MSDS 7333 – 406

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Probabilistic Indoor Positioning Systems

Case Study

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Martin Garcia, Michael Catalano, Jeremy Otsap, and Christian Nava

May 20, 2020

Introduction

Indoor positioning systems (IPS) can help enhance, or identify, critical business needs where location-based monitoring is available. The need to know where certain items or people are in real-time can improve asset management, increase security, reduce operating expenses, and optimize resources. Large facilities or campuses can provide users with dynamic, turn-by-turn navigation. Retailers can utilize IPS technology for location-based marketing (couponing), provide a user with a customized experience, or guide a shopper to the most convenient and available parking space. Hospitals, warehouses, and airports can employ IPS technology to track their assets, make responses to maintenance issues more efficient via push notifications, or optimize the flow of traffic.

Technologies used by IPS include beacons that transmit radio signals using Bluetooth technology to locate the user, Wi-Fi signals that locate a user based on the signal strength, and light-based positioning that uses LED lighting to position people and objects. As per the Nolan and Lang text, GPS tracking may pose challenges when used inside buildings.[[1]](#footnote-2) However, by utilizing Wi-Fi signals detected from network access points, near real-time positioning can be realized with minimal training, calibration, and equipment.

In their 2006 paper, King, Knopf, Haenselmann, Lubberger, and Effelsberg created a probabilistic indoor positioning system using the Wi-Fi signal strength between a wireless local area network (LAN) access point and mobile devices to locate users.[[2]](#footnote-3)

In this case study the approach of King et al was reviewed and the experiment modified to see if any improvements in location accuracy were generated. More specifically, it incorporates clustering techniques to predict the location of the “online” dataset, it includes measurements from an access point (MAC ID) that was initially excluded, it applies a distance-based weight to a k-nearest neighbors classification model, and it compares the results of three different scenarios: (1) using the excluded MAC ID instead of the original used by King et al, (2) using both the original from King et al and the excluded MAC ID, and (3) using the original MAC ID from King et al.

Summary of Experiment

The King et al experiment uses the 802.11 beacon signal-strength readings from six stationary access points (Wi-Fi routers) located in a 36- by 15-meter area on a single floor of an office building at the University of Mannheim. Each access point is uniquely identified by a 48-bit hexadecimal MAC address. Offline measurements were taken for reference using a hand-held device on a grid of 166 points spaced one meter apart. This location grid offers a point of reference to compare future movements and give an estimate of its location with respect to the signal strength of the stationary access points.

Given that the human body will block the signal of 802.11 radios, signal strength will vary depending on the user's orientation.[[3]](#footnote-4) To account for this, signal strength readings or each (x, y) position were recorded at eight different orientations in 45-degree increments from 0 to 315 degrees.

The offline recordings were used to train the positioning system. An additional set of data called the “online” data and composed of 60 randomly selected location-orientation combinations, was used to test and make predictions. The location of a new “online” source was estimated with a clustering algorithm that classifies location based on comparative signal strength of the reference signals. It uses the closest match, or approximation, of Euclidean distance to the “offline” reference neighbors to estimate the location of the nearest group, or cluster, of reference (x, y) coordinates.

Note the word *estimate* above. Given the nature of the signal latency, the distance between access points, phase cancellation, environmental factors, variance in signal amplitude and distortion, slight variations in the angle of measurement, translate to the fact an exact location cannot be given. In short, there is variance between redundant signal measurements.

The goal is to find the model that most accurately estimates the actual position.

Data Description

Over one million measurements of signal strength were analyzed, termed the "offline" dataset, which contains signal strength measures on a grid of 166 points, spaced a meter apart, within the building floor. A second dataset, termed the "online" data, is comprised of 60 locations and orientations chosen at random throughout the floor, and is used to test the model for predicting location.

As seen in the Nolan and Lang text and throughout the replication of King’s work, the data is noisy, requires a considerable amount of formatting, and requires some assumptions to be made on behalf of the modeler. Multiple iterations of KNN runs need to be run to assess the fit and performance based on the characteristics found in the data.

A description of variables taken from the Nolan text is found in Table 1.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| time | date/time | Timestamp in milliseconds since midnight, January 1, 1970 UTC |
| id | string | MAC address of the scanning device (wireless router) |
| pos | numeric (continuous) | Physical coordinate of the scanning device |
| orientation | numeric (continuous) | Orientation of the user carrying the scanning device (angle in degrees) |
| MAC | string | MAC address of a responding peer (e.g., an access point or a device  in adhoc mode) with the corresponding values for signal strength in dBm (Decibel-milliwatts), the channel frequency and its mode (access point = 3, device in adhoc mode = 1) |

*Table 1. Description of study variables. Measurements include online and offline datasets.*

Data Cleaning & Preparation

The data must go through several transformations, not only for interpretability, but for modeling reasons. Trying to model the initial raw data would end in disaster as it is only one long string of ‘tokens’ separated by semicolons and other notations. The format consists of uneven rows of signals recorded from multiple sources received by our scanning device per location and orientation. The data is split into two respective data matrices, one for our scanning device, position, and angle and a separate matrix for signals detected by our device at given angle and position. These matrices are combined to create a data frame.

Extensive data munging is required for our k-nearest neighbor model to process. The raw data has several signals recorded at every combination of access point, location, and orientation. This requires us to create a single summary signal strength for each location, orientation, and access point combination. We utilize the mean value of the signal strength for each location by orientation and access point. Not all our data is collected in 45-degree increments and must be rounded to the minimum absolute difference of the actual orientation minus the range the range of angles, from 0-315.

Several irrelevant data are also scrubbed from our final dataset including scanning device mac and position z. Time is also simplified from milliseconds to seconds. These changes help contribute to the interpretability for our KNN model and our team. To further simplify readability the structure of the data frame is flipped from long to short format. This creates a new column for each access point. Important to note when looking at distributions of signals strength while holding location constant is the impact of the orientation and access point distance. This translates to controlling the angles used in training to match our test set of mixed angles. The angle we decide to use act as parameters which can be tuned with our function selectTrain.

Model Description

*Performance Metric*

For the model to make its best prediction, a loss function, L(y, f(x)), that will penalize prediction errors is needed to quantify how well the model performs. In this case, the sum of squared errors (SSE) is the best choice to evaluate model performance since it provides a good measure of error when using Euclidean distance.



The sum of squared errors measures the sum of the difference between the predicted values, yi , and the actual values, xi. The best performing model, or the best fitting model, will minimize the loss function, i.e., it will have the smallest error between the predicted and the actual values. The SSE will be used as the metric to compare the performance of the clustering models. It will provide a measure of the error between the predicted signal location and the actual signal location.

*Clustering Model One*

With our data prepared and functions created, we’re ready to train the model which most accurately estimates position. Our approach is to loop through iterations of values for *k* with the goal of minimizing the sum of square error. Specifically, we’ve built three KNN models: one using all seven access points (KNN Test 1), and a combination of leaving out access points ending in *c0* and *cd* (KNN Tests 2 and 3, respectively).

|  |  |  |
| --- | --- | --- |
| **KNN Test 1** | **KNN Test 2** | **KNN Test 3** |
| 00:0f:a3:39:e1:c0 | 00:0f:a3:39:e1:c0 | ~~00:0f:a3:39:e1:c0~~ |
| 00:0f:a3:39:dd:cd | ~~00:0f:a3:39:dd:cd~~ | 00:0f:a3:39:dd:cd |
| 00:14:bf:b1:97:8a | 00:14:bf:b1:97:8a | 00:14:bf:b1:97:8a |
| 00:14:bf:3b:c7:c6 | 00:14:bf:3b:c7:c6 | 00:14:bf:3b:c7:c6 |
| 00:14:bf:b1:97:90 | 00:14:bf:b1:97:90 | 00:14:bf:b1:97:90 |
| 00:14:bf:b1:97:8d | 00:14:bf:b1:97:8d | 00:14:bf:b1:97:8d |
| 00:14:bf:b1:97:81 | 00:14:bf:b1:97:81 | 00:14:bf:b1:97:81 |

*Table 2. Access points associated with each KNN model*

*Clustering Model Two*

As posited by the King et al study, the previous methods use the *k* nearest neighbors to a set of signal strengths, and the known (x,y) values for these neighbors are averaged. Our second model modifies the KNN function and generates a weight value for each neighbor. The weights are inversely proportional to the “distance” (in signal strength) from the test observation. We perform this on the best result from above using all seven access points and a *k* of five.

*Results*

The weighted average prediction, while not groundbreaking, manages the lowest sum of squares error prediction (**209.2** with a k=5). Table 3 describes the error results of each model.

|  |  |
| --- | --- |
| **KNN Model** | **Sum of Squares Error** |
| Using All 7 MAC IDs | 209.61 (*k = 5)* |
| Removal of c0 | 249.92 (*k = 5)* |
| Removal of cd | 273.68 (*k = 7)* |
| Weighted Average Prediction | 209.25 (*k = 5)* |

*Table 3. Results*

*Explanation of Access Point Swap*

The MAC IDs ending in *c0* and *cd* are located at the same position; we need to understand the impact each combination of runs has on our model performance. When comparing the swapped models, leaving out the *c0* MAC ID from the KNN model gave us a prediction error of **249.9** (k=5) as opposed to a prediction error of **273.7** (k=3) with MAC ID ending in *cd* left out. By keeping all seven MAC IDs in the weighted average KNN model, we achieve an even lower error prediction (**209.6** with a k=5).

Conclusion

*Drawbacks of methods employed and recommendations for improvements*

While the methods of this case study suggest a weighted-average approach for prediction, there are few drawbacks and challenges with this exercise.

**Wireless Technology: 802.11 Radio Signal**

The 802.11 radio signal tracking triangulates location using the signal strength from three or more access points. This measure of signal strength is called Received Signal Strength Indicator (RSSI), a technology that is imperfect and has some inaccuracies in the way it triangulates location. The RSSI signal will vary slightly when measuring from the same position. Typically, this is addressed by taking the average of repeated measurements. Note there are other environmental factors such as EMI interference, signal distortion, etc. that are inherent in the 802.11 standard, however, these are not statistically significant compared to the RSSI triangulation mechanism itself. Alternatively, there are other mechanisms for triangulating location such as angle of arrival (AOA), time of arrival (TOA), and time difference of arrival (TDOA). While they may provide more accurate performance, the supporting technology does have a higher cost.

**Measurement Inaccuracies**

With respect to the measurement variations, such as variation in the angles of measurement not being exactly 45 degrees, or the positional measurement not being exactly at a particular (x, y) coordinate, the only way this could be corrected would be some sort of automated mechanism, such as a GPS droid similar to what NASA has on Mars. However, implementing such an automated mechanism would be impractical for most use cases given the costs of such a technology far outweigh the potential benefit from a gain in accuracy.

**Loss Function is Sensitive to Outliers**

The method of detecting the nearest neighbor uses a point’s Euclidean distance, which was derived based on calculating the sum of squares. This is the same method used in other models such as ANOVA and regression, which also lack robustness to outliers. A single outlier could have a significant impact on the model. An outlier can be extreme in two ways. The first would be an extreme input value. A simple example to illustrate the point would be selecting an (x, y) position that is significantly farther away from the group of access points than the other measurements. The second would be an extreme output value. This could be the result of inaccurate signal strength due to a faulty access point, user error, signal distortion, etc.

A different approach that would be more robust to outliers would be to use a Manhattan-distance algorithm, which uses the absolute value rather than the sum of squares. Using a different modeling technique altogether such as hierarchical clustering would be an additional approach. The most effective solution, however, would be to intelligently filter the data with the help of subject matter experts such as an RF engineer, for example, that would provide the ranges for expected values and cut-offs for what is likely an erroneous measurement. We did this to a certain extent by cleaning out data and removing measurements that likely came from type 1 ad-hoc connection requests.

**Real-time Implementation of Model**

Real-time location tracking via the methods described in this case study are likely impractical since classification using KNN clustering depends on comparing values against the entire data set. A reduced model cannot be derived, i.e., the data is the model. Real-time implementation of the model would require an extremely low latency between data input and prediction output. As such, any location prediction would have to be run in a “batch” (i.e., asynchronous) fashion comparing new data points against the original, large, reference data set. A potential solution would employ a more costly active-scanning technology that is capable of real-time tracking, which would replace passive RSSI technology.

1. D. Nolan and D. T. Lang, "Predicting Location via Indoor Positioning Systems," in Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving, Boca Raton, FL, CRC Press, 2015, pp. 1-43. [↑](#footnote-ref-2)
2. T. King, S. Kopf, T. Haenselmann, C. Lubberger and W. Effelsberg, "COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses," in Proc. of the First ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization (WiNTECH), Los Angeles, CA, USA, 2006. [↑](#footnote-ref-3)
3. T. King, S. Kopf, T. Haenselmann, C. Lubberger and W. Effelsberg, "COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses," in Proc. of the First ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization (WiNTECH), Los Angeles, CA, USA, 2006. [↑](#footnote-ref-4)