A Machine Learning-Based Technique for the Classification of Indoor/Outdoor Cellular Network Clients

Kareem Abdullah[‡], Sara Attalla[⋆], Yasser Gadallah[⋆], Ayman Elezabi[⋆], Karim Seddik[⋆], Ayman Gaber[†] and Dina Samak[‡]

[‡]*Eventum IT Solutions*, [⋆]*The American University in Cairo*, [†]*Vodafone Egypt*

***Abstract*—In this paper, we propose a machine learning-based indoor/outdoor (IO) user classification algorithm in cellular systems as pertains to 3G networks. We consider different scenarios. The experimental results show that the best machine learning algorithm for IO classification is the boosting algorithm with an accuracy that reaches 88.9%.**

I. Introduction

Indoor users of mobile cellular networks constitute a large percentage of network clients [1]. Many such users suffer poor signal conditions due to obstacles and walls. Therefore, a reliable indoor-outdoor (IO) classification of users is needed on the network side to potentially target indoor users with specific actions in order to improve their user experience. Another benefit of IO classification is that indoor users are typically low-mobility users, which may be used in mobility pattern prediction by the network, a topic of current interest in its own right. Other use cases for IO classification include targeted content by the service providers to indoor users.

Another need for IO classification is that in several developed countries, cellular operators are just starting to perform "3G shutdown" [2], [3]. This operation is performed, hand-in-hand with the deployment of 4G in order to keep up with customer demand by re-farming the spectrum used by 3G networks to be used by 4G networks instead [4]. This will occur over many phases, and here stems the importance of selecting the right spectrum to allocate to 3G to serve the remaining 3G customers throughout these phases. This would basically depend on the nature of the 3G traffic in each cluster and whether it is generated from indoor or outdoor users thus guiding the right matching with the allocated band for the 3G technology. For all the above reasons, devising a technique for classifying 3G indoor and outdoor users is of utmost importance. We use TEMS measurement tools to collect data for 3G networks from a variety of environments and multiple mobile operators within several locations in Egypt.

The rest of this paper is organized as follows. Section II explains the data structure used in our IO classification. Section III describes the proposed method. Section IV presents the experimental results. Finally, Section V concludes the study.

II. Data Structure

We use a set of 3G measurements collected from three major mobile operators serving an entire country.

All measurements are collected from the mobile user point of view. The data we use is collected from multiple clusters spanning different cities all over the country. Several rigorous drive-test and walk-test surveys were performed in the designated cities/clusters to build the data used to address the problem at hand. Each of the data measurements is collected over a two-second interval through different hours of the day.

The used data do not contain sensitive or identifiable information about the operators' base stations. Measurements were captured during idle mode, voice and data sessions with different scenarios from multiple bands where 3G is allocated and from multiple carriers. In addition, in the connected mode, measurements were taken for all the serving cells in the active set (A1, A2,...) in soft handover regions, as well as the monitored cells (M1, M2,...) at the same collection intervals.

The data include around 237,000 records before cleaning, where each record represents the measurements of a certain handset including some cell identification features that provide important information about the serving and neighboring cells IDs as well as the carrier frequency and band. Using multiple feature selection algorithms, the most impacting features are: Received Signal code Power (RSCP), Signal-to-noise ratio (Ec/Io), Channel Quality Indicator (CQI), Block Error Rate (BLER), Modulation and Coding Scheme (MCS) and the frequency band.

III. Proposed Method

In this problem, we use a variety of classical, bagging and boosting machine learning algorithms to compare their accuracy score and select the best fitting algorithm that learns information about the given data to output a binary result whether the provided reading belongs to the indoor class (0) or the outdoor class (1).

In our model comparison, we use Decision Tree, Random Forest, Xgboost, Adaboost and KNN classifiers [5]. As we show later, boosting algorithms such as XGboost and bagging algorithms such as Random Forest show best performance.

One of the issues we consider is the unbalanced data of the indoor and outdoor data sets. We use two methods to overcome this issue. The first method uses stratified splitting for the train and test sets and applies balanced accuracy [6] as an evaluation metric. Another approach is to use oversampling techniques [7] to balance the training set and evaluate the common accuracy on a balanced non-oversampled testing set.

IV. Experimental Results

We conduct some experiments that illustrate the ability of the selected techniques to differentiate between indoor and outdoor mobile users. We evaluate our tests on 30% of the available data. We implemented our technique using Python and we used scikit-learn [5] in implementing the classical machine learning approaches. We have four scenarios of data collections.

A. Data Scenarios

1) First Scenario : We use data collected from one venue through a walk test inside and outside a building. We captured 224 indoor readings and 60 outdoor readings. We used the following KPIs in our classification:
(RSCP: A1, Ec/Io: A1, A2 & M1, DL frequency band: A1, A2 & M1, MCS 1, 2 & 3, CQI-Mean 1 & 2, and BLER)
Simple classification models such as Decision Trees and Random Forest Classifiers can capture the difference and achieve 100% accuracy.

2) Second Scenario: Outdoor data is collected from a drive test covering most of the city, whereas indoor data was collected from selected venues where we know do not have a good 3G mobile coverage. The data belong to a single operator covering one cluster of a major city. We conduct our experiment using only Ec/Io and RSCP of the first 2 active cells (A1 and A2) and drop the rest of the features (due to the lack of some KPIs from those cells). The accuracy in this case reached a balanced accuracy of 98.4% as shown in Figure. 2.

3) Third Scenario: The high accuracy observed from the second scenario is due to the biased data of the indoor venues that have bad 3G connectivity coverage. To overcome the problem of biased data, we collected more data from different places such as highways, rural areas and industrial areas to ensure the inclusion of all data coverage cases. Due to the large data size here, we choose 3 cells to conduct our experiment, namely, A1, M1 and M2 using the EC/IO and RSCP KPIs of each cell. The size of the collected data is 173,754 indoor records and 63,981 outdoor records reaching 70563 indoor records and 26603 outdoor records after cleaning. The results show an expected drop in the balanced accuracy to reach a maximum of about 76.58% as shown in Figure. 2.

4) Fourth Scenario: In this scenario we use the same data set that is used in the third scenario and add CQI - mean and the top 3 modulation schemes for the first active cell.These criteria lead to data size reduction, reaching 32367 indoor records and 11025 outdoor records after cleaning. The balanced accuracy increased to 86.9% as shown in Figure 2. Figure 1 shows the confusion matrix of predicted and actual test data that are predicted by the best model.

B. Over-sampling

As explained in Section III, we tried to solve the unbalanced data issue. One solution to this is to use oversampling techniques. The data size in this technique is balanced to reach 25,965 records for both indoor and outdoor data sets. Using Random Over Sampling, we enhanced the accuracy of the fourth scenario to 88.9%.

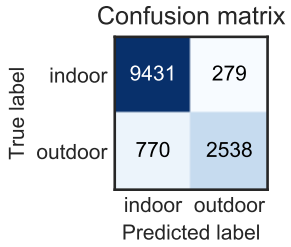

Figure 1. Fourth Scenario's Confusion Matrix

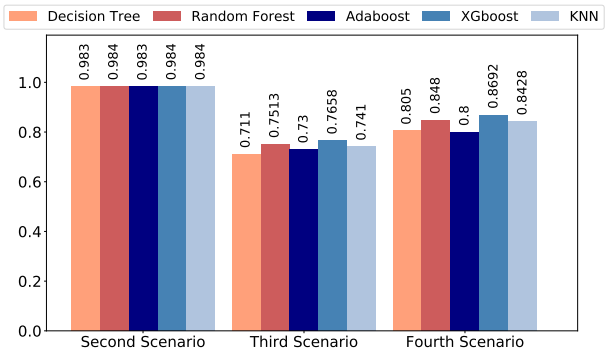

Figure 2. Machine learning Algorithms balanced accuracy results

V. Conclusions

In this paper, we proposed an indoor/outdoor classification algorithm for 3G networks. Using the TEMS technology, KPIs are collected from multiple clusters under different conditions during the day. We applied feature selection to choose the best features to use. We then tuned our model to select the best machine learning algorithm with the best model hyper-parameters. Results show that boosting algorithms such as XGboost showed best performance with an accuracy of 88.9% under many scenarios in different environments.

References

[1] A. Landström, "Indoor mobile classification and coverage analysis," M.S. thesis, Luleå University of Technology, 2009.

[2] Alan Weissberger, "AT&T to shut down 3G network in 2022; Verizon at end of 2019," IEEE ComSoc Tech Blog, 2019, http://techblog.comsoc.org/2019/02/22/att-to-shut-down-3g-network-in-2022-verizon-at-end-of-2019/.

[3] Alex Choros, "Australian 3G Network Shutdown: Everything you need to know," Whistle Out, 2019, https://www.whistleout.com.au/MobilePhones/Guides/Australian-3G-network-shutdown-what-you-need-to-know.

[4] Shiying Han, Ying-Chang Liang, and Boon-Hee Soong, "Spectrum refarming: A new paradigm of spectrum sharing for cellular networks," *IEEE transactions on communications*, 2015.

[5] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[6] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," pp. 3121–3124, Aug 2010.

[7] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.