3/15/2022

Kathy Narim-Nomstrom

VP, IOT Analytics

Overview:

As per your request, I have compiled an internal report detailing the evaluation of multiple machine learning models in order to accurately navigate indoor locations using WiFi positioning.

**Goal of the Project:**

This project aims to provide models accurate enough to be incorporated in a mobile phone in-door navigation app.

**Where is the data**:

The data set from 2013 is available at the UCI Machine Learning Repository: <http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc>

**Information about the data**:

The data set contains no missing values. The UJIndoorLoc database surveys three buildings of Universitat Jaume I, in Valencia, Spain, with at least 4 floors and an area of nearly 110,000m². There were more than 20 different users and 25 Android devices used to track location. There are 529 attributes containing the WiFi fingerprint, including the coordinates where it was taken. Other attributes contain location data (latitude, longitude, and floor), Building ID, Space ID (offices, labs, etc), relative position (inside/outside the space) where the recording was captured, User ID with user height, and timestamps.

**POA**:

The initial task entailed deciding which R libraries to utilize. I loaded ‘readr’ in order to read the csv dataset. The remaining libraries with be mentioned as the POA progresses. In order to build models faster, ‘doParallel’ was loaded and a 4 core cluster was created. I then used ‘plyr’, ‘dplyr’, and ‘explore’ to perform the EDA. I double checked for missing values and found none.

The dataset was very large, therefore I created a subset to focus only on Building 2, which contained 47.6% of all observations.

I then needed to create a new attribute which merged FLOOR, SPACEID, and RELATIVEPOSITION to LOCATION. LOCATION needed to be converted to a new data type, a factor.

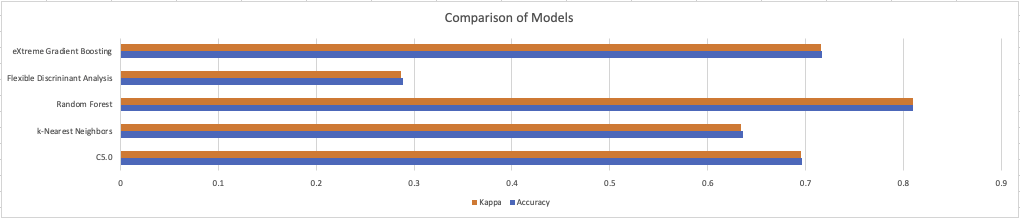
I nulled the following attributes: LONGITUDE, LATITUDE, FLOOR, BULIDINGID, SPACEID, RELATIVEPOSITION, USERID, PHONEID, TIMESTAMP. The only attributes needed for building classification models are the 520 Wireless Access Points and LOCATION, the latter being the dependent variable.

To further reduce the size of the dataset, I used near zero variance (NearZeroVar) to remove WAP attributes with a value of 100, which means ‘not used’.

I used CARET for training, testing, and classification modeling. I set the seed to ‘123’ with a 70/30 split for all models to keep the modeling consistent. The cross validation was set to 10 with 5 repeats. I created the following models: C5.0, k-Nearest Neighbors, Random Forest, Flexible Discriminant Analysis, and eXtreme Gradient Boosting. Each model had a tune length of 5.

Then I used the predict function on each model in order to create a confusion matrix to visualize the model results.

**Comparison of models:**



Out of the 5 models built, the best performing models were Tree models. Random forest had the highest Accuracy and the highest Kappa scores. According to Kappa value interpretation Landis & Koch (1977), a Kappa score between .81-1.0 is considered perfect. The Random Forest model produced a Kappa score of .809 with a slightly higher Accuracy. Based on this interpretation, eXtreme Gradient Boosting and C5.0 come close but with Kappa scores registering as substantial, being between .61-.80. According to the resulting data, Tree models perform the best with the given dataset.

**Recommendations:**

Classification modeling was used for this task. This meant that only location, as the dependent variable, was used against 520 wireless access points. Adding Channel Status Information (CSI) would greatly improve upon the limitations of measuring just the RSS signal strength by providing more fine-grained physical layer information. This is only one method of modeling indoor navigation. Another method would be to build regression models utilizing actual longitude and latitude estimation.

It would also be possible to calculate indoor positioning using the timestamp. This is done in a number of ways: Time of Flight (TOF)/Time of Arrival (TAO), Time difference of arrival (TDOA)

In regards to methods not possible with the current data set, Angle of Arrival (AOA) and Angle of Departure (AOD) could possibly be more accurate but require Bluetooth hardware.

GRANT JOHNSON