selecting only the stable APs rather than considering all APs tend to give better results.Thus, identification of such APs is very important as considering the less contributing APs will only decrease the overall performance of the system. We tried to find a combination of grid size and stable APs for which the performance of the system is best.

1.6 Organization of the thesis

The rest of the thesis is organized as follows:

Chapter 2 contains few research works based on the different types of machine learning techniques. Further, few works are discussed based on Wi-Fi and Wi-Fi fused with other technologies in Indoor Localization.

Chapter 3 investigates the importance of granularity. The effect of change in sizes of grids has also been elaborately discussed. Also, the importance of stable AP selection based on granularity has been illustrated.

Chapter 4 contrasts on the effect of device heterogeneity on ILS and how this problem area can be addressed has also been talked about in this section.

Chapter 5 presents the whole experimental setup for carrying out the experiments. Storing of data, followed by preprocessing and classification of data are also explained.

Chapter 6 discusses the results obtained from the algorithms proposed in chapter 3 and 4.

Chapter 7 finally draws the conclusion of the thesis and discusses the limitations and scopes of future work.

Chapter 2

Related Work

2.1 Overview

This chapter discusses the representation of the research works on ILS. The research works can be broadly divided into two categories based on the approach used i.e statistical approach and machine learning. However, nowadays the current trend is towards machine learning approaches because of it's learning capability and robustness towards variation in features. Machine learning techniques can also be divided into three sub-categories, supervised learning, unsupervised learning, semi-supervised learning.

2.2 Machine learning techniques

2.2.1 Supervised learning

In this learning, all the instances of data obtained during the data collection phase are labeled properly. Regression and classification approach is generally used in supervised learning. Some of the supervised learning algorithms used in ILS are as following:

1. *Bayesian Network* : Bayesian network is based on Bayes theorem and is used for classification. Bayesian network is formed by the calculation of conditional probability at each node. The topological structure of the Bayesian network and conditional probability table (CPT) are two main parts of Bayesian network [14]. The Bayesian network can be represented by a directed acyclic graph. Each node of the acyclic graph represents the attribute of data whereas CPT of each node enumerates the corresponding probability for each value of the node.

2. *Support Vector Machines (SVM)* : SVM algorithm is applied in both classification and regression problems. It categorizes new data coming into the system by creating an optimal hyperplane. Centroid algorithm is used for localization after the creation of a hyperplane separating training data points which are nearest to the test data point.

3. *k-nearest neighbour (kNN)* : kNN [15] is a lazy learning technique that classifies instances based on their similarity. It is called lazy because approximation of function happens locally and during the classification phase computations are done. A new data point is classified based on the classes of the nearest neighbors. For e.g., if the value of k is 3, it means the class which is common among the closest 3 neighbors of the data point will be assigned as the class of the data point.

4. *K** : K* algorithm is an instance based algorithm, which defines the distance metric by using entropy concept. The entropy is calculated by finding mean of transformation complexity from one instance to another. The probabilities of a new instance to all the members of the category are summed up. The highest probability is finally selected [16].

Research works in the field of ILS using the above mentioned supervised machine learning techniques are discussed as below:

Martin Azizyan et al. [17] proposed a classification technique based on ambiance fingerprinting with Wi-Fi triangulation. In ambiance fingerprinting, data of Wifi, light, sound, compass, and accelerometer are stored. The model was trained using the mean and variance of the accelerometer readings as features. In the testing phase, classification based on accelerometer readings is done using SVM to classify whether the user is in moving state(+1) or stationary state (-1). However, the application is not feasible in real life as it is time-consuming and power hungry in nature.

An environment learning approach was proposed by JM Akre et al. [18] for the localization of UHF passive tags. An aggregate function of all RSSI values from possible RFID readers is calculated during the training phase. kNN algorithm is used to determine the class of test data. The k nearest neighbors of the test data tag are found, among which the dominant tag is determined as the tag corresponding to test data.

Luca Calderoni et al. [11] proposed an indoor localization system in a hospital environment using Random Forest classifier to predict room level accuracy. Patients can be located with the help of RFID tags and antennas. In the training phase, a bootstrap sample with n times replacement is chosen from a total number of available samples. Error of the tree is estimated with the remaining samples. In the test phase, a new observation is used as input to all the trees in the forest. The room of the device is determined by the prediction of the majority of trees in the forest.

2.2.2 Un-supervised learning

This learning technique involves unlabeled training data. Clustering approach is used in case un-supervised learning, to get the essence of the data by clubbing those instances of data having the same characteristics. One of the most common unsupervised learning algorithm used in ILS is mentioned as following:

1. *k-means* : k-means algorithm use clustering technique to categorize un-labeled data. The value of k is determined by the number of initial centroids representing each cluster. The distance between the data points and the centroids are calculated following which the data points are assigned the cluster corresponding to the centroid having least distance. The centroid is recomputed until there is no change.

A few research works using un-supervised learning are described as following:

Philipp Bolliger [19] proposed Redpin which eliminates pre-deployment effort of fingerprint generation. Room level accuracy is provided by measuring the strength of currently active GSM cell, signal strength from all available APs, Bluetooth signals from non-portable Bluetooth devices. The position is identified by comparing the current measurement with all the known fingerprints stored in the database. The location corresponding to the fingerprint whose distance to the current measurement is smaller than the threshold is known as the location of the device. However, if an unknown or wrong location is detected, users are prompted to correct the fingerprint. Hence, the location database is not updated without the explicit input from the users which may not be feasible in real life implementation.

H Wang et al. [20] proposed an unsupervised learning scheme, named UnLoc. UnLoc eliminates the need for traditional fingerprinting. The naturally existing landmarks in indoor space such as a unique pattern of accelerometer, a unique set of Wi-Fi APs from a particular point, unusual magnetic fluctuation from a particular point are used to identify specific locations. Using Unloc, mobile devices can sense these 'landmarks' in indoor. Recalibration of the mobile devices can be done with on the basis of sensing the 'landmark'. k-means clustering is used for clustering the signatures. The value of k is varied to identify the clusters which are highly dissimilar than other clusters to create good signatures. However, the signatures of specific locations need to be modeled as the method is based on the information of those specific locations. Hence, only after modeling and identification of the landmarks, creation of unsupervised Wi-Fi fingerprint is possible.

Valentin Radu and Mahesh K. Marina proposed HiMLoc [21] which is able to predict the location in indoor from the crowdsourced data. Pedestrian dead reckoning (PDR), Wi-Fi fingerprinting, and activity recognition data are combined together to give the estimation. The database contained Wi-Fi fingerprint along with corresponding location annotation. For classification of the activity recognition data, three classifiers J48, NaiveBayes and FT are used in WEKA tool. The obtained signal is compared using Euclidean distance with the Wi-Fi data already stored in database. The database keeps on updating as the fingerprints are annotated with time, and the new fingerprint gets higher priority and replaces old ones. The estimations of activity classifier, Wi-Fi positioning component and PDR's variable are integrated using particle filter. Weights are assigned to each of the predictors which get updated over time based on their performances. However, the weight of all of the predictors is normalized to keep the summation 1. However, usage of the proposed method is not battery friendly, hence turning on sensors depending on the available landmarks at a particular point can be a remedy to it.

2.2.3 Semi-supervised learning

Semi-supervised learning uses a mix of both labeled and unlabeled data. Labeling data is a tedious job, so to avoid this only a few data is labeled whereas rest of the data is unlabeled. The possible classes can be identified from the labeled data, and hence, the classes of unlabeled data can be easily identified. A semi-supervised learning approach has been described below:

1. *Manifold Learning* : Manifold learning reduces the number of features from a high dimensional data set in order to find the defining features of the dataset. This makes a high dimensional data manageable without losing the essence of the dataset. The data points may have a lot of features, however only a few underlying parameters can describe these features. Data points embedded in high dimensional space are actually samples from a low dimensional manifold. Manifold learning tries to find these features of the data to find a reduced dimensionality of the data.

Few works in indoor localization using the semi-supervised approach are described as following:

The authors [22] proposed semi-supervised learning based on label propagation algorithm (LP algorithm). In the training phase, few labeled data with a huge amount of unlabeled data is used to obtain a completely labeled dataset. In the testing phase, unlabeled new signal strength data is labeled to estimate the current location. Labeled and unlabeled data are represented as vertices of a connected graph. LP algorithm is used to propagate the label from a labeled vertex to another unlabeled vertex through the weighted edges. The process of propagation converges after inferring labels of unlabeled vertex. The data points are divided into K parts, out of which only 1 part is labeled and rest K-1 parts are unlabeled. During construction of connected graph, the edge is constructed between two data points i, j if i is present among the j's 20 nearest neighbors.

Teemu Pulkkinen et al. [23] proposed a manifold learning based WLAN positioning approach. The dimensionality of dataset is reduced into more manageable one to find the important defining features of the high dimensional dataset. The precise location of the previously obtained small subset of fingerprints is used to map points on Isomap manifold to geographical coordinate system.

A semi-supervised hybrid fingerprint method is proposed by Mu Zhou et al. [24] to reduce calibration effort needed for traditional fingerprinting. The database consists of a small number of known fingerprints along with the locations and a comparatively higher number of unlabeled fingerprints with no location attached to them. The irregularly collected user traces are labeled using semi-supervised manifold alignment approach. A reduced number of RSS samples are collected at each reference point/location by applying Execution Characteristic Function (ECF). The ECF populated the radio map by interpolating the new fingerprint from its neighbors. Localization performance is estimated using Bray-Curtis distance and A-star algorithm based Manifold alignment radio map along with four other types of radio maps.

2.3 Wi-Fi based Indoor localization

One of the most popular technologies used in Indoor localization is Wi-Fi. Some of the research works in this field using Wi-Fi are as following:

Lim et al. [25] proposed a system that does not require offline RSSI fingerprinting. WiFi APs are used as reference node as their location is known apriori. An online RSSI map is created when an AP obtains RSSI values from other APs. The user location is estimated by mapping the RSSI to distance between client and APs. A median accuracy of 1.76 meters is obtained. However, an user location is predicted by estimating the number of samples which might cause a delay.

Vasisht et al. [26] proposed a system based on single Wi-Fi AP known as Chronos. ToF is used by Chronus for accurate localization. ToF is calculated by the certain beacon messages received at the AP from the user device. A wideband system is emulated by employing the inverse relationship between bandwidth and time. Different channel measurements are obtained from the hopping of transmitter and receiver between different frequency bands of Wi-Fi. Accurate ToF is estimated from the combined information obtained. After the computation of accurate TOF at the AP, distances between each antenna pair on AP and user devices is obtained. An error minimization process is used to obtain 2D locations relative to the AP with the help of the measured distances. A median accuracy of 0.65 meters is attained by Chronos. However, the approach may not be feasible in real life as it consumes a lot of energy and not scalable.

2.4 Fusion of Wi-Fi with other technologies

Wi-Fi based Indoor localization can further be improved if it can be fused with other technologies. However, Bluetooth is found to be the most appropriate technology to be fused with Wi-Fi as both are having common characteristics. In fusion of technologies, one technology is considered to be relevant for estimating user location whereas another technology is used to complement and enhance the features of the system. Some of the research works using Wi-Fi along with Bluetooth is discussed as follows:

Matthew Copper et al. [27] proposed a new framework named LoCo which provides high accuracy in indoor conditions. LoCo combines the RSSI signals from Wi-Fi and signals of Bluetooth low energy (BLE). LoCo framework collects signal strength from Wi-Fi APs and BLE beacons with the help of a client service running on a smartphone. A classification engine is used to process the scanned data. Processing of data can be done in two ways: one way of processing allows classification to be performed in the device only, as classifier along with trained definition is loaded into the device beforehand. The second way is that the classification is done in cloud-based server and the scan results are send to the cloud server as a JSON structure. The device's room is estimated using ensemble learning method boosting [28][29] and high-level accuracy is provided by the use of straightforward methods such as [19]. Three methods named 'MAX BLE', 'Redpin System' and boosting classification illustrated the increase in accuracy when both BLE and Wi-Fi RSSI data are considered instead of only BLE or only Wi-Fi RSSI data. The improved accuracy in case of using Wi-Fi and BLE together is 0.966 as compared to the individual accuracy of 0.937 and 0.943 in BLE and WiFi.

Another research by Ju-Hyeon Seong et al. [30] on an environment adaptive localization method by using Wi-Fi and BLE. The authors scanned areas where BLE and Wi-Fi are installed. The WiFi and BLE signals are recorded with the help of smartphones. In training phase, they preprocessed all the BLE and Wi-Fi signals collected with Hausdorff distance algorithm and removed the outliers and eliminated abnormally recorded signal strengths from various APs. The authors proposed a log-distance path loss model which is able to predict position more precisely than the existing fingerprint method. The experimental results showed a reduction in existing method average error of 2.758m to 2.063m.

2.5 Summary

The chapter gives a brief idea of the machine learning techniques used in Indoor Localization. Various research works based on those machine learning techniques are illustrated in detail. Further, a few research works using Wi-Fi and fusion technologies such as BLE are discussed in the later part. In the next chapter, a proposed work dealing with granularity has been discussed in a detailed manner.

Chapter 3

Varying Granularity

3.1 Overview

Grid size selection becomes an important parameter while predicting location with high precision. Indoor activities like navigation require very high precision. Also, the need for identifying the important APs with respect to different grids arise as the APs change while navigating from one grid to another.

This chapter focuses on the importance of varying positional granularity and AP selection while determining the location indoors.

3.2 Challenge

The WiFi signal strengths depend on certain factors such as configuration of device, time and surrounding conditions. Signal strengths vary when there is a change in the above mentioned factors. Along with these factors, variation in the area covered by one coordinate location (that is, one grid) can be another important aspect behind changing WiFi signal strengths.

However, most of the research work in this field is done by fixing the granularity (grid size) for a given context. Different granularity along with context might not play a crucial role while achieving room level accuracy, but can be of the utmost importance when positioning is even more precise. To address the changing behaviour of signal strengths due to granularity and context, there is a need for the selection of stable APs.

Localization error can be minimized by selecting the most stable APs for appropriate positioning granularity.

3.2.1 Grid-cell size selection

The research works in this field mainly focus on room level accuracy and fine grain level accuracy. Room level accuracy [27] helps us in determining whether a person is in a room or not. However, room level accuracy may not be appropriate considering the fact that nowadays, most of the location based services and applications require even more precision i.e fine grain accuracy. In fine grain accuracy, square grids of certain area (say 0.5m x 0.5m) is used to represent location points of the experimental region. As the grid size gets altered, the positioning capability of the system also changes. Hence, fine grain level accuracy at various granularity levels must be compared for different state-of-the-art classifiers. Results across these classifiers help in identifying the most appropriate positioning granularity.

3.2.2 Important APs selection

Estimation of incorrect location is possible at a particular time if the various factors mentioned above are not taken into account. It is therefore very much significant to identify those APs that are steady in nature, as a result of which the change in time has a negligible impact on the identified APs. Identification of stable APs also helps in reducing computational time and more accurate location prediction. Hence, the robust APs should be considered while performing experiments. The importance of stable AP selection is shown by Jiang et. al. [31]. However, the selection of these APs should not be concentrated in a particular region, as accuracy may get affected while considering other regions. So, with respect to each and every region, important APs needs to be identified.

The framework to address various context and different grid sizes is illustrated in Figure 3.1.

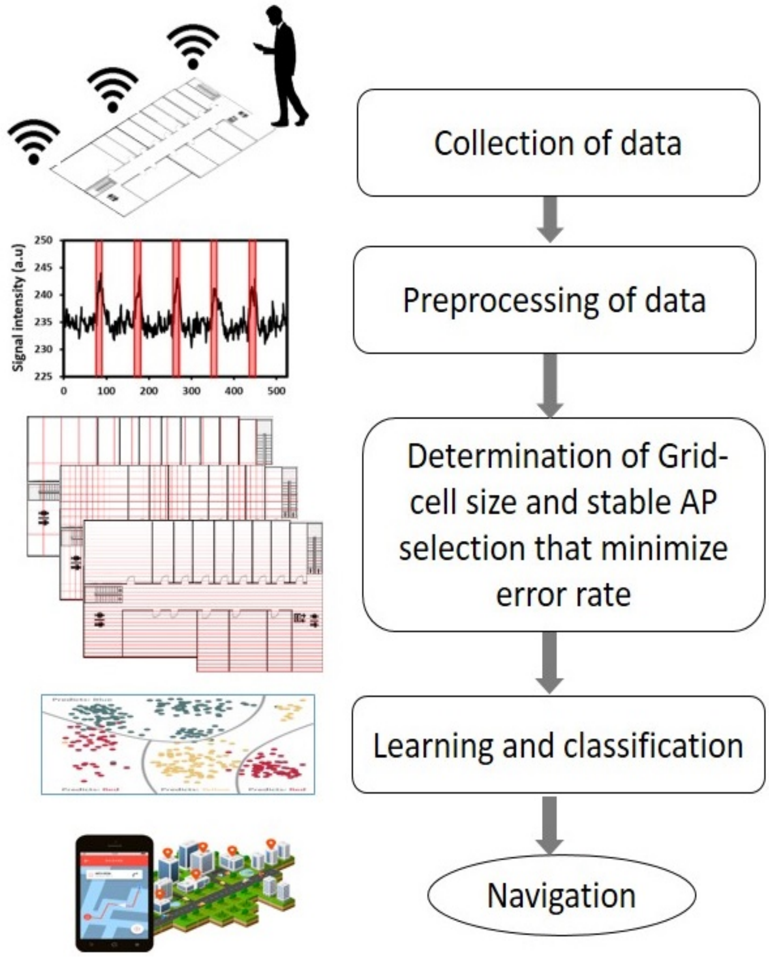

Collection of data
Preprocessing of data
Determination of Grid-cell size and stable AP selection that minimize error rate
Learning and classification
Navigation
Signal intensity (a.u)
250
245
240
235
230
225
0
100
200
300
400
500

FIGURE 3.1: Proposed framework dealing change of context and grid sizes

3.3 Mechanism

Minimization of localization error by the selection of grid-cell size and determination of stable APs has been discussed as follows:

1. *Grid-cell size selection*: The main aim of different grid-cell size selection is to understand which grid size reduces the prediction error. Algorithm 1 summarizes the grid selection method. The algorithm takes preprocessed RSS from the APs and the respective APs as input. In step 2 and 3, the entire experimental region is divided into ixi sq.m. cells followed by, assigning each cell a label, say $C^i = \{c^i_1, c^i_2, \cdots, c^i_n\}$. Then, with respect to each cell, the corresponding stable APs are found in step 4. In steps 5 and 6, Signal strength from all the stable APs is considered for making the training set T^i_r whereas, certain cells of C^i are considered for generating the test set T^i_e. State-of-the-art classifiers are used to determine localization accuracy E^i_{acc} and localization error E^i_{meter} by classifying T^i_r and T^i_e in steps 7 and 8. The grid-cell size in which the achieved localization accuracy is maximum and localization error is minimum is the output of the algorithm in step 9.

2. *Stable AP selection*: The different steps of selecting stable APs has been summarized in Algorithm 2. The algorithm takes a set of grid cells, say $\{c_1, c_2, \cdots, c_n\}$, preprocessed RSS from APs and a set of all APs as input. Collection of multiple samples has been done with respect to time (T_s), device (D_{id}), and ambient conditions (Am_{hu},Am_{hr}). Mean, μ and standard deviation, σ are calculated corresponding to each AP and each cell c^i_j in step 5 and 6. A weak signal has an RSSI level measuring less than -80dBm. These type of signals are generally not heard during experimentation. Hence in step 8, an AP a is added to set of stable AP only when mean of the RSS received at AP is greater than -80dBm and the standard deviation is less than a threshold.

   The mean is calculated using the formula:

   $$\mu = \frac{\sum_{i=1}^{s} R^{t,am,d}_{c,a}}{s} \tag{3.1}$$

   where, c represents the cell from where s number of RSSI instances are collected. And, median is calculated by the formula:

   $$\sigma = \sqrt{\frac{\sum_{i=1}^{s} (R^{t,am,d}_{c,a} - \mu)^2}{s}} \tag{3.2}$$

Algorithm 1: Determination of Grid-cell size

```
input : $C$: grid-cells set $\{c_1, c_2, \cdots, c_n\}$,
        $R$: preprocessed RSS from AP,
        $AP$: a set of all APs
output: $cs$ : cell size having maximum localization accuracy and minimum
        localization error
for $i = 1$ to $m$ do
    divide the experimental area into $i \times i$ sq.m. cells
    assign label to each cell, say $C^i = \{c^i_1, c^i_2, \cdots, c^i_n\}$
    goto Algorithm 2, identify the stable APs, $AP^i_{stable}$
    prepare train set $T^i_r$ from $R$ considering RSSIs from $AP^i_{stable}$
    prepare test set $T^i_e$ considering RSSIs from $AP^i_{stable}$
    determine localization accuracy $E^i_{acc}$ by classifying $T^i_r$ and $T^i_e$
    determine localization error in meter $E^i_{meter}$
return cell size, $cs$, for which $E^i_{acc}$ is maximum and $E^i_{meter}$ is minimum
```

Algorithm 2: Stable APs selection

```
input : $C^i$: a set of grid-cells $\{c^i_1, c^i_2, \cdots, c^i_n\}$,
        $R$: preprocessed RSS from AP,
        $AP$: a set of all APs
output: $AP^i_{stable}$ : a set of stable APs of $C^i$
initialize $AP^i_{stable}$ with $\phi$
foreach $c \in C^i$ do
    foreach $a \in AP$ do
        if $a \notin AP^i_{stable}$ then
            calculate mean $\mu$ of RSSIs received
            calculate standard deviation $\sigma$ of RSSIs received
            if $\mu > -80dBm$ and $\sigma < threshold$ then
                add $a$ in $AP^i_{stable}$
return $AP^i_{stable}$
```

3.3.1 Summary

This chapter summarizes a challenge of IPS, i.e optimal grid selection and the effect of it on the performance of the system. Moreover, the importance of stable APs with respect to each grid has been discussed in this chapter.

But, the robustness of the system due to the inclusion of a new device to it, on which the model has not been trained is still a challenge. This challenge of device heterogeneity is discussed in the next chapter.

Chapter 4

Handling Device Heterogeneity

4.1 Overview

In indoor positioning, multiple devices are used for the collection of data. However, it is not possible to collect data with all the devices available. So, RSSI data corresponding to all the devices is not available. When a new device comes into the system, having different vendor, configuration than the ones used during the training phase, it may affect the overall accuracy of the system.

In this chapter, first we discuss the effect of devices having different configurations on indoor localization. In the next section, we proposed an ensemble of conditional classifiers which can handle device heterogeneity.

4.2 Challenge

Data collection in ILS is mostly done with the help of RSS fingerprinting. Fingerprinting method is robust towards the uncertainties of the propagation environment. Location is determined with the help of a device, whose current received RSS values are compared to the stored values in the fingerprint database. In a real scenario, there cannot be any constraint on the model of the device or number of devices. As the devices are heterogeneous in terms of hardware and software, RSS varies by 25-30 dB between the devices [32]. An ILS may fail if a new user device is used to test the system, due to the reason that the RSS fingerprint contains the signal strengths of other different devices.

4.3 Conditional Ensemble

The individual state-of-the-art classifiers used under same conditions for training and testing achieves good accuracy. However, when an unknown test set is used, the individual classifiers fail to address the change in condition, as a result of which the accuracy of the classifiers dropped significantly. Hence, it is quite evident that the individual classifiers are unable to generalize while different devices are used.

This drastic fall of accuracy with changing conditions motivates us to use an ensemble based approach. The ensemble can be done either by combining the predictions made by all classifiers for a particular condition, also can be called as a classifier combination or by combining all the predictions made by a particular classifier under different changes in contexts, called as a conditional ensemble. So, to handle changing conditions i.e change of devices we proposed a conditional based ensemble. Majority voting among the predictions made by a particular classifier over various contexts will help us handle any unknown condition correctly.

The error predicting capability of an ensemble model gets increased if we choose the individual classifiers having low correlation among themselves. Combining the opinion of classifiers with different predicting abilities help in handling the diverse characteristics of data. For this reason, we have chosen Bayesian classifier, function based classifier, and two different lazy learning based classifiers.

Wi-Fi fingerprinting and inertial sensor readings are integrated in [9]. Three different machine learning techniques namely, k nearest neighbor (kNN), Linear Regression (LR), Non-Linear regression based Neural Network (NL-NN) is used so as to achieve better accuracy while predicting location. However, specific indoor path is considered during data collection which is not practically possible as the user may move using some other path. The hybrid approaches also can speed up the time required for computation as illustrated in [33]. Scalability problem of indoor localization was solved by the authors [34] using a novel joint clustering and ELM method.

The working of an ensemble model dealing device heterogeneity has been illustrated in Figure 4.1.

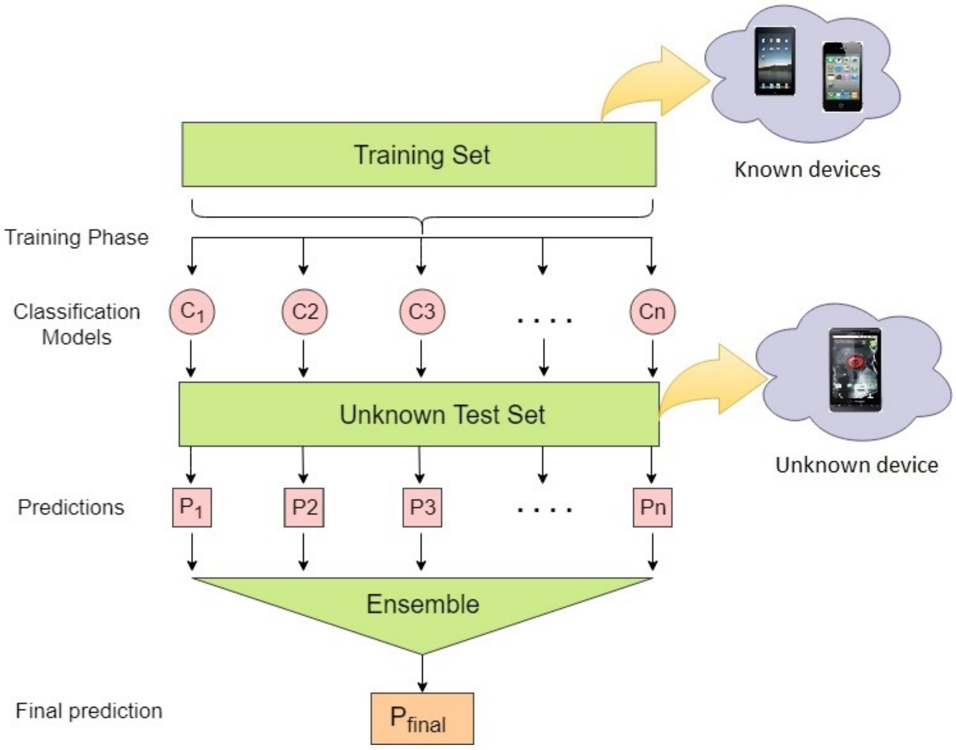


FIGURE 4.1: Ensemble dealing device heterogeneity

4.4 Mechanism

An algorithm based on classification with the unknown test set for an individual classifier and classification by majority voting among different classifiers has been discussed as follows:

Algorithm 3: Prediction with unknown test set

```
input : Set of Classifiers, $CF$= {$cf_1$,$cf_2$,....,$cf_n$}
        Train set, $T_r$,
        Test set, $T_e$
output: Accuracy, $AC$= {$ac_1$,$ac_2$,....,$ac_n$}, of all classifiers in $CF$
1 foreach $cf_i$ in $CF$ where i=1 to $n$ do
2     Output_Prediction ← classify $T_r$ and $T_e$ with $cf_i$
3     foreach instance in Output_Prediction do
4         find predicted location, $l_{pred}$
5         find actual location, $l_{act}$
6         if $l_{pred}$ = $l_{act}$ then
7             correct_instance ++
8     $ac_i$ ← (correct_instance / total number of instances)x100
9 return $AC$
```

Algorithm 4: Majority Voting

```
input : Set of Conditions, $CN$= {$cn_1$,$cn_2$,....,$cn_m$},
        An individual classifier, $cf$
        Test set, $T_e$
output: Accuracy, $AC^{mv}$, of majority voting in changing conditions
1 foreach $cn_j$ in $CN$ where j = 1 to $m$ do
2     prepare $T_r$ for $cn_j$
3     Output_Prediction ← classify $T_r$ and $T_e$ with $cf$
4 foreach instance of Output_Prediction do
5     find predicted location, $l_{pred}$, predicted in maximum number of $cn_j$
6     find actual location, $l_{act}$
7     if $l_{pred}$ = $l_{act}$ then
8         correct_instance ++
9 $AC^{mv}$ ← (correct_instance / total number of instances)x100
```

In algorithm 3, the individual classifiers are used for classifying the unknown test set. The output prediction file is generated in step 2 of the algorithm. Output prediction file contains all the predicted instances by the individual classifier used. In step 6, if the actual location is same as the predicted location then the number of correct prediction is increased by 1. Finally, in step 8 the accuracy is calculated by total obtained correctly predicted instances over the total number of instances.

In algorithm 4, after the classification of train set and unknown test set by an individual classifier under various conditions in step 2, output prediction files are generated in step 3. The predicted and actual location corresponding to all instances of prediction files are found in step 5 and 6 respectively. If the predicted location is found same as an actual location in step 7, the number of correctly classified instance is increased by 1 in step 8. Accuracy calculation is same as in algorithm 3.

4.5 Summary

In this chapter, first we identified the problem of device heterogeneity while predicting location in ILS and discussed the problem in a detailed manner. In the latter part of the chapter, we talked about using Majority Voting among classifiers which can resolve the issue by taking a collective decision so as to reduce the effect of device heterogeneity.

Chapter 5

Experimental Setup

5.1 Overview

This chapter discusses the implementation of our work in detail. First, the collection of data is described. In the next section preprocessing and preparation of the data is discussed in a detailed manner.

5.2 Data collection

An Android application was developed to record RSSI from corresponding to all the APs. The data collection covered all the cells of a particular floor 5.1 of our University building. The building consists of various classrooms, labs, faculty rooms having different dimensions. Most of the rooms were covered while collecting data. Data [35] were recorded during various times of day (temporal), using different devices (device heterogeneity) and different ambiances (closed or open door, change in no. of users present in the vicinity).

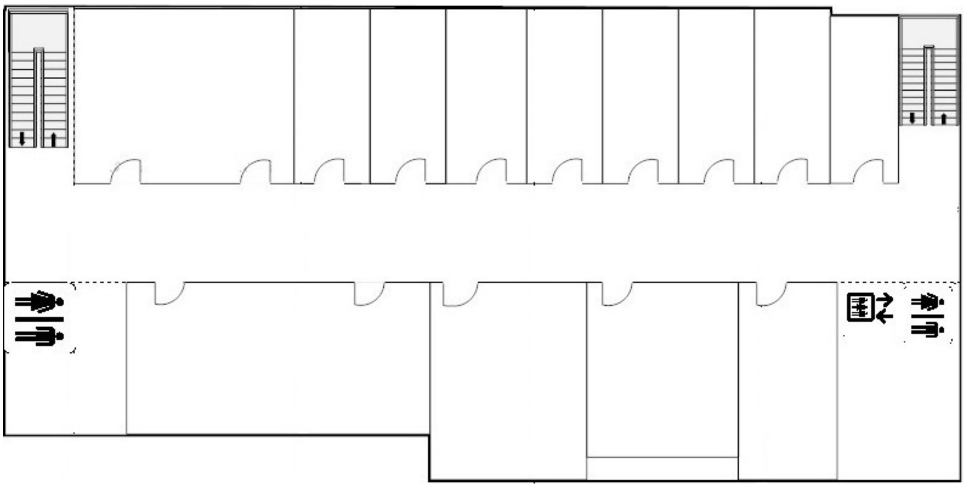
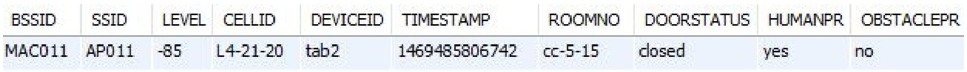


FIGURE 5.1: Experimental region

MySQL database was used to store the collected data because of it's lightweight and high performing nature. The database contains the following fields:



BSSID	SSID	LEVEL	CELLID	DEVICEID	TIMESTAMP	ROOMNO	DOORSTATUS	HUMANPR	OBSTACLEPR
MAC011	AP011	-85	L4-21-20	tab2	1469485806742	cc-5-15	closed	yes	no

BSSID is a unique identity no. assigned to the WAPs. There are 72 unique BSSIDs present in the database. SSID identifies the APs from which signals were received. There may be multiple APs having the same SSID. LEVEL represents the value of RSSI received.

CELLID uniquely identifies each cell of the experimental region. It represents a particular grid from where RSSI from the APs was collected during the data collection phase. The database is having 119 unique cells.

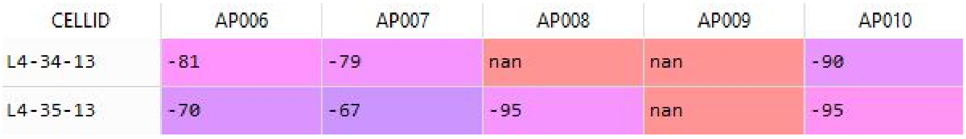
The data was collected with respect to change in context (DOORSTATUS, HUMANPR, OBSTACLEPR) from various rooms (ROOMNO) of the chosen floor, at a different time (TIMESTAMP) of the day. 5 devices (DEVICEID) having different vendors and different specifications were chosen while recording data.

5.3 Data pre-processing

Pre-processing of collected RSSI needs to be done before analyzing. During data collection, RSSI from all the APs is received which are in range. However, some APs are personal hotspots too, which are not present all the time during data collection. These type of AP hotspots are removed as they are not stable. The RSSI data from these unstable APs incur noise during data analysis and also, are not heard from all the locations leading to missing entries. Duplicate entries of fingerprint data are also removed because they don't add any new dimension to data analysis. Also, reduced data helps in much easier and less complex analysis.

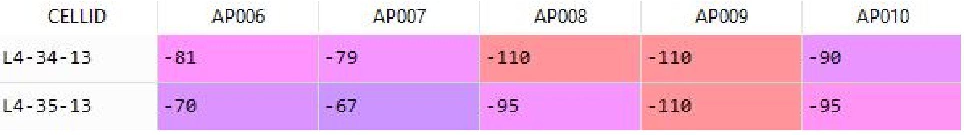
Due to the presence of obstacles in the indoors, it may happen that all RSSI values are not received from all APs and cells. As a result of which there will be missing entries in the collected data. These missing entries are filled with a dummy value of -110dBm. However, the range of RSSI received from the APs lie between -40dBm to -100dBm.

Pre-processing of collected data is illustrated with the help of following figures 5.2, 5.3 and 5.4:



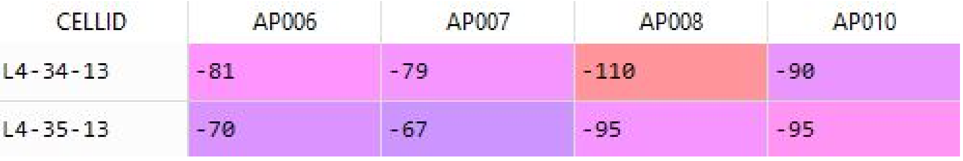
CELLID	AP006	AP007	AP008	AP009	AP010
L4-34-13	-81	-79	nan	nan	-90
L4-35-13	-70	-67	-95	nan	-95

FIGURE 5.2: Collected raw data



CELLID	AP006	AP007	AP008	AP009	AP010
L4-34-13	-81	-79	-110	-110	-90
L4-35-13	-70	-67	-95	-110	-95

FIGURE 5.3: Filling the missing entries

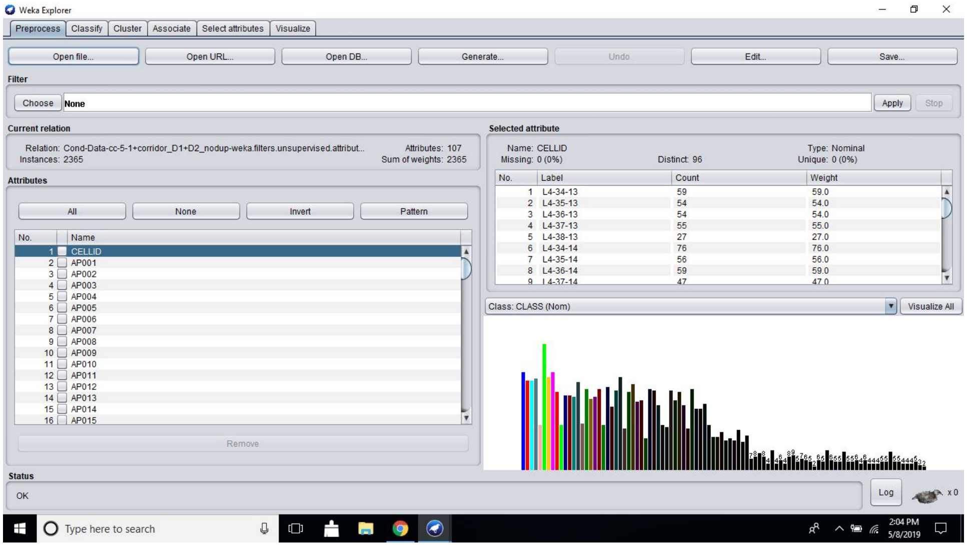


CELLID	AP006	AP007	AP008	AP010
L4-34-13	-81	-79	-110	-90
L4-35-13	-70	-67	-95	-95

FIGURE 5.4: Removal of unstable AP

5.4 Data analysis

The processed data is analyzed with the help of the Waikato Environment for Knowledge Analysis (WEKA). Machine learning algorithms available in WEKA are used for prediction. Execution of the machine learning algorithm kNN in WEKA environment is illustrated in 5.5 and 5.6.



Filter

Choose None

Current relation

Relation: Cond-Data-cc-5-1+corridor_D1+D2_nodup-weka.filters.unsupervised.attribut... Attributes: 107
Instances: 2365 Sum of weights: 2365

Attributes

All None Invert Pattern

No.		Name
1	□	CELLID
2	□	AP001
3	□	AP002
4	□	AP003
5	□	AP004
6	□	AP005
7	□	AP006
8	□	AP007
9	□	AP008
10	□	AP009
11	□	AP010
12	□	AP011
13	□	AP012
14	□	AP013
15	□	AP014
16	□	AP015

Selected attribute

Name: CELLID Type: Nominal
Missing: 0 (0%) Distinct: 96 Unique: 0 (0%)

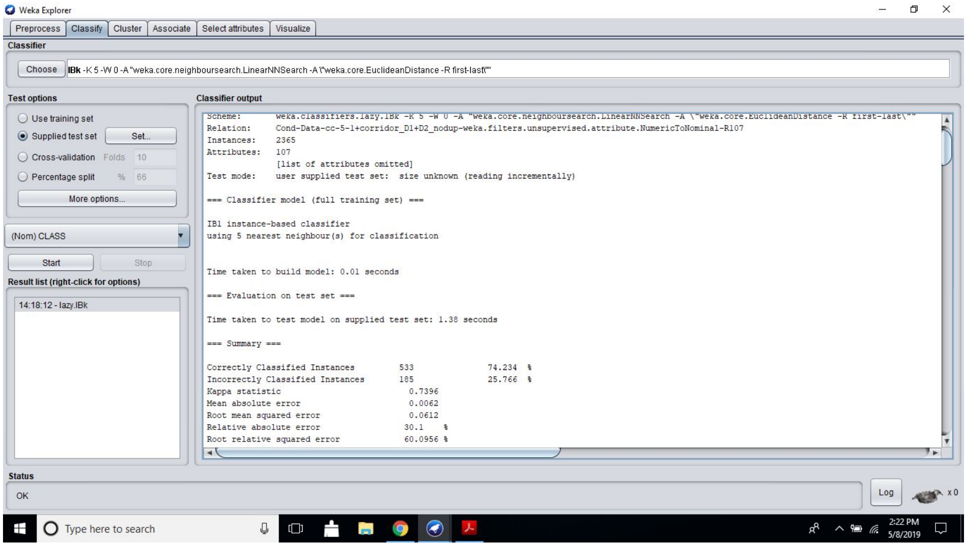
No.	Label	Count	Weight
1	L4-34-13	59	59.0
2	L4-35-13	54	54.0
3	L4-36-13	54	54.0
4	L4-37-13	55	55.0
5	L4-38-13	27	27.0
6	L4-34-14	76	76.0
7	L4-35-14	56	56.0
8	L4-36-14	59	59.0
9	L4-37-14	47	47.0

Class: CLASS (Nom)

Status

OK

FIGURE 5.5: Execution of kNN algorithm in WEKA



Classifier: IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) CLASS

14:18:12 - lazy.IBk

```
Scheme:       weka.classifiers.lazy.IBk -K 5 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     Cond-Data-cc-5-1+corridor_D1+D2_nodup-weka.filters.unsupervised.attribute.NumericToNominal-R107
Instances:    2365
Attributes:   107
              [list of attributes omitted]
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.38 seconds

=== Summary ===

Correctly Classified Instances         533               74.234  %
Incorrectly Classified Instances       185               25.766  %
Kappa statistic                          0.7396
Mean absolute error                      0.0062
Root mean squared error                  0.0612
Relative absolute error                 30.1     %
Root relative squared error             60.0956 %
```

Status: OK

FIGURE 5.6: Execution of kNN algorithm in WEKA

WEKA receives the pre-processed ARFF training set. A machine learning algorithm is chosen for prediction. Here, we have chosen the kNN algorithm (known as IBk in WEKA) for demonstration purpose. After the selection of algorithm, either we can perform cross validation, else we can supply an unknown test set for which accuracy will be predicted.

5.5 Architecture

The overall flow of data collection followed by data pre-processing and analysis is illustrated with the help of a process flow diagram in 5.7

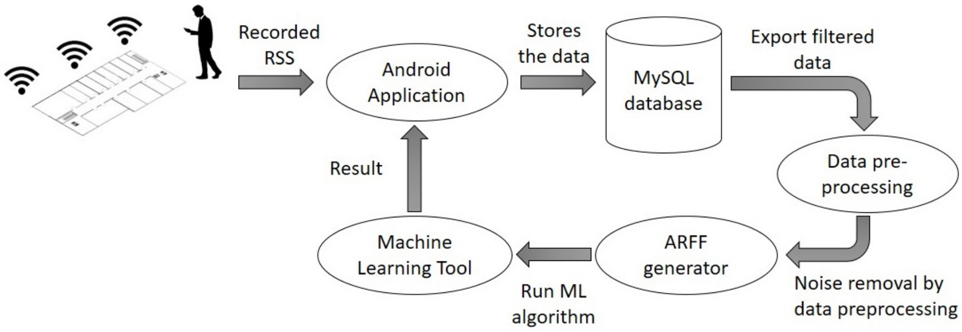


FIGURE 5.7: Process flow

The process starts from a collection of RSSI from various APs available with the help of an android application. The collected data gets stored in a MySQL database. The data is filtered, followed by pre-processing. In preprocessing, noise and other undesirable attributes are removed from the data and the data is generated in ARFF file format. WEKA is a machine learning tool which uses the ARFF file format. Different machine learning algorithms such as BayesNet, kNN, LibSVM, KStar are used to get the localization accuracy. Finally, the result obtained from the WEKA tool is returned back in the android application.

5.6 Summary

In this chapter, we first discussed the data collection in indoor positioning from various APs followed by the storing of the collected data in the database. We illustrated the pre-processing phase with the help of figures. Running of a machine learning algorithm in WEKA is discussed in a detailed manner. At last, we demonstrated the overall flow of the whole process in indoor localization.

Chapter 6

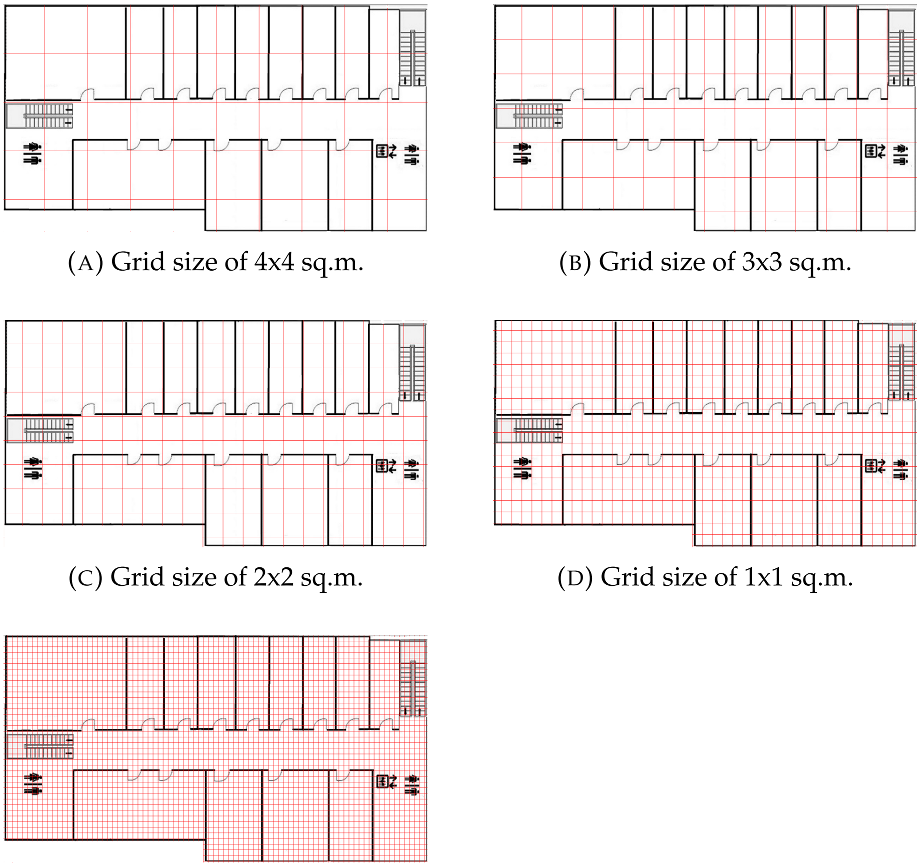
Result and Discussion

6.1 Overview

In this chapter, we will discuss the experimental results obtained from various granular levels. We will try to find the no. of stable APs, which varies along with a change in granularity.

6.2 Result of Granularity and Stable AP selection

Thorough experiments have been performed on different granularity levels as shown in figure 6.1. WEKA tool on Intel Pentium quad core machine with 1.60GHz processor and 8GB RAM is used while carrying out the experiments. Three different types of devices are used namely, Samsung Galaxy Tab 10 (*android version 4.0*), Samsung Galaxy Tab E (*android version 5.0*) and Motorola Moto E (*android version 5.1*).



(A) Grid size of 4x4 sq.m.

(B) Grid size of 3x3 sq.m.

(C) Grid size of 2x2 sq.m.

(D) Grid size of 1x1 sq.m.

(E) Grid size of 0.5x0.5 sq.m.

FIGURE 6.1: Different grid sizes in experimental region

Four state-of-the-art classifiers namely, BayesNet, LibSVM, IBk(k=5) and K* are used for classification. Java based code of two algorithms, Algorithm 1 and Algorithm 2 is interfaced with WEKA 3.9 to run the machine learning algorithms. The parameters of all the classifiers are default except IBk which is having 5 as the value of parameter k.

The results are obtained on cross validation and unknown test set. Unknown test results depict the robustness of the system when an unknown device comes into the experimental region.

6.2.1 Cross-Validation

Table 6.1 shows the result of 10 fold cross-validation result of the machine learning algorithms for each grid size that consider RSSIs as features. The training set consists of RSSI data collected with the help of two devices, Samsung Galaxy Tab 10 and Samsung Galaxy Tab E.

TABLE 6.1: Result of 10 fold cross-validation (%)

Size of cell (sq.m.)	**BayesNet**	**LibSVM**	**IBk**	**K***
4x4	98.71	93.08	95.16	94.70
3x3	94.69	86.65	87.02	91.95
2x2	88.70	87.32	89.40	89.40
1x1	95.46	85.79	88.40	91.32
0.5x0.5	30.92	17.10	50.45	19.07

6.2.2 Unknown Test Set

The result of experiment where the Test data set is supplied is shown in Table 6.2. The train set contains the RSSI data collected using Samsung Galaxy Tab 10 and Samsung Galaxy Tab E whereas in the test set contains the RSSI data collected using device Motorola Moto E.

TABLE 6.2: Result of unknown test set (%)

Size of cell (sq.m.)	**BayesNet**	**LibSVM**	**IBk**	**K***
4x4	96.43	92.90	95.16	94.71
3x3	94.69	89.35	89.35	92.71
2x2	91.91	87.18	90.33	91.13
1x1	95.46	85.81	88.76	91.32
0.5x0.5	38.88	11.11	33.33	22.22

Table 6.1 and 6.2 shows the results obtained by 10 fold cross-validation and unknown test set respectively. It is quite evident from the tables that localization accuracy tends to decrease as the size of grid becomes smaller. There is a drastic decline in accuracy when the cell size reduces to 0.5x0.5 sq.m. from 1x1 sq.m. RSSIs does not vary significantly when the size of the cells are so small, and hence even for the classifiers it is difficult to identify these cells. The cells are very small to even collect data if any obstacle is present, hence it is not feasible to collect data from all cells having an area of 0.5x0.5 sq.m.

The state-of-the-art classifiers can easily provide better accuracy in bigger grid sizes as it is easier to label each cell correctly as compared to smaller sized grids. Also, the presence of obstacles in bigger grids will not affect data collection from any grid. Wi-Fi fingerprint characteristics will be also varying significantly as the grid size increases.

6.2.3 Stable APs selection

Each and every state-of-the-art classifier used for classification achieves better accuracy when 22 stable APs are considered instead of all the 43 APs when the grid size is 4x4 sq.m. The localization accuracy varies from 92% to 96% when 22 APs are considered across all the classifiers.

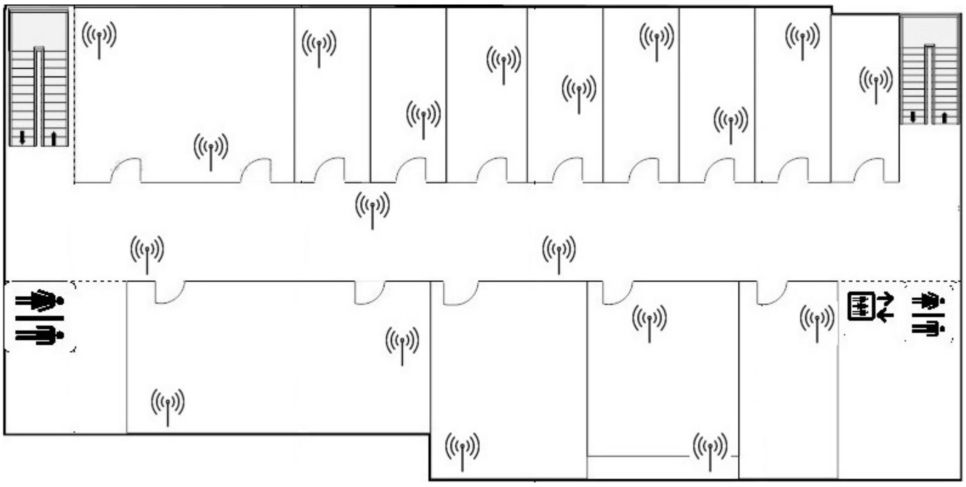


FIGURE 6.2: Stable APs when grid size considered is 1x1 sq.m.

As shown in figure 6.2, 20 stable APs are obtained from algorithm 2 when the experimental bed is divided into grid cells having 1x1 sq.m. size. This cell size achieves significant accuracy at a fine grain level.

6.3 Result of Device heterogeneity

Intensive experiments have been carried out using four devices D1, D2, D3, and D4. BayesNet, LibSVM, K* classifiers are used with default parameters and IBk is used with the value of k=5. Table 6.3 and 6.4 shows the results obtained after performing 10-fold cross validation and classification with unknown test set.

TABLE 6.3: Result of 10 fold cross validation (% ± Standard deviation)

Train Set	BayesNet	LibSVM	IBk	K*
D1,D2	94.25±1.13	95.41±1.12	92.23±1.59	97.39±0.85
D1,D3	89.66±2.01	92.69±1.59	85.50±2.30	94.30±1.39
D1,D4	92.82±1.57	93.08±1.34	92.35±1.64	95.85±1.20
D2,D3	94.28±1.44	96.02±1.12	90.92±1.79	97.52±0.98
D2,D4	92.78±1.76	94.61±1.49	91.91±1.62	97.35±1.11

The accuracy is shown by 10-fold cross validating with standard deviation is shown in table 6.3. Mean (equation 6.1) is calculated for accuracy obtained from each fold while cross validating. Further, the calculation of standard deviation (equation 6.2) using the following formula:

$$\mu = \frac{\sum_{i=1}^{10} acc_i}{10} \tag{6.1}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{10} (acc_i - \mu)^2}{10}} \tag{6.2}$$

where, *i* is the cross validation fold no. and *acc* is the accuracy achieved that particular fold.

TABLE 6.4: Result of unknown test set (%)

Classifier	D1,D2	D1,D3	D1,D4	D2,D3	D2,D4	D3,D4	MV
BayesNet	76.38	59.37	91.31	79.51	88.19	88.19	93.16
LibSVM	75.69	57.98	91.31	84.37	89.93	88.19	93.38
IBk	76.38	57.63	85.06	77.43	82.98	82.98	87.42
KStar	80.55	61.80	94.44	86.11	91.31	92.01	96.26

*MV = Majority Voting

Table 6.4 shows the result obtained after supplying an unknown test set to the system. Different training sets are used while performing the experiments. The training sets are a combination of data obtained from the mentioned devices. For e.g, D1,D2 means the training set contains RSSI values recorded in D1 and D2 device during the data collection phase.

Table 6.3 shows that all the classifiers achieve a fair accuracy under same conditions for training and testing. However, the accuracy drops under changing conditions. It is quite evident from table 6.4 that the performance of the system gets enhanced by the proposed method of majority voting. Classification with individual classifiers BayesNet, LibSVM, IBk, KStar obtained an average accuracy of 80.49%, 81.24%, 77.07%, 84.37% under different conditions. However, in Majority voting a significant increase of 10% to 13% is observed.

The reason behind the increase in accuracy is that even if the individual classifiers are failing to predict the data taken under different conditions, ensemble based approach is covering all the conditions and hence the accuracy of the overall system increases.

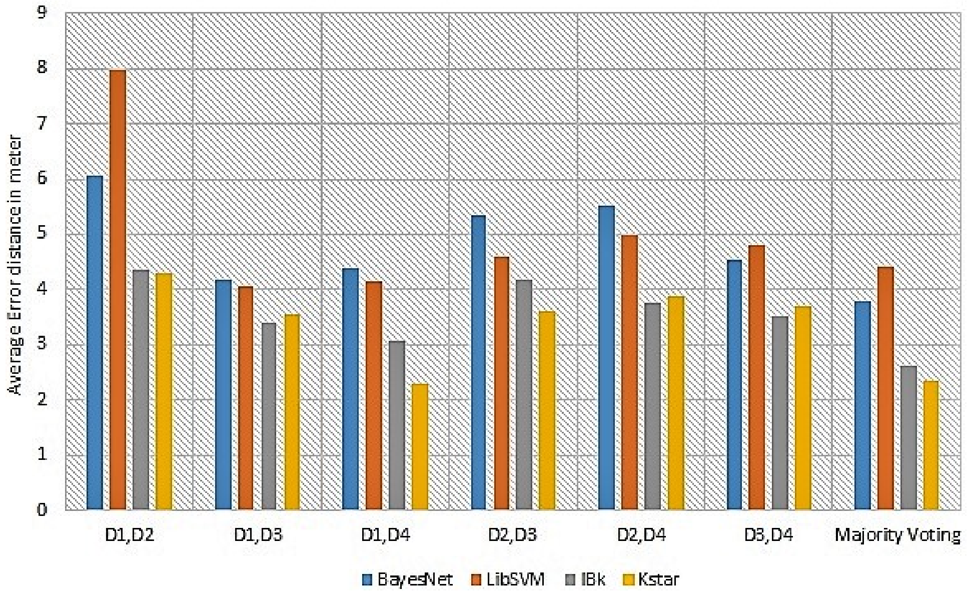


FIGURE 6.3: Average error distance (in meter) for different individual classifiers and Majority voting

The average error distance obtained across various classifiers is shown in figure 6.3. Majority voting obtained a relatively better result as compared to the individual classifiers. This signifies that Majority voting even when predicts a wrong location, it predicts a very nearby location to the original location due to which the error distance gets minimized.

6.4 Summary

This chapter gives an insight into how changing granularity can affect the capability of system. The dataset collected for different granular levels with respect to various heterogeneity results in varying signal strengths. Stable AP selection subject to heterogeneity and change in grid size can solve the issue.

The performance of individual classifiers while changing devices during data collection is also illustrated in this chapter. A brief discussion of results by the four machine learning algorithms with different parameter settings is provided. Among the individual classifiers K* performs the best however, overall results show that majority voting is able to handle the change in devices in a very robust manner and as a result, the accuracy of the system is also increased as compared to the accuracy obtained from individual classifiers.

Chapter 7

Conclusion and Future Work

Conclusion

Researches in the field of Indoor localization are still going on as there is no such proposed method which can address the changes of context and stabilize the effect of the changes.

The various heterogeneity such as context, temporal and device are discussed in detail. Further, the effect of this heterogeneity on ILS has also been illustrated. An ensemble based method is proposed which can handle the change in devices quite effectively. We have compared the results of classification accuracy obtained in individual classifier such as BayesNet, LibSVM, kNN and K* under changing conditions with the ensemble based approach.

We have also shown how granularity can affect the positional capability of ILS. The increase in size of the grids increases the accuracy however, certain activities in indoor space requires very fine grain level accuracy and hence, the grid size should be optimally set. Important APs in an ILS needs to be identified to reduce the re-calibration effort. The stable APs change as there is a change in granularity. So, it is important to identify APs with respect to each and every grid size and then an optimal grid size must be obtained which provides significant accuracy at a fine grain level.

We have supported the observations above with comparative classification results obtained across various state-of-the-art classifiers.

Future Work

The conditions in an indoor space vary during the whole day. Due to this change in conditions, the signal strengths also vary. In indoor navigation while the user is moving, it is very difficult to label the instances during data collection correctly because user may be at the boundary of the cell or even moved to another cell.

We plan to investigate Multiple Instance Learning (MIL) which allows a number of instances to be batched into bags. The bag is labelled correct even if one instance of the bag belongs to the current location of the user. The characteristics of signal strength from APs can be captured by using the bagging of the instances and hence, the classification will be appropriate. State-of-the-art individual classifiers such as BayesNet, kNN, LibSVM may not be able to handle data collected from a dynamic environment and hence, we will study few MIL based classifiers such as CitationKNN, MISVM and use them for classification.

Bibliography

[1] Punith P Salian et al. “Visible light communication”. In: *2013 Texas Instruments India Educators’ Conference*. IEEE. 2013, pp. 379–383.

[2] Yusnita Rahayu et al. “Ultra wideband technology and its applications”. In: *2008 5th IFIP International Conference on Wireless and Optical Communications Networks (WOCN’08)*. IEEE. 2008, pp. 1–5.

[3] Zhu Jianyong et al. “RSSI based Bluetooth low energy indoor positioning”. In: *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE. 2014, pp. 526–533.

[4] Paramvir Bahl et al. “RADAR: An in-building RF-based user location and tracking system”. In: (2000).

[5] Moustafa Youssef and Ashok Agrawala. “The Horus WLAN location determination system”. In: *Proceedings of the 3rd international conference on Mobile systems, applications, and services*. ACM. 2005, pp. 205–218.

[6] Priya Roy and Chandreyee Chowdhury. “Smartphone based indoor localization using stable access points”. In: *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking*. ACM. 2018, p. 17.

[7] Feng Yu et al. “An indoor localization of wifi based on support vector machines”. In: *Advanced Materials Research*. Vol. 926. Trans Tech Publ. 2014, pp. 2438–2441.

[8] Yungeun Kim, Hyojeong Shin, and Hojung Cha. “Smartphone-based Wi-Fi pedestrian-tracking system tolerating the RSS variance problem”. In: *2012 IEEE International Conference on Pervasive Computing and Communications*. IEEE. 2012, pp. 11–19.

[9] Sudeep Pasricha et al. “LearnLoc: a framework for smart indoor localization with embedded mobile devices”. In: *Proceedings of the 10th International Conference on Hardware/Software Codesign and System Synthesis*. IEEE Press. 2015, pp. 37–44.

[10] Sunkyu Woo et al. “Application of WiFi-based indoor positioning system for labor tracking at construction sites: A case study in Guangzhou MTR”. In: *Automation in Construction* 20.1 (2011), pp. 3–13.

[11] Luca Calderoni et al. “Indoor localization in a hospital environment using random forest classifiers”. In: *Expert Systems with Applications* 42.1 (2015), pp. 125–134.

[12] 2019. URL: https://indoo.rs/amazon/.

[13] Lauren Davidson. *Trafford Centre owner unveils indoor mapping and deals app*. 2019. URL: https://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/11858755/Trafford-Centre-owner-unveils-indoor-mapping-and-deals-app.html.

[14] Dong-Peng Yang and Jin-Lin Li. “Research on personal credit evaluation model based on bayesian network and association rules”. In: *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE. 2007, pp. 3677–3680.

[15] David W Aha, Dennis Kibler, and Marc K Albert. “Instance-based learning algorithms”. In: *Machine learning* 6.1 (1991), pp. 37–66.

[16] John G Cleary and Leonard E Trigg. “K*: An instance-based learner using an entropic distance measure”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 108–114.

[17] Martin Azizyan, Ionut Constandache, and Romit Roy Choudhury. “SurroundSense: mobile phone localization via ambience fingerprinting”. In: *Proceedings of the 15th annual international conference on Mobile computing and networking*. ACM. 2009, pp. 261–272.

[18] Jean-Michel Akré et al. “Accurate 2-D localization of RFID tags using antenna transmission power control”. In: *2014 IFIP Wireless Days (WD)*. IEEE. 2014, pp. 1–6.

[19] Philipp Bolliger. “Redpin-adaptive, zero-configuration indoor localization through user collaboration”. In: *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*. ACM. 2008, pp. 55–60.

[20] He Wang et al. “No need to war-drive: Unsupervised indoor localization”. In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM. 2012, pp. 197–210.

[21] Valentin Radu and Mahesh K Marina. “Himloc: Indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced wifi fingerprinting”. In: *International conference on indoor positioning and indoor navigation*. IEEE. 2013, pp. 1–10.

[22] Shaoshuai Liu, Haiyong Luo, and Shihong Zou. “A low-cost and accurate indoor localization algorithm using label propagation based semi-supervised learning”. In: *2009 Fifth International Conference on Mobile Ad-hoc and Sensor Networks*. IEEE. 2009, pp. 108–111.

[23] Teemu Pulkkinen, Teemu Roos, and Petri Myllymäki. "Semi- supervised learning for wlan positioning". In: *International Conference on Artificial Neural Networks*. Springer. 2011, pp. 355–362.

[24] Mu Zhou et al. "Semi-supervised learning for indoor hybrid fingerprint database calibration with low effort". In: *IEEE Access* 5 (2017), pp. 4388–4400.

[25] Hyuk Lim et al. *Zero-configuration, robust indoor localization: Theory and experimentation*. Tech. rep. 2005.

[26] Deepak Vasisht, Swarun Kumar, and Dina Katabi. "Decimeter-level localization with a single WiFi access point". In: *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*. 2016, pp. 165–178.

[27] Matthew Cooper et al. "LoCo: boosting for indoor location classification combining Wi-Fi and BLE". In: *Personal and Ubiquitous Computing* 20.1 (2016), pp. 83–96.

[28] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[29] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[30] Ju-Hyeon Seong and Dong-Hoan Seo. "Environment adaptive localization method using Wi-Fi and Bluetooth low energy". In: *Wireless Personal Communications* 99.2 (2018), pp. 765–778.

[31] Pei Jiang et al. "Indoor mobile localization based on Wi-Fi fingerprint's important access point". In: *International Journal of Distributed Sensor Networks* 11.4 (2015), p. 429104.

[32] Kamol Kaemarungsi. "Distribution of WLAN received signal strength indication for indoor location determination". In: *2006 1st International Symposium on Wireless Pervasive Computing*. IEEE. 2006, 6–pp.

[33] David Mascharka and Eric Manley. "Machine learning for indoor localization using mobile phone-based sensors". In: *arXiv preprint arXiv: 1505.06125* (2015).

[34] Wendong Xiao et al. "Large scale wireless indoor localization by clustering and extreme learning machine". In: *2012 15th International Conference on Information Fusion*. IEEE. 2012, pp. 1609–1614.

[35] Priya Roy et al. "JUIndoorLoc: A Ubiquitous Framework for Smartphone- Based Indoor Localization Subject to Context and Device Heterogeneity". In: *Wireless Personal Communications* (2019), pp. 1–24.