Tran, Tony

Wi-fi Indoor Location Evaluation

**Introduction**

Our client hopes to implement an indoor positioning service an average consumer can use to simplify navigation of indoor spaces. Currently, GPS (Global Position Systems) provide this service for navigating locations outdoors. Research of how to implement a system indoors has shown substantial progression in Wi-fi fingerprinting. As Wireless Local Area Networks are ubiquitous in public spaces, along with widespread use of Wi-fi capable devices, the adoption of an application to provide indoor navigation should likely increase. To estimate the position of a user with a connected device, the application uses RSS (Received Signal Strength) of the receiving device to Wi-fi access points within range. The strength of the signal approximates where the receiving device is located. We created and compared models to investigate whether a reliable machine learning approach could be applied.

**Objective**

The goal of our evaluation is to create a reliable system for indoor positioning that can be easily accessed and used through a smart phone application. This will be available through Wi-fi fingerprinting of indoor spaces. The main application is for a user to be able to find a location indoors such as a student looking for a classroom, or a shopper looking for a retail store in a shopping center.

**Overview of Process**

1. Frame the problem – Specify what the business goal is and what we are trying to answer.
2. Collect and assemble data – Gather the possible data necessary to achieve the goal.
3. Evaluate data – Understand the data and analyze. Discern the applicable parts of the data to run models that can produce desired results.
4. Pre-process data – Organize, rearrange, filter, transform, remove, combine data as necessary in order to fulfill the needs of the models applied.
5. Train and test models – Put the data through models and produce metrics to be assessed.
6. Evaluate model performance – With performance metrics, evaluate how well the models addressed the problem.
7. Communicate results – Provide recommendations from the model evaluation for business.
8. Continue to collect data – Use new data for assessment and updating model.

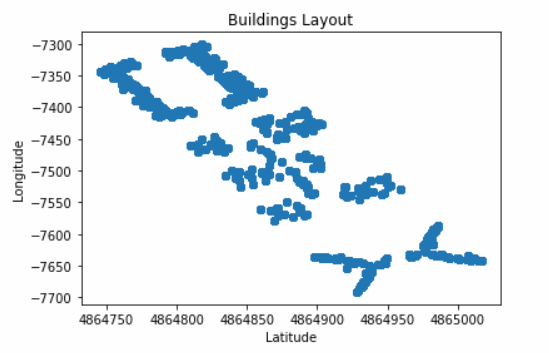
**Data Source and Structure**

Data was provided from the UJIIndoorLoc Data Set available from the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc>). The validation data set was not used in our evaluation. The training data set has 19,937 records with 529 attributes. There were no Missing Values found in the data set. Attributes 1 to 520 are displayed as WAP### and are the observed signal strength for the Wi-fi access points. These are recorded as values from -104 to 0, and +100. A +100 indicates WAP### was not detected. Attributes 521 to 529 are described in the following table.

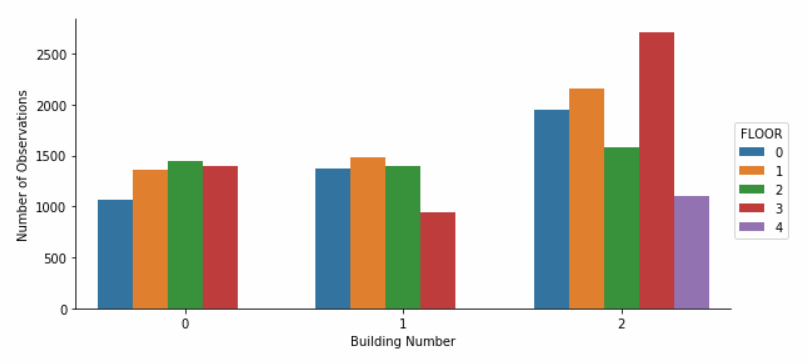
|  |  |
| --- | --- |
| Attribute | Description |
| WAP 001 - 520 | Signal strength values from -104 to 0, and +100 |
| LONGITUDE | Position by longitude ranging from -7691.338 to -7300.819 |
| LATITUDE | Position by latitude ranging from 4864746 to 4865017 |
| FLOOR | Floor levels inside of building with values from 0 to 4 |
| BULDINGID | Building Identity of the three buildings with values from 0 to 2 |
| SPACEID | Identification number assigned to spaces in building (classroom, office, etc.) |
| RELATIVEPOSITION | Position in relation to SpaceID (1 – inside space, 2 – outside of door of space) |
| USERID | User identifier with values from 1 to 18 |
| PHONEID | Device identifier with values from 1 to 24 |
| TIMESTAMP | Time of capture of recorded observation |

**Data Management and Processing**

For the purpose of our evaluation, we modified the data to address the intended use and fulfilling our objective. The intended user is a smart phone owner entering an unfamiliar indoor space. With a phone application, they could navigate to the desired location within the space. Outside of a building or complex, GPS can handle navigation. A unique dependent variable was created by combining FLOOR and SPACEID. The LONGITUDE, LATITUDE, RELATIVEPOSITION, USERID, PHONEID, and TIMESTAMP attributes were removed, since we are testing the ability of a model to find a position by using existing Wi-fi infrastructure. The original WAP values were altered to change the +100 to 0 to signify no signal was detected. The values -104 to 0 were altered to 1 to 105 for weak to strong signal strength. We filtered the data by Building number and created models for Building 2. The results of one building provide a framework for the process of applying to the other buildings and other indoor spaces outside this dataset.



Observations per Floor by Building



**Model Performance and Comparisons**

We tested five different algorithms for the dataset on the initial run through of 1000 samples. The algorithms tested were Random Forest, k-Nearest Neighbors, Support Vector Machines, and Gradient Boosting Classifier. The Accuracy was used to select the models to move forward with on full set. Below we have displayed the values of accuracy produced from each model.

|  |  |
| --- | --- |
| Accuracy of 1000 Samples for Each Model | |
| Random Forest | 71.0% |
| k-Nearest Neighbors | 72.0% |
| Support Vector Machines | 52.7% |
| Gradient Boosting | 52.3% |

The Random Forest and k-Nearest Neighbors produced the greatest accuracy of the algorithms. They were used to created models on full Building 2 data set. Random Forest was tuned by setting a higher number of estimators to 100. The Random Forest model produced 86% accuracy. K-Nearest Neighbors was tuned producing best results with neighbors at 1, jobs at 2, and metric set to ‘manhattan’. The k-Nearest Neighbors model produced 82.8% accuracy. A Neural Network model was also created and run on the entire Building 2 dataset. The Neural Network yielded 78.4% accuracy. The results for Random Forest were accurate to a greater degree than other algorithms. Furthermore, the cross validation applied to the Neural Network increased computing resources and time. We also looked at precision, recall, and f1-scores for Random Forest and k-Nearest Neighbors. With .91 precision, .86 recall, and .87 f1-score, the Random Forest model outperformed k-Nearest Neighbors in these metrics as well.

**Recommendations**

The model we created produces a relatively reliable ability to find a position based on the signal strength of Wi-fi access points. Based on the performance metrics of accuracy, we found the Random Forest algorithm to produce the best results. Because of existing Wi-fi infrastructure, wide adoption of smart phones, and captured data, we can simplify implementation for a large population of users. Predicting the position at a room level with this approach can be accomplished and addresses how the goal of our evaluation has been framed.

However, there are some issues that need to be addressed. The use of captured data means there needs to be updated information for training models. This is particularly vital if existing infrastructure changes or has alterations. Building rooms may change or access points might be situated elsewhere as time goes on. To adjust for this, continual monitoring and labelling of the Wi-fi fingerprints would need updating.