This is a report on developing Machine Language models to locate an individual’s indoor location within three buildings using the buildings’ wifi-modem network. The dataset is sourced from UCI’s Machine Learning Repository and contains 21,048 records and 528 Attributes.

Data Set / modeling challenges:

1. Wide data set with 529 Attributes.
   * 519 of the attributes are wifi-modem (WAP) intensity readings (-104 to 0, or +100).
     + + 100 meaning no intensity reading
   * 2 attributes are longitude and latitude
   * 3 attributes are location (building, floor, room)
   * 1 attribute is relative position (inside the room or outside-the-door).
     + Approximately 80% of the readings were outside-the-door.
   * 1 attribute is UserID
   * 1 attribute in Phone Type ID
   * 1 attribute is UNIX time when intensity reading was taken
2. Sparce data set: Most values in the wifi intensity reading (WAP) columns did not have an intensity reading (i.e. 100)
3. High number of levels for the dependent variable (i.e. large factor model) (737)

Data Prep:

To overcome above stated challenges the following steps were taken.

1. All three location attributes (building, floor, room) were combined (concatenated) into one column named LOCATION. Relative Position was not included into this combined value since 80% of the values were the same (outside the door) and it is believed that the cost in model performance by including it would not be offset by the value gained.
2. Feature Selection: All the wifi-modem intensity features were used along with the new combined attribute LOCATION as the dependent variable. The remaining attributes were dropped for reasons as follows:

*However, if lat/long were accurately obtained from phone readings, then isn’t this exercise academic?*

* + - Longitude – correlated to target variable
    - Latitude – correlated to target variable
    - Floor – concatenated into LOCATION
    - BuildingID – concatenated into LOCATION
    - SpaceID – concatenated into LOCATION
    - RelativePosition – value not worth complexity
    - UserID – not relevant
    - PhoneID – not relevant
    - Timestamp – not relevant

1. The data were split by each of the three buildings, creating three sub-data-sets to be modeled. This reduced the number of attributes and the number of levels of the dependent variable for each of the models.
2. WAP columns with no variability were removed (i.e. wifi-modems that did not have any intensity readings) further reducing the total number of attributes for each data set.
3. Structure of the final data sets are:
   * Building 0 = df1 (5249 obs. of 201 variables)
   * Building 1 = df2 (5196 obs. of 208 variables)
   * Building 2 = df3 (9492 obs. of 204 variables)

Modeling

Due to the size of the datasets, the parallel processing capabilities of my computer were activated such that 6 cores were used in parallel to train the data on the modeling algorithms.

R programing within the R-carat package is used to execute the ML.

Four categorical ML algorithms are explored to develop the best model using the metrics of Accuracy and Kappa for performance evaluation. These are:

* Random Forest (rf) => took between 17 and 35 min. to run each of the data-sets
* C5.0 (C5.0) => took between 20 and 30 min. to run each of the data-sets
* Stochastic Gradient Boosting (gbm) => system crashed twice, no results yielded
* K-Nearest Neighbors (knn) => took between 5 and 10 min. to run the data-sets

All modeling was done with 10-fold cross validation. Repeating of cv 5 and 3 times on the first run of the Random Forest algorithm was explored, however, metric performance improvement was not offset by increased computational cost, so all final models used a repeat of 1.

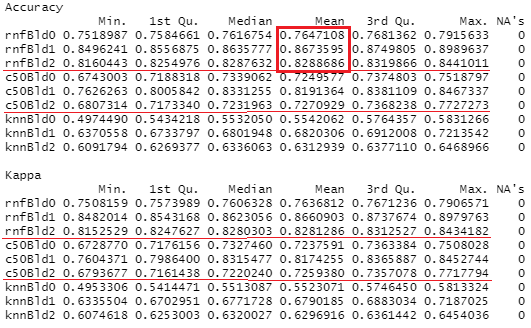
The parameter optimization feature (tuneLength) of the R-carat package is used to tune all of the training runs. A tuneLength of 5 is used for all models and results showed that a further tuning was not needed (e.g. tuneLength > 5 would not yield better performance). Most models found optimal parameters within 3 iterations of tuneLength. Manual fine tuning for further optimization was not performed due to computational cost and time.

Model Performance Results:

Results of each of the models is shown in the table below. As mentioned above, gbm is not in the table as my system crashed when attempting to train on gbm due to machine limitations (memory and parallel processing power) and the complexity of running wide data-sets across a boosting algorithm.

The table below shows that Random Forest performed the best on all three data-sets in terms of mean accuracy and mean kappa. The rf models also show more consistency across the 10-fold cv’s as they also have the least variability (i.e. Max – Min) for both accuracy and kappa.

Training Model Performance



rnf = Random Forest, c50 = C5.0, knn = K-Nearest Neighbors

Building-1 data set performed the best across all models while Building-0 was the worst performer in all cases. Not sure why as Building-1 has the least observations and the most variables which is counter intuitive to data-set complexity.

Model Selection and Performance on Test Data-Set

Because of the above stated performance, the Random Forest models are selected and applied to the test data set. Results of predictions are listed in table below.

Predicted Performance of Test Data Sets (rnf)

|  |  |  |
| --- | --- | --- |
| **Building** | **Accuracy** | **Kappa** |
| Bldg - 0 | 0.708 | 0.707 |
| Bldg – 1 | 0.820 | 0.818 |
| Bldg – 2 | 0.741 | 0.740 |

Performance of predicted values came in below the performance of the training data between 5% and 8%.

Conclusion:

From current modeling results of wifi-localization: depending on building, we can determine a person’s location in the building by room and floor with a confidence level between 70 and 80%. Granularity of location being from just outside the door to inside the room. Building 1 having the higher confidence level 80% and Building 0 having the lower at 70%.

Recommendations:

To improve modeling the following recommendations are provided:

1. Take several readings from perimeter of each room. This will minimize potential of cross signaling with other rooms and better define a room’s location.
2. Use consistent equipment to take readings (i.e. same phone make/model). This will remove any differences in data collection equipment signals.
3. Further research/understanding is needed on how to best model large factor models. That is, did we apply the best algorithms for large factor models?
4. Improve machine hardware to run more complex modeling methods.
5. Investigate other technologies that are used for indoor positioning. These technologies include:
   1. Ultra Wide Band (UWB): Available on later phone models with 5G capability.
   2. Bluetooth - would require phone capability be turned on
   3. RFID – also requires some receiver to be worn by the individual.

One company that is very involved in developing indoor localization system is omlox and has systems that can work with multiple technologies at once.