Evauate Techniques for Wifi Locationing

IOT Analytics

Prepared by – Prakash Balakrishnan

Introduction

One of our clients is planning to develop a Indoor positioning system to be deployed on large industrial campuses, in shopping malls, et cetera to help people to navigate a complex, unfamiliar interior space without getting lost. On further research, we found that a lot of significant establishments had been made in the area of Wi-Fi Fingerprinting for Indoor Localization techniques. This is primarily due to increase in the accessibility of Wireless Local Area Networks and increased usage of mobile devices with Wi-Fi capability. Wi-Fi fingerprinting is RSSI based (Received Signal Strength Indication) and it simply relies on the recording of the signal strength from several access points in range and storing this information in a database along with the known coordinates of the client device in an offline phase. During the online tracking phase, the current RSSI vector at an unknown location is compared to those stored in the fingerprint and the closest match is returned as the estimated user location. This report evaluates different machine learning models to see which produces the best result, enabling us to make a recommendation to our client for incorporating into smart phone app that they will develop to determine a person’s location in indoor spaces.

Goal Summary

- Develop a system which can be incorporated with a smart phone app for Indoor locationing and/or help people to navigate unfamiliar Interior space without getting lost via wifi fingerprinting.

Data Source

Data was collected from the UCI Machine Learning Repository - **UJIIndoorLoc Data Set**

Link for the data – http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc

Structure of Data Set

Data Set contains ~ 21 K records(Training + Validation) with 529 attributes as mentioned below.

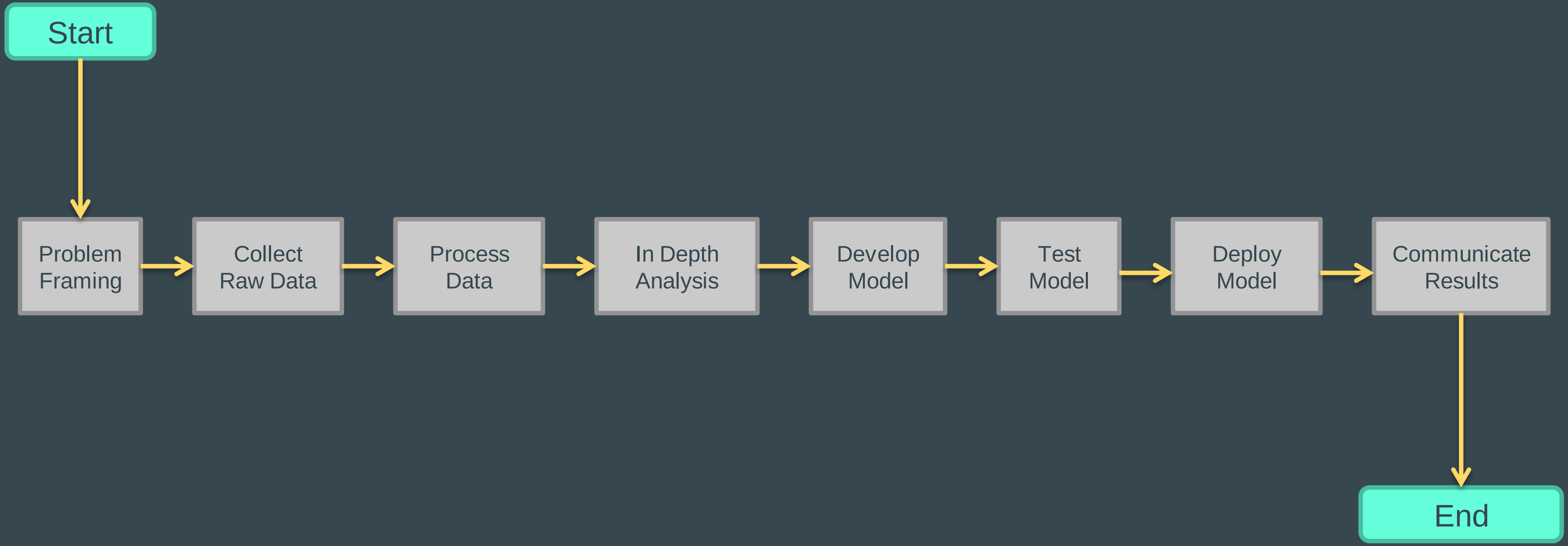
- **Attribute Information:**

Attribute 001 (WAP001): Intensity value for WAP001. Negative integer values from -104 to 0 and +100. Positive value 100 used if WAP001 was not detected.
....
Attribute 520 (WAP520): Intensity value for WAP520. Negative integer values from -104 to 0 and +100. Positive Vvalue 100 used if WAP520 was not detected.
Attribute 521 (Longitude): Longitude. Negative real values from -7695.9387549299299000 to -7299.786516730871000
Attribute 522 (Latitude): Latitude. Positive real values from 4864745.7450159714 to 4865017.3646842018.
Attribute 523 (Floor): Altitude in floors inside the building. Integer values from 0 to 4.
Attribute 524 (BuildingID): ID to identify the building. Measures were taken in three different buildings. Categorical integer values from 0 to 2.
Attribute 525 (SpaceID): Internal ID number to identify the Space (office, corridor, classroom) where the capture was taken. Categorical integer values.
Attribute 526 (RelativePosition): Relative position with respect to the Space (1 - Inside, 2 - Outside in Front of the door). Categorical integer values.
Attribute 527 (UserID): User identifier (see below). Categorical integer values.
Attribute 528 (PhoneID): Android device identifier (see below). Categorical integer values.
Attribute 529 (Timestamp): UNIX Time when the capture was taken. Integer value.

- **Missing Values :**

No missing Values were found in the data set.

Data Science Process Framework



Data Science Process Framework – Cont'd

Step 1 - Frame the problem

First thing to do before solving a problem is to define exactly what it is. The end goal here is that the system which is being developed and incorporated with smart phone app should be able to estimate Users location inside a building, campus etc. as accurate as possible.

Step 2 – Collect Raw Data

Once the problem is identified, data is required to provide some insights to turn the problem around with a solution. The source of data and high level insights were outlined in previous slides.

Step 3 – Process the Data for Analysis

This step involves processing of data before any analysis can be done on it. Often, data can be quite messy especially if its not maintained well. Such instances might corrupt actual analysis. Examples include Missing values in the data set, corrupted values like Invalid entries etc. For this particular dataset, there were no missing values.

Step 4 – In depth Analysis

This step of process is where deeper analysis is done on the data based on initial hypothesis by applying various statistical, mathematical and technical knowledge and leveraging on data science tools to crunch the data and find insights as much as possible to meet the business goal.

Data Science Process Framework – Cont’d

Step 5 - Train Model

This is the step where various Models are trained upon the training data and the most accurate model is selected. As part of this task, following 4 classification models were trained – KNN, SVM,Bagged CART and Random Forest

Step 6 – Test / Deploy Model

The trained model in the previous step is now tested on unseen data and made sure that it is able to predict the output with accepted level of accuracy. Once the model is tested, it can be deployed in real time environment to provide the actual functionality to the users.

Step 7 – Communicate results

This is the final step of the process where the data insights/ Visualizations / model performances will be communicated to the technical / Non – technical audience.

Data Preparation

Following modifications were done on the data set in order to find an optimized model for evaluating a person's indoor location -

- Removed the Timestamp feature / attribute from the data set
- The data sample was restricted to Building 2, Floor 2 due to limited availability of processor speed and memory.
- Attributes building Id, Floor, Space Id and Relative position were all combined into single attribute and later converted into 'Factor' data type. There were 73 levels in total.
- This sample database for Building 2 Floor 2 after the above changes resulted in 1577 observations with 523 features which was later split into 80% training set (1277 observations) and 20% test set (300 observations).

Model Performance - Summary

"Random Forest" was the most optimized model due to its highest Accuracy and Kappa value.

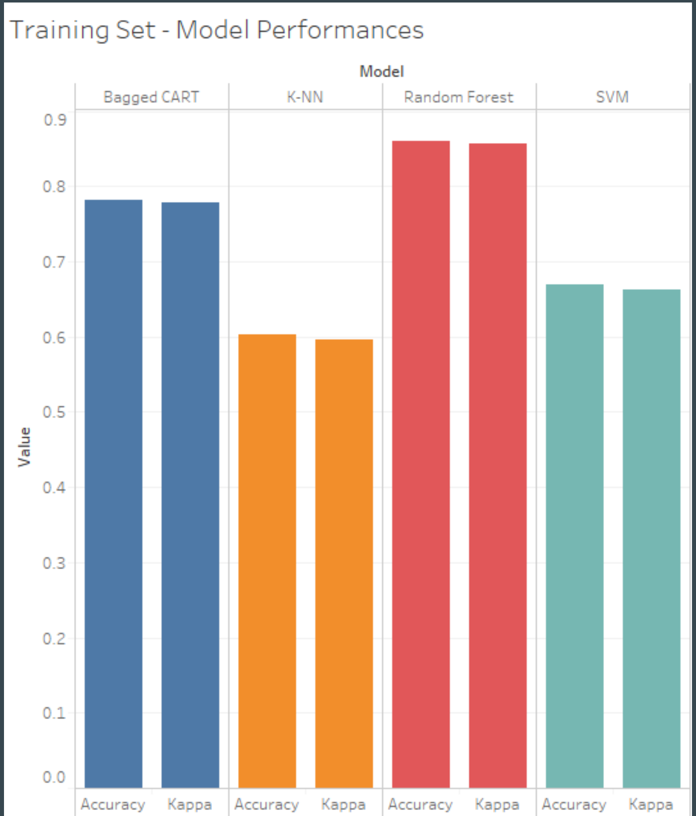
K-Nearest Neighbor (k=5)		
	Accuracy	Kappa
Training Set	0.6301655	0.6236861
Test Set	0.6733333	0.6678454

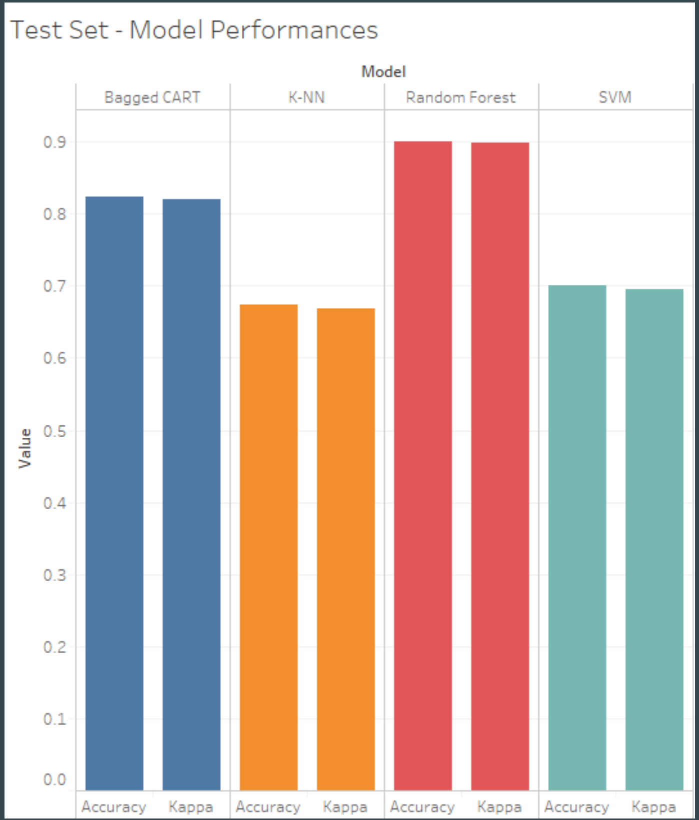
SVM		
	Accuracy	Kappa
Training Set	0.6732503	0.6675838
Test Set	0.7	0.6950359

Random Forest (mtry=32)		
	Accuracy	Kappa
Training Set	0.86640241	0.86399842
Test Set	0.9	0.8982855

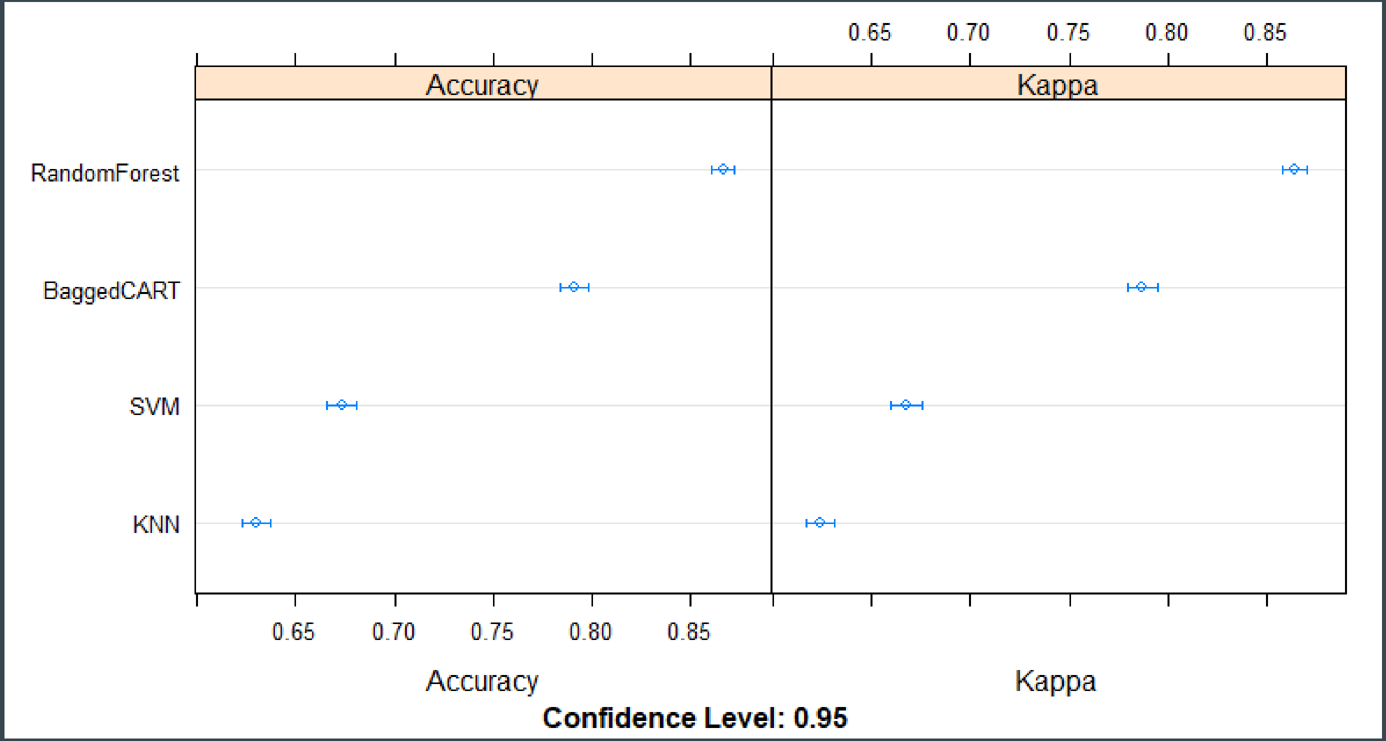
Bagged CART		
	Accuracy	Kappa
Training Set	0.7906982	0.7869962
Test Set	0.8233333	0.8203512

Model Performances - Comparison





Model Performances – Comparison using dotplot



Conclusion

The growing commercial interest in indoor location-based services (ILBS) has spurred recent development of many indoor positioning techniques. Due to the absence of Global Positioning System (GPS) signal, many other signals have been proposed for indoor usage. Among them, Wi-Fi emerges as a promising one due to the pervasive deployment of wireless LANs (WLANs). In particular, Wi-Fi fingerprinting has been attracting much attention recently because it does not require line-of-sight measurement of access points (APs) and achieves high applicability in complex indoor environment.

Pros –

- Random Forest model was able to predict the relative location of user inside a building with ~90% accuracy .
- Wi-Fi fingerprinting based approach is cost effective due to presence of existing infrastructure
- Room level details of an user's location can be efficiently calculated with Wi-Fi Indoor positioning system.

Cons –

- Model is unable to predict the direction in which a person is moving inside a building.
- Any changes of the environment such as adding or removing furniture or buildings may change the "fingerprint" that corresponds to each location, requiring an update to the fingerprint database. However, the integration with other sensor such as camera can be used in order to deal with changing environment

Other Localization Techniques

- RSSI and lateration based –

RSSI localization techniques are based on measuring signal strength from a client device to several different access points, and then combining this information with a propagation model to determine the distance between the client device and the access points. Trilateration (sometimes called multilateration) techniques can be used to calculate the estimated client device position relative to the known position of access points.Though one of the cheapest and easiest methods to implement, its disadvantage is that it does not provide very good accuracy, because the RSSI measurements tend to fluctuate according to changes in the environment.

- Time of flight (Tof) based –

This localization approach takes timestamps provided by the wireless interfaces to calculate the ToF of signals and then use this information to estimate the distance and relative position of one client device with respect to access points. The granularity of such time measurements is in the order of nanoseconds and systems which use this technique have reported localization errors.

- Angle of Arrival(AoA) Based –

With the advent of MIMO WiFi interfaces, which use multiple antennas, it is possible to estimate the AoA of the multipath signals received at the antenna arrays in the access points, and apply triangulation to calculate the location of client devices. SpotFi,[8] ArrayTrack[6] and LTEye[14] are proposed solutions which employ this kind of technique.

Questions?

Contact Information

Name : Prakash Balakrishnan
Email Id : prakash_balakrishnan@IOTanalytics.com
Contact No – 971-269-4970