Radio Beacon Base Context Identification

Jana Koteich, Nathalie Mitton

Radio Beacon Base Context Identification

Jana Koteich[1] et Nathalie Mitton[1]

[1]*Inria, 40 Avenue Halley, 59650 Villeneuve-d'Ascq, France*

L'étude de la mobilité humaine devient de plus en plus cruciale de nos jours dans les études sur les transports, la planification urbaine, les comportements de mobilité des foules et bien plus encore. Dans cet article, nous proposons une nouvelle approche pour étudier le type de mobilité en en construisant un modèle léger d'apprentissage machine (ML) basé sur l'observation des informations de réseau sans fil pour déterminer la situation réelle de l'appareil, qui peut être fixe (bureau, domicile, maison, etc), ou mobile (bus, piéton, voiture, etc). Le modèle est entraîné sur un ensemble de données réelles de balises WiFi et BLE collectées pendant environ 90 heures cumulées dans différents scénarios et conditions. Les résultats montrent que les algorithmes ML d'assemblage basés sur des arbres de décision, comme Random Forest, ont donné les meilleurs résultats, en termes de précision (94%) et de score f1. Nous pensons qu'une telle approche pourrait être utilisée pour étudier la mobilité humaine et constituer une étape importante vers le déploiement à grande échelle d'applications basées sur la mobilité en exploitant les téléphones mobiles de tous les jours.

Mots-clefs : Datasets, Machine Learning, Wireless Network, IoT, Mobility Model

1 Introduction

Understanding and modeling humans and device mobility have fundamental importance in mobile computing, with implications ranging from network design and location-aware technologies to urban infrastructure planning [Tri21]. So, inferring mobility states such as being stationary, walking, or driving is critical for several applications. The fact that nowadays, users carry several connected devices such as smartphones, laptops, and smartwatches equipped with radio communication technologies offering a different set of services resulting in different usage and mobility, provides new approaches for studying human mobility.

Several researchers have made the effort to study human mobility to determine people's fine-grained activities like using GPS positioning [VPBS+13], but GPS-based mobility characterization raises many issues such as spotty coverage and battery consumption. Other attempts and tools have been developed to predict global mobility [TVAA21] in general or only the next step [TAA18] of a trajectory. Such approaches are different in the sense that they mainly aim to trace contacts between devices and not necessarily their mobility. In our approach, we provide a novel technique to determine the real-life situation of a device. The novelty of such an approach is mainly characterized by its applicability, since nowadays almost all devices support WiFi and BLE, so no need for external hardware or module. Thus, by monitoring the behavior of WiFi and BLE in a device's range, we are able to infer a device's actual status.

In this paper, we investigate how wireless networks can be leveraged to infer a device's actual status by analyzing the behavior of wireless links within its range. Thus we propose a joint ML-based method. The B-model identifies if a device is stationary or mobile, and its output aids the M-model in determining the device's real-life situation. The models are trained using WiFi and BLE beacons jointly, and evaluated with a dataset of 90 hours collected under various conditions over a year. The results show that decision-tree-based ensemble ML algorithms like LGBMClassifier and XGBClassifier gave the best results, in terms of accuracy and f1 score for both models with an accuracy of 99% and 94% respectively, confirming the capability of determining mobility context from only WiFi and BLE data.

2 Data Collection

Data collection is considered as the foundation of the Machine Learning model building. The dataset is collected in different scenarios, with different variations using FiPy micro-controllers from pycom to

scan for wireless beacons. In [KM23], the dataset description is illustrated in details, but mainly we are concerned with the following features: 1) Node W (N_W) that scans for WiFi APs every two seconds, and 2) Node B (N_B) that scans for Bluetooth devices every second. The goal is to observe and record the variations of the wireless technologies in different mobility contexts, which are mainly categorized into two: *Static* and *Mobile* scenarios. For *Static* we define the following cases: Home, Office, Restaurant, Bus station, University and Meetings. For *Mobile*, we have the following scenarios: Pedestrian, Car, Bus, Metro, and Trains. The data is collected using diverse scenarios, including both rural and urban areas, to ensure that it was representative of a wide range of environments and populations, and was not biased towards any particular group or location. The data is collected and uploaded to github [†] with a timestamp and a label from where it was collected as Comma-separated values (CSV) files to be ready for the pre-processing stage.

3 Model Architecture

The aim is to extract knowledge from natural crowd mobility through radio beacons, by translating each "real-life" situation into a network model. As 'real-life' situation refers to the context of a network, which is the mobility of the environment of a device. Thus first we simplified the use-case by classifying the context into two main categories: *Static* and *Mobile*. To this end, to achieve the main goal which is building a ML model that can determine the real-life situation of a device, two models are defined. We started with the binary model called *B-model* that can determine whether the device is static or mobile. The output of this model helps improve the performance of the second main model which is called *M-model*, that stands for multi-class classification model, which aims to determine given the output of B-model, what is the exact status of a device inside a category among ten different possibilities: *i)* either between home, office, conference, restaurant or university; *ii)* or train, bus, car, metro or pedestrian.

3.1 Data Pre-processing and Feature Engineering

For each labeled dataset, we have a file that includes the data from WiFi and BLE at a certain time-slot. Each dataset is in the form of an n x 6 matrix, where $n > 0$ is a variable number that equals to the number of received beacons/probe-responses from all detected APs over the scanning time t. But feeding data into a model must be a column matrix (not an n x m matrix). So, we need to transform the n x 3 matrix into a 1 x $(f + 1)$ where f is the number of selected features from both WiFi and BLE and the *label* added at the end. Thus, in this case, each dataset will be represented by a single row. To transform the shape of the dataset from 2D to 1D, the dataset undergoes through two main phases:

— ϕ_1: Get the main statistics for each AP, thus as a result we will get an n x f_1 matrix, where n is the number of unique AP that appeared during scanning, and f_1 is the number of extracted features. Contact duration will be calculated for each AP, with the mean and standard deviation of the RSSI. As a result, we will end up with two dataframes that represent the statistics of the collected data for each wireless technology. So, the dataset still has the 2D form at this phase.

— ϕ_2: Transform each 2D dataset from ϕ_2 to a 1D. The dataframes will be transformed to a 1D vector by displaying the statistics of APs based on their contact duration (Δ_t), since the contact duration is an important metric for differentiating between scenarios. The features are mainly categorized into three main conditions: Long Δ_t, medium Δ_t, and short Δ_t based on the following criteria: First, get the percentage of contact duration as follows:

$$\%\Delta_t(m) = \frac{\Delta_t(m)}{t} \times 100 \quad \forall m \in M \tag{1}$$

Then, define three sets (L, M, S), where L is the set of AP that has $\Delta_t > 70\%$ of the total time t, M: Set of AP where $30\% < \Delta_t < 70\%$, and S: Set of AP that has a $\Delta_t < 30\%$ of the total time t. The mean and standard deviation of the RSSI of the access points that belongs to each set is calculated.

As a result, we end up with a vector (V_i) that includes 24 features (12 features extracted from WiFi and 12 extracted from BLE), plus the label of the scanned scenario i. The same process is done for all the datasets

†. https://github.com/Janakoteich/PILOT-Dataset-Collection-of-Multi-communication-Technologies

that are collected in different scenarios, thus we end up with a dataframe of V_n input vectors, where n is the total number of datasets. Now the dataset is ready for training.

4 Models Evaluation

The dataset is a labeled and collected over one year in different scenarios. The size of the dataset is 44,4 MB. The distribution of the datasets is unbalanced, to this end, with the available dataset for the moment, classical machine leaning algorithms will be suitable for our case since the dataset is small, and since it is a labeled dataset, supervised learning techniques will be applied to meet the final goal for constructing a classification model to guess the categorical label. In this section, several simulations are done to evaluate our model to select the one with the best performance.

4.1 B-Model: Binary Classification

The aim of this model is to determine whether the device is in a static or mobile context as defined in Section 3. The dataset is of 4764 input vectors. Each input is a result of one minute scanning in each scenario. The percentage of each label is as follows: 33.9% train, 17% Home, 10% office, 9% conference, 7.5% bus, 7.3% pedestrian, 7.3% university, 3.5% restaurant, 1.6% metro, and 2% car. 27 classification models are selected for training. From the 27 models [‡], we have selected the most known models to compare. Due to paper limitations table 1 displays only the three models that gave the highest accuracy.

As a result, LGBMClassifier and XGBClassifier gave an accuracy of 99%, and knowing that the trained dataset is not balanced, the *Balanced Accuracy* is calculated as it also gives a 99% accuracy, and same for F1 score which indicates a good performing classifier.

To justify the importance of using both information from WiFi and BLE jointly for such a model, we repeated the training using only WiFi data, then only BLE, and compared the results with the models trained on both technologies jointly. Figure 1 displays the accuracy value (from cross validation) for each trained model for the three scenarios. The green curve represents the values from models that are trained with only BLE input data, and the red curve for WiFi input data only, thus we can see that WiFi information gives better accuracy than BLE input data in almost all trained models. The violet curve represents WiFi and BLE trained data jointly, the results shows a higher accuracy from such data, thus we can conclude that both WiFi and BLE jointly gives better accuracy for estimating the output.

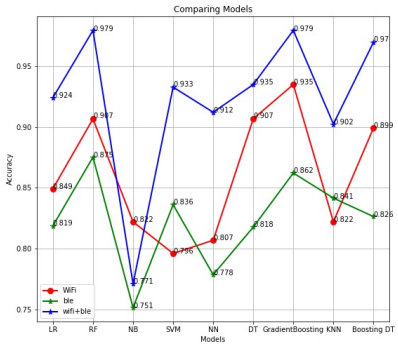

FIGURE 1: Accuracy comparison for three different input datasets

4.2 M-model: Multi-class Classification

The aim of this model is to determine a specified context of the network as mentioned in Section 3. We have tested 26 different machine learning classifiers to train the multi-class model. The dataset is of 2375 input vectors, each input is a result of two minutes scanning in each scenario.

First, to improve the accuracy, we added the results of B-model as an input feature to M-model to see how could the static/mobile information enhance the performance of the models. We have selected the best three models that gives the highest accuracy among the 26 tested models that are: LGBMClassifier, XGBClassifier and RandomForestClassifier, as shown in table 2. We will test our chosen models again by getting the accuracy after cross validation and calculating the Time consumed for training and prediction.

As shown in Table 2, XGBClassifier achieves the highest accuracy among the other selected models after cross validation, with the shortest prediction time, then it will be selected for hyper parameters tuning. After

‡. J. Koteich and N. Mitton, "Machine Learning Approach for Mobility Context Classification using Radio Beacons," in MASCOTS2023 IEEE, New York, United States, Oct. 2023.

hyper parameters tuning, the accuracy remains the same, leading that XGBoost classifier features the best accuracy with approximation to 94%.

TABLE 1: B-model models evaluation and comparison

Model	*Accuracy*	*Balanced*	*F1 Score*
LGBMClassifier	0.99	0.99	0.99
XGBClassifier	0.99	0.99	0.99
RandomForestClassifier	0.98	0.98	0.98

TABLE 2: Comparison between best three models for M-model

Model	*Accuracy*	*Training Time*	*Prediction Time*
XGBClassifier	93.62%	0.868	0.00436
LGBMClassifier	92.95%	0.907	0.00706
RandomForestClassifier	93.38%	0.433	0.01808

5 Conclusion and Future Work

In this paper, we have proposed a novel approach for inferring human mobility status by determining a device's status within its network context through WiFi and BLE, working in conjunction. Firstly, we trained a model to ascertain whether a device is in a mobile or static network context. Then a complementary model was trained, for detailed real-life context classification. These models were trained using real datasets collected over one year, for 90 hours, across diverse scenarios and conditions.The initial method achieved 94% accuracy in distinguishing ten scenarios with a simple machine learning algorithm (XGBClassifier). As part of our future work, we plan to investigate the addition of additional data already present in our datasets such as LoRa traces and accelerometer information to improve the accuracy of our classification. Then, we intend to leverage these results to design a dynamic and reliable opportunistic routing protocol.

References

[KM23] Jana Koteich and Nathalie Mitton. PILOT Dataset: A Collection of Multi-Communication Technologies in Different Mobility Contexts. In *CoRes*, 2023.

[TAA18] Douglas Teixeira, Mário Alvim, and Jussara Almeida. On the predictability of a user's next check-in using data from different social networks. In *Proc. of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*, 2018.

[Tri21] Amee Trivedi. *Human Mobility Monitoring using WiFi: Analysis, Modeling, and Applications*. PhD thesis, University of Massachusetts Amherst, USA, 2021.

[TVAA21] Douglas Teixeira, Aline Carneiro Viana, Jussara Marques Almeida, and Mário S Alvim. Revealing challenges in human mobility predictability. *ACM Transactions on Spatial Algorithms and Systems*, 2021.

[VPBS+13] Gonzalo M. Vazquez-Prokopec, Donal Bisanzio, Steven T. Stoddard, Valerie Paz-Soldan, Amy C. Morrison, John P. Elder, Jhon Ramirez-Paredes, Eric S. Halsey, Tadeusz J. Kochel, Thomas W. Scott, and Uriel Kitron. Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PLOS ONE*, 8.4:1–10, 2013.