Received Signal Strength Indicator based geolocation, a machine learning approach

Alexandre Bec, Samuel Cohen, Anthony Houdaille, and Maël Fabien

Télécom ParisTech, 2018

1 The training set

The aim of this study is to provide a location estimation using Received Signal Strength Indicator for Internet of Things (IoT) sensors. The aim is to allow a location of low consumption connected devices that use the Sigfox network. Typically, IoT sensors do not use GPS networks due to the high energy consumption this would imply. State of the art models are able to be precise to the nearest kilometer in urban areas, and around ten kilometers in less populated areas.

1.1 Building the feature matrix

One of the first task is to build a data set that can be exploited for machine learning applications.

In the first data set, each rows corresponds to one message received by one base station. Therefore, there are duplicated rows that could be concatenated using a one-hot encoder. We created one-hot encoders for all the initial features (RSSI, base stations latitude and longitude, "n-seq", "time ux") and also included additional distance measures, and a count of the number of base stations which received the message.

Finally, we scale the data set, group the rows by message ID, and split the data into a training and a test set. The final training data set is a 6068 x 1569 matrix.

1.2 Exploratory data analysis

On average, six base stations received each message. This gives us an idea of how many points we will be able to rely on for each message geolocation estimation.

On figure 1, we observe that some base stations (in red) are not lying within the same geographic area than the messages. This is problematic since those base stations have all received at least one signal. Outliers have been relocated at the barycenter of the observations.

2 Building the predictions

We need to predict both the latitude and the longitude of the geolocation of the message. This is a multi target regression case. We have 3 ways to approach this problem :

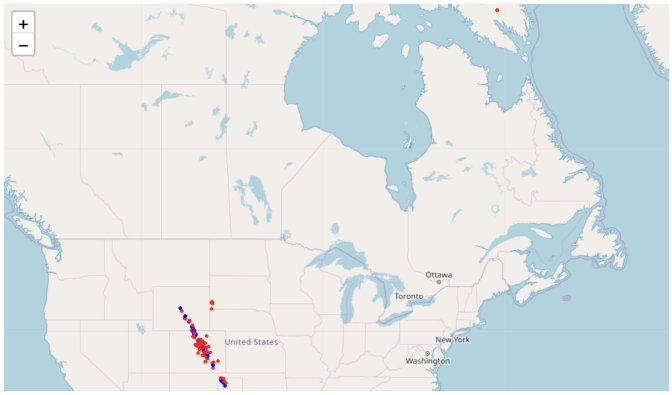

Fig. 1. Map of the location of the base stations (red) and messages (blue)

- Considering both the latitude and the longitude as independent
- Considering both the latitude and the longitude as dependent, and modeling the longitude first to explain the latitude
- Considering both the latitude and the longitude as dependent, and modeling the latitude first to explain the longitude

Visually, the independence hypothesis would hold if the targets (latitude and longitude) were more or less uncorrelated. However, in our case, we can observe a clear relationship. Predicting the longitude first improves the accuracy of the model by 3.5% on average. Graphically, this can be justified. The longitude is the biggest source of position variation in the data. Therefore, if we have an idea, more or less precise of the longitude of the base station, it is easier to predict the latitude.

The results of the different regressors tested are summarized in the Table 1.

Table 1. 80% percentile error achieved per model

Model	Error (m)
Extra Trees Regressor	2332
Light GBM	2698
Random Forest Regressor	2834
KNN Regressor	3056
Decision Tree Regressor	3347

The extra trees regressor is used to predict our final model.