Project Report on
Indoor Localization Using Machine Learning

Anshul Roonwal
SBUID: 110554783
Stony Brook University,
Stony Brook, NY11790
aroonwal@cs.stonybrook.edu

Aman Raj
SBUID: 109939872
Stony Brook University,
Stony Brook, NY11790
amraj@cs.stonybrook.edu

Neeraj Dixit
SBUID: 109951838
Stony Brook University,
Stony Brook, NY11790
ndixit@cs.stonybrook.edu

1. Abstract

Cellular (network-side) localization refers to positioning mobile phone users using solely cellular signals. Many other terminologies like Indoor Positioning system (IPS) are coined for the same concept. The goal is to identify user's indoor locations (granularity to a room level) using such MR (Measurement Reports) signals in real time. One might argue that we have GPS measurements available which can accurately find a location with an accuracy of a few meters, but not all mobile devices carry a GPS chip and even if they carry there can be errors in the signal due to atmospheric disturbances which can distort the signal before they reach a receiver. [1].

Once we rule out GPS we are only left with terrestrial signals to prediction the location. Experience has shown us that a straightforward application of measuring cellular cell strength (MR signals) does not work very well indoors. The obvious shortcomings are again any obstructions such as buildings or trees and also the proximity to the cell phone tower. With all this being mentioned here, do we have any other technology that we can combine with cellular strength to better estimate the location of the user inside a building, or in fact inside a room. We try to answer these questions through this project.

2. Problem Statement

Cellular strength measurements may or may not work well indoors owing to numerous reasons. The measured cellular cells and Wi-Fi signal strengths can be deteriorated by civil structure like walls, poles. So we cannot use the relation between measured signal strength and location as we might read a corrupted signal strength. This would make the triangulation process prone to errors because distance estimates itself may be erroneous and the circles may not intersect at a single point. This makes it difficult to estimate approximate location considering only signal strength

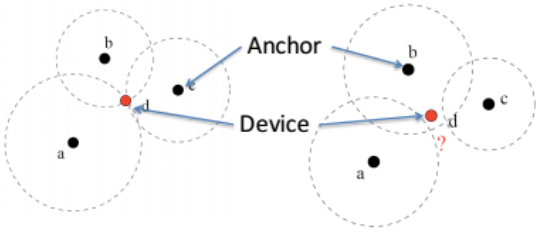

Fig. 1 Trilateration using distance estimates

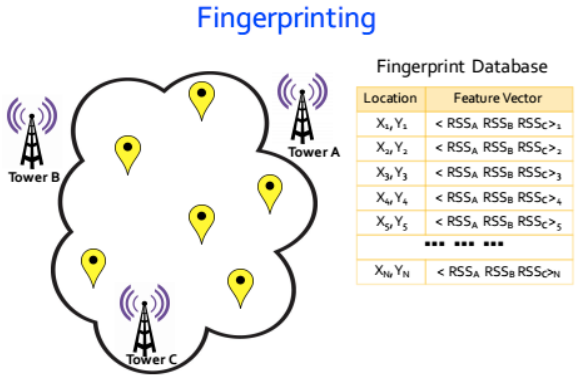

Fig. 2 Fingerprinting

3. Relevant Prior Work

In a paper by the name Network side positioning of cellular-band devices with minimal effort by, the researchers have addressed the problem of network-side localization where cellular operators are interested in localizing cellular devices by means of signal strength measurements alone. While fingerprinting-based approaches have been used recently to address this problem, they require significant amount of geo-tagged (‘labeled’) measurement data that is expensive for the operator to collect. Their goal was to use semi-supervised and unsupervised machine learning techniques to reduce or eliminate this effort without compromising the accuracy of localization.

4. Dataset Descriptions

The dataset is obtained from a mobile application which reports the signal strengths from cellular cells and nearby Wi-Fi devices. The application records the received signal strength from multiple cellular towers as well as Wi-Fi signal strengths at a duration of 1 second. Along with the signal strength this data is labeled with the known location from where the observation is recorded.

5. Approach for Solution

In this project we have used a Machine Learning based solution proposed by Ayon Chakraborty, Smir R. Das, Luis E. Ortiz in his research paper *Network-side Positioning of Cellular-band Devices with Minimal Effort* [2]. The benefit of using Machine learning algorithm is that it eliminates the errors due to variations in signal strength measurement. The classifier focuses only on prior radio signal survey on the area of interest and align this towards Fingerprinting based approach. We are using Naïve Bayes classification for this purpose. We are interested in the granularity up to which correct results can be obtained by Naïve Bayes classification. The aim of this project is to determine the accuracy by which we can predict the location of a device using Cellular signal strength. We have also analyzed the gain in accuracy that can be achieved by supplementing the data with Wi-Fi signals. The major focus would be to rely only on cellular cell strength as there is more probability of cellular signal available at any location. Wi-Fi access points may not be available at all locations all the time.

6. Design of Classifier

The measured signal strengths constitute the feature set for Naive Bayes classification. A detailed analysis of the prediction is presented on the basis of subset of features like cellular signal strength only, Wi-Fi signal strength only and a combination of cellular cells and Wi-Fi signal strengths.

6221	5518	....	....	00:24:6c:31:b2:73	00:24:6c:31:b1:42
0	0	...	---	-56	-71
-72	0	...	---	0	-33
0	-35	...	---	0	0
:	:	:	:	:	:

Fig. 3 Classifier structure for Naïve Bayes

Figure 3 depicts the feature vectors that were used to be given as the input to the classifier. The column headings '6221', '5518' and son on shows the cell Ids of the base stations that were being observed at a particular location of data collection. Similarly, the Mac Id of the Wi-Fi have been used as the combined feature vector besides the cell Ids. Each record in the data frame represents an observation that collects the signal strengths and Wi-Fi strengths at a given location. We have trained Naïve Bayes classifier after splitting the data into 80% and 20% for training and testing respectively. Though we have captured the Wi-Fi signal strengths, we have calculated the performance of our model after using only cellular strengths, after adding a few most contributing Wi-Fi and finally after using all the Wi-Fi.

7. Results

We have measured the accuracy of the classifier using 2 different approaches:

1. Single label per room

Data is collected at different locations of one room and the same label is given to readings obtained in same room. This data is used just for training the Naive Bayes classifier. The testing data is collected at random locations inside the room.

1. Accuracy of 60% when we use only the cellular signal strength
2. Accuracy of 90% when we use the top 10 Wifi with highest signal strength
3. Accuracy of ~100% when we use all Wifi signal strength along with cellular signal strength

2. Multiple labels per room

Data is collected at different locations and different labels are given to each reading. The data collected for analysis is split into ratio of 80% for training and 20% for testing the trained model

1. Accuracy of 52% when we use only the cellular signal strength
2. Accuracy of 88% when we use the top 10 Wifi with highest signal strength
3. Accuracy of 97% when we use all Wifi signal strength along with cellular signal strength

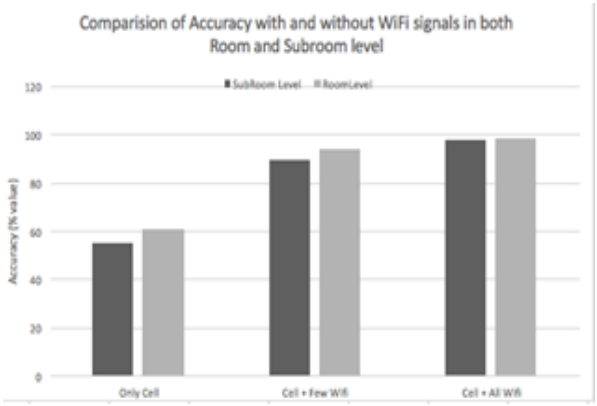

Fig. 4 Comparison of accuracy with using Wifi and without using Wifi

Figure 4 shows clear comparision of the two appraches we used here. Accuracy with single room per label has been observed to be better than the approach where we are using multiple labels. Also, the more information we added besides the cellular strengths, the more accurate results we got. But the aim to make a trade off between the amount of Wi-Fi we use and the accuracy that we obtain. The results for which are shown in the next plot.

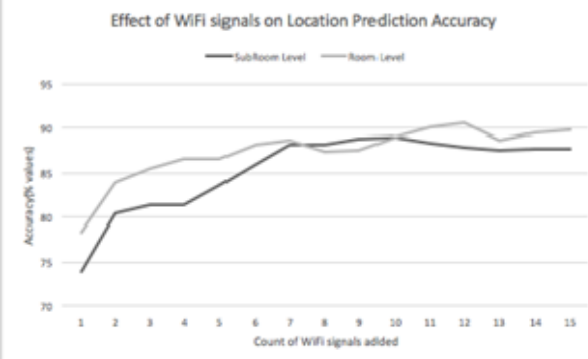

Fig. 5 Accuracy on the basis on number of Wi-Fi signals

Figure 5 shows that after adding around 6 prominent Wi-Fi signals, we are getting an accuracy that doesn't improve much over addition of any more Wi-Fi.

8. Analysis & Findings

The above analysis shows that while Cellular Signal Strength all by itself is able to predict the location with a commendable accuracy. Adding the data of even a few Wi-Fi signal would shoot the accuracy to a near perfect levels.

9. Conclusion

The above experiment has proved that given sufficiently large training data set, we can use machine learning approach towards solving the problem of Indoor Localization. While cellular signal strength provides a rough estimate, adding a few Wi-Fi signals for localization boosts the accuracy of location prediction to almost ~90%

9.1 Assumptions

We have assumed following things in the above approach:

1. N signal strengths uniquely identify a single point inside a premises.
2. Although the measured signal strengths may be incorrect, they are consistent at the same location.
3. The location samples given to train the classifier model are evenly distributed.
4. The location prediction can be real time only if sufficient training data is provided to classifier.

10. Further Improvements

Naive Bayes classifier is most basic algorithm that can be used for classification. The classification accuracy depends highly on the training data. Inaccurate and noisy training data will lower the accuracy of the classifier and so will the uneven distribution of data samples from different locations.

As a solution to avoid the shortcomings of Naive Bayes can use 'Unsupervised Learning' algorithms like K means to remove the necessity of manual labelling the training data. We can further use 'Hidden Markov Model' to probabilistically determine the next location based on history of mobility obtained from the sequence of data points. The cellular signal strengths would act as evidence variable to further ratify the prediction. This model would personalize the location history on the basis of existing historical data as a predictive model, considering people don't change their lifestyle often.

11. References

[1] http://www.mio.com/technology-gps-accuracy.htm

[2] Network-side Positioning of Cellular-band Devices with Minimal Effort.

[3].https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[4]. http://scikit-learn.org/stable/modules/naive_bayes.html