IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 2, 2019, accepted April 27, 2019, date of publication May 2, 2019, date of current version May 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914451

An Ensemble Learning Scheme for Indoor-Outdoor Classification Based on KPIs of LTE Network

LEI ZHANG[1,2], QIN NI[3], MENGLIN ZHAI[1,2], JUAN MORENO[4], AND CESAR BRISO[4]

[1]College of Information Science and Technology, Donghua University, Shanghai 201620, China
[2]Engineering Research Center of Digitized Textile and Apparel Technology, Ministry of Education, Shanghai 201620, China
[3]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China
[4]School of Telecommunications Systems and Engineering, Technical University of Madrid, 28031 Madrid, Spain

Corresponding author: Menglin Zhai (mlzhai@dhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61801107, in part by the Shanghai Sailing Program under Grant 19YF1436800 and Grant 17YF1400900, and in part by the Fundamental Research Funds for the Central Universities under Grant 2232019D3-55 and Grant 17D110410.

ABSTRACT Wireless Big Data has aroused extensive attention, as mass mobile devices have been developed and deployed for the upcoming 5G era. The context information of these devices is of importance for personalized services in a smart environment. Nevertheless, the constant change of scenes challenges to the network operator. In this paper, we propose an ensemble learning scheme for indoor-outdoor classification for a typical urban area, based on the cellular data captured in a commercial LTE network. The variables are extracted by network key performance indicators (KPIs) and radio propagation knowledge. Based on these main variables, the decision trees grow and split by the Gini index of sampled features. Then, all decision trees are assembled as weak learners to build the ensemble scheme, thus improving the discrimination ability. The self-validation results show the ensemble model achieves extreme accurate (with an out-of-bag error lower than 1%) classification for indoor and outdoor environments. Moreover, the prominent variables are selected based on the variable importance of in the initial training. The reconfigured model based on fewer variables and less weak learners also gains the highest accuracy and relative short compute time, compared with other classical machine learning methods.

INDEX TERMS Propagation measurement, ensemble learning, LTE, channel model, scene classification.

I. INTRODUCTION

Trends like the Internet of Things(IoT), big data and real-time process are poised to impact the current information society a variety of ways. In the meantime, the next generation (5G) cellular network is expected to offer upgraded services with diverse requirements. On the one hand, updated communication infrastructures such as small cell base stations are assumed to be densely deployed to improve radio coverage and network capacity [1]. On the other hand, to support the low-latency services like Internet of vehicles (IoV), catching and computing resources are extended from cloud to the local or edge of network architecture, refers to Fog/Edge computing [2]. Moreover, with the boom in personalized services, service-oriented recognition and prediction evolve to basic requirements to improve user experience. To address these challenges, next-generation wireless networks must be more intelligent to provide user-driven solutions. Machine learning based big data analytics unearth the in-depth knowledge regarding the network, then gradually transforming it from automated to autonomous, and fully realizing network potential for the enjoyable user experience and excellent network performance [3].

Context awareness is the foundation towards more advanced techniques aiming at personalization and intelligence of service provisioning in the wireless big data era [4]. The context information in general and location information, in particular, can aid in addressing several of the key challenges in 5G, complementary to traditional and disruptive technological developments [5]. By understanding human behaviors and related contextual information, the operators can achieve much higher efficiency in resource management and larger economic benefit [6]. Large-scale scene understanding is one of the basic computer vision problems. It finds applications in robotic navigation, image/video indexing, archiving and retrieval, etc. [7]. For location-based services (LBS) and real-time location systems (RLS), an attractive topic is a seamless coverage from outdoor to indoor. By combining behavior track with

The associate editor coordinating the review of this manuscript and approving it for publication was Ke Guan.

activity recognition technology, the high-level functional assistant services can be provided in the smart environment [9]. Therefore, the indoor-outdoor(IO) classification is a fundamental issue for various smart applications and services.

A. RELATED WORK

To achieve an accurate classification between indoor and outdoor environments, a lot of scholars have tried different methods on various sensor based datasets. The most common method relies on Global Navigation Satellite Systems (GNSS) such as the Global Positioning System (GPS) is usually blocked by the building structure in the indoor environment. In [8], Lee *et al.* collect GPS alongside temperature data in different indoor locations during three seasons. Results show that IO classified by GPS alone obtained about 73% agreements. Nevertheless, No matter the standalone GPS strength or the signal-to-noise ratio (SNR) of the GPS signal are unreliable on IO classification, considering the time delay and the energy hungry [10] [11], as well as the possible deviations in semi-indoor environments. Besides the temperature sensor, other lightweight sensors are implemented to obtain additional features for indoor positioning and pedestrian navigation, e.g., illuminance [12], accelerometer [13], magnetic variance [14], or multisensor fusion [15] methodologies. Due to the user privacy and data catching& processing complexity, the deployment of these sensors-based approaches are limited. The characteristics of radio signal propagation are of importance in designing a positioning system, While in free space a simple inverse-square law can be used to determine the signal strength as a function of range [16]. A series of research focuses on radio pattern on the indoor-outdoor channel, e.g., outdoor-to-indoor path loss model [17], channel characterization for LTE small cells [18], penetration loss onboard high-speed train [19]. These studies reveal the obvious differences between outdoor and indoor scenarios, from the radio aspect. A current work named IODetector observes not only the variance of light intensity and magnetic field but also introduces the cell radio pattern [20]. Moreover, SenseIO system [21] consist of four sensing modules (serving cell tower, Wi-Fi based, activity recognition, and light intensity) based on the on-board lightweight sensors of smartphone states an IO detection accuracy above 92%. Yet the hard-coded thresholds of these systems may hurt the accuracy/efficiency of IO detections across different environments [22]. Thus, machine learning based approaches are proposed to adaptively sense the contexts [24] [25]. Authors in [26]and [27] use machine learning to generate the binary indoor/outdoor classifiers rely on the signal strength of GSM cellular network and Wi-Fi RSSI, respectively.

B. MAIN CONTRIBUTIONS

Motivated by the above observations, we propose an ensemble learning scheme for IO classification based on empirical cellular data. Compared to conventional work, our main contributions are as follows:

- The real-world measurement data are captured in the present 4G Long Term Evolution(LTE) cellular network covers typical urban areas. The measurement report(MR) are matched and converted with the cell data to provide the key performance indicators (KPIs).
- Various variables are extracted from the KPIs, based on the radio propagation knowledge. Then an ensemble learning scheme is proposed to classify the MRs from UE(User Equipment) side with extremely high accuracy($> 99\%$).
- The out of bag (OOB) error of the initial ensemble model is calculated along with the number of weak learners. Also, the variable importance is evaluated to select predominant features. The ensemble scheme is reconfigured and simplified by the trade-off of OOB errors and the number of weak learners, as well as the predominant features.
- The ensemble approaches are self-compared and cross-compared with classical machine learning algorithms, which proves the proposed methods can efficiently classify the indoor-outdoor.

The rest of this paper is organized as follows: Section II describes the measurement campaign and captured data structure; Section III explains the ensemble scheme design with main processes; Section IV shows system modeling, evaluation, and reconfiguration based on the detailed analysis; Then the reconfigured model are compared with the initial model and other machine learning algorithms in section V; Section VI draws the conclusion and future works.

II. MEASUREMENT & DATA

In LTE evolved universal terrestrial radio access network (E-UTRAN), eNBs (evolved NodeBs) as base stations are deployed to manage radio resource and mobility in the cell and sector to optimize all the UE's communication in a flat radio network structure (see Fig. 1). Therefore, the performance of an LTE eNB depends on its radio resource management algorithm and its implementation. Different eNBs

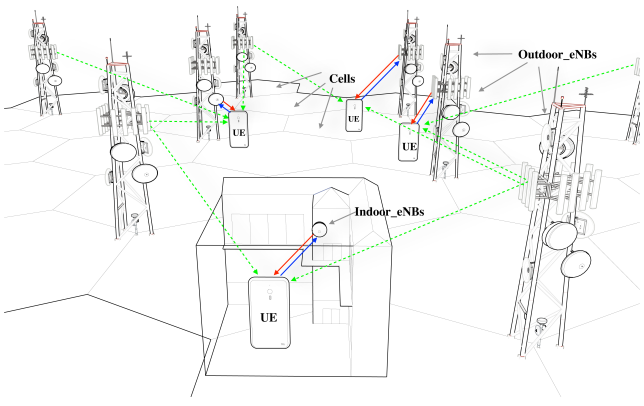

FIGURE 1. The sketch of LTE E-UTRAN deployment and serving for UEs.

TABLE 1. Measurement configurations.

Item	Description
Cellular network	FDD-LTE
Frequency band	0.9/2.1 GHz
RSTP	40 dBm
Hearable Serving cell	≤ 1
Hearable Neighboring cell	≤ 6
Handset device	Samsung S5 G900I
Platform	Android v4.1
Software	TEMS Application v16
Timing Advance	Available
GPS	Available

RSTP: Reference Signal Transmit Power

have a different range, i.e. a limited size in which a receiver can successfully hear the transmitter. In Fig. 1, we present the track of a UE moving among the cells. Due to the eNB's limited coverage, the serving cell of UE has been changed at a different position (handover), as well as the heard neighboring cells. Therefore, the radio patterns are very different in multiple dimensions, which results in different KPIs in different scenes. This physical phenomenon motivates the measurement campaign for further context classification.

A. MEASUREMENT CAMPAIGN

The cellular Data are collected by the field test with a test mobile system in a typical urban environment, as shown in Fig. 2. Specifically, the indoor measurements are conducted in stillness at five shopping malls as indoor scenarios, In an outdoor environment, a large amount of MR samples are recorded with GPS coordinates in outdoor cases distributed over five square kilometers.

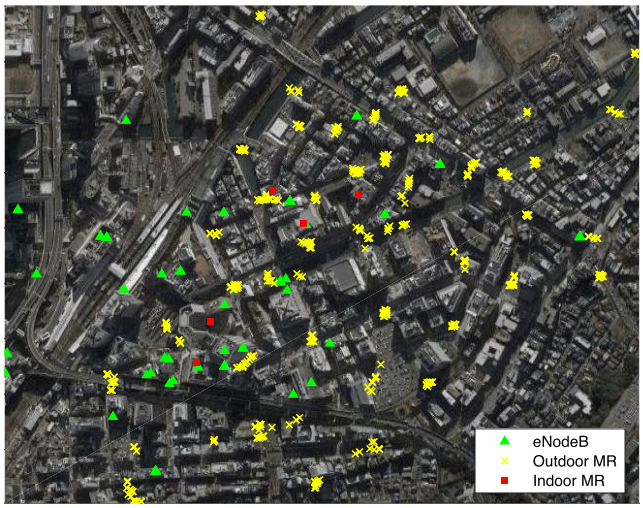

FIGURE 2. Birdview of the measurement campaign in a typical urban environment.

In the measurement process, the handheld mobile devices(Samsung S5 G900I smartphone with TEMS application installed) are continually apply data services in FDD-LTE (Frequency Division Duplexing) network, capture the MRs include serving and neighboring cells information and record their physical indexes by the intra-frequency measurements within E-UTRAN. Note the moving speed of field test is restricted with an up-limit(30 km/h) to avoid the distortion of the radio measurement due to high speed as well as to secure the space between consecutive measurements.

B. DATA DESCRIPTION

After cleaning and converting, 24912 data samples are filtered out from the raw data. Each MR sample includes the following informations [23]:

- *Time Stamp*: The time at which an intra-frequency measurement report event is recorded by TEMS, with the precision of a second.
- *Timing Advance (TA)*: TA is used to control the uplink transmission timing of individual UE. In the case of FDD, the uplink radio frame structure is the same as the downlink radio frame. Thus, the distance between a UE and the network antenna can be approximated by the propagation speed and TA.
- *Latitude & Longitude*: The real-time latitude and longitude geographic coordinates recorded by the GPS system on UE. In most cases of indoor, the GPS signal is shielded by the roofs or walls, so the latitude and longitude would be missed in the MR.
- *E-UTRAN cell identifier (ECID)*: The uniqueness of a cell is marked by ECID. To ensure uniqueness, the E-UTRA absolute radio frequency number(EARFCN) and Physical layer cell identifier (PCI) recorded in MR are combined to get a unique ECID for each cell.
- *Reference Signal Received Power (RSRP)*: The RSRP denotes the linear average received power on the resource elements that carry cell-specific reference signals (CRSs). The reference point for the RSRP shall be the antenna connector of the UE. Moreover, the interference and noise components on the downlink signal are included.
- *Reference Signal Received Quality (RSRQ)*: The RSRQ is calculated by the ratio of RSRP to the total received power including interference from all sources and thermal noise (received signal strength indicator, RSSI). Due to the consideration of interference and noise in RSRQ measurement, a UE may experience different received signal qualities at different locations.

Combining the above indexes, each MR is constructed as Fig. 3 shows. The maximum hearable neighboring cells are six, and the missed values (not heard by UE) are filled with NaN.

III. ENSEMBLE SCHEME DESIGN

A. ALGORITHM DESIGN

The decision tree is a classical supervised machine learning algorithm that commonly used in operations research, specifically in decision analysis. In such a tree-like graph, the condition is represented as the "leaf" (node) and the decision as

Time Stamp	Timing Advance	Latitude	Longitude
Serv_ECID	Serv_RSRP	Serv_RSRQ	
1st_Neig_ECID	1st_Neig_RSRP	1st_Neig_RSRQ	
2nd_Neig_ECID	2nd_Neig_RSRP	2nd_Neig_RSRQ	
⋮			
6th_Neig_ECID	6th_Neig_RSRP	6th_Neig_RSRQ	

FIGURE 3. Data structure of each measurement report.

"branches" (edges). This splitting process continues until no further gain can be made or a preset rule is met, e.g. the maximum depth of the tree is reached. Commonly used decision trees include ID3, C4.5, and Classification & Regression Tree (CART). For classification, the CART constructs a binary tree based on a numerical splitting criterion recursively applied to the data, thus it is generally superior to other decision trees, in terms of classification capabilities [28]. Gini Index is the commonly used metric for the impurity measure to build CART, which is given by:

$$G = \sum_{k=1}^{K} P_{mk}\,(1 - P_{mk}) = 1 - \sum_{k=1}^{K} P_{mk}^2 \tag{1}$$

where P_{mk} represents the proportion of observations in the $m - th$ region from the $k - th$ class in a total of K classes. In essence, the Gini index gives a measure of how good a split is by how mixed the classes are in the groups created by the split. The higher the variance, the more misclassification there is. Therefore, the lower values of the Gini Index yield better classification. It is calculated for every row and split the data accordingly in the tree to determine the best way to reduce error. Unfortunately, this greedy method leads to over-fitting and model over generalization. To combat this problem, the ensemble meta-algorithm can be used to offer a more robust result. It combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Therefore, it is less prone to overfitting. This study introduces an ensemble scheme called random forest (RF) based on "bootstrap aggregation (bagging)" idea. RF constructs multiple individual decision trees at the training process. Predictions from all trees are pooled to make the final prediction based on the majority vote. The complete ensemble scheme based on cellular big data for IO classification is described in Alg. 1. Furthermore, the main procedure steps are presented detailedly in the following subsections.

B. FEATURE EXTRACTION

Feature engineering is fundamental to build an intelligent system. Specifically, feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Therefore, it is difficult and expensive.

Algorithm 1 Ensemble Scheme for Indoor Detection

PRE-PROCESS

- Data clearing & converting: table lookups from the eNB database to match the MR and serving/neighboring cells.
- Data labeling: feature extractions based on the feature engineering rules from the cellular big data.
- Each training sample is formed by M features, and n training samples are prepared in total.

INPUT:

- Set feature dimension as $\mathbf{M} = \{x_1, ..., x_M\}$;
- Set training data $\mathbf{S} = \{s_1; s_2; ...s_n\}$ with correct labels $\Omega = \{\omega_{indoor}, \omega_{outdoor}\}$ representing 2 classes {indoor, outdoor};
- Set B specifying the number of trees in the RF.

PROCEDURE:

for $b = 1, ..., B$

do

- Create a bootstrapped sample $\{s_b, \omega_b\}$, by randomly row sampling $\{\mathbf{S}, \Omega\}$ with replacement;
- Collect the ignored samples $\{s'_b, \omega'_b\}$from the bootstrap operation as OOB (out of bag) data;
- Pre-generate a CART based on the bootstrapped sample $\{s_b, \omega_b\}$;
- On each node of the CART, randomly select m features from all M features ($m \ll M$). Then select an optimal segmentation feature as the best split point to set the node by Gini Index;
- Let each tree grows to the maximum extent without any pruning;
- Receive the hypothesis(classifier) h_b;
- Add h_b to the ensemble, ϵ;
- Set an acceptable test error threshold η;
- Set an acceptable compute time consuming T.

endfor

OUTPUT: A composite model, random forest $\epsilon = \{h_1, ..., h_B\}$, and an internal test data set $\mathbf{O} = \{\mathbf{S}', \Omega'\}$.

TEST:

- Give unlabeled OOB samples $\mathbf{S}'$;
- Evaluate the ensemble ϵ on $\mathbf{S}'$;
- Let $v_b = \begin{cases} 1 & if\ h_b\ picks\ class\ \omega_{indoor} \\ 0 & otherwise \end{cases}$ be the vote of each tree for indoor detection;
- Obtain the total vote and made the final classification by the majority vote: $C_{RF}(\mathbf{S}') = Majority\ Vote\{v_b\}_{b=1}^{B}$;
- Computes the misclassification probability (OOB error e) of OOB samples, compared with Ω';
- Record the compute time t.

repeat

Feature selection based on the variable importance;
Reconfigure B specifying number of trees in the ensemble.

until $e < \eta$ & $t < T$

In this study, the expert domain knowledge of the radio propagation is crucial for feature extraction. Generally, in FDD-LTE cellular system, radio propagation follows some solid rules: 1) UE is generally served by the eNB provide the strongest power; 2) up to 6 neighboring cells can be heard (beyond the threshold) and sorted by received power strength on the UE side; 3) the building structures lead to penetration on the radio propagation link from outdoor to indoor, and vice versa; 4) a classical large-scale channel model can estimate the path loss and the corresponding received power, but the fluctuations caused by the multipath effect cannot be neglected. In summary, the RSRP and RSRQ values of both *Serving cell* and *Neighboring cells* are directive features to judge the scenarios of UE, e.g, outdoor UE can be served by the outdoor eNB with high RSRP value and good RSRQ. Besides, the UE in the outdoor environment has a strong possibility to be served by the outdoor eNB and hear more neighboring cells. Also, the RSRP/RSRQ differences between *Serving cell* and strongest *Neighboring cell* are two useful features to detect if the UE is located in the edge of the cell. Last but not least, the estimated RSRP by standardized channel model can be implemented to compare with the measured value as a critical feature. Taken together, we extract twenty variables from the cellular big data as follows:

- **Variables 1-7:** setting as the RSRP values of one *Serving cell* and six *Neighboring cells* as the first 7 features.
- **Variables 8-14:** setting as the RSRQ values of one *Serving cell* and six *Neighboring cells* as the following 7 features.
- **Variable 15:** setting as the class of serving eNB: the indoor class *Serving cell* is denoted by 1 and outdoor class *Serving cell* is denoted by 0;
- **Variable 16:** setting as the number of all sensed *Neighboring cells*;
- **Variable 17:** setting as the number of indoor *Neighboring cells* in the MR;
- **Variable 18:** setting as the RSRP difference between *Serving cell* and *Neighboring cell* with strongest RSRP;
- **Variable 19:** setting as the RSRQ difference between *Serving cell* and *Neighboring cell* with highest RSRQ;
- **Variable 20:** setting as the difference between measured RSRP of *Serving cell* and estimated RSRP based on channel model. since widely used Cost 231 Hata Channel Model extends Hata's model to 2GHz for urban area, we use Cost 231 Hata Model to estimate the path loss (PL) as

$$PL = 46.3 + 33.9\log(f) - 13.82\log(h_B) - a(h_R, f) + [44.9 - 6.55\log(h_B)]\log(d) + C \quad (2)$$

where $f, h_B, h_R, d, and\ C$ denote the carrier frequency, base station and mobile station antenna effective heights, link distance, and correct factor(= 3 for metropolitan areas), respectively. In our study, $h_B and h_R$ are ignored, due to the difficulty in obtaining for each eNBs and MR. Moreover, d is hard to calculate, especially for the indoor case without GPS data. Thus, we roughly estimate the link distance by TA from the MR and propagation speed. As defined in 3GPP 36.213 [29], $1\ TA = 16 * Ts = 16 * (0.5ms/15360) = 0.52\mu s$, so 1 TA maps to the distance between eNB and UE is $300,000km/s * 0.52\mu s = 78m$. Eventually, the difference of estimated and measured RSRP can be roughly given by

$$\triangle RSRP = RSTP - (46.3 + 33.9\log(f) + 44.9\log(78TA) + 3) - RSRP \quad (3)$$

C. SELF-VALIDATION

Something has to modify, be attention on the gini index and gini impurity

In the ensemble model, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run. As shown in Alg. 1, about one-third of the samples are left out of the bootstrap sample and not used in the construction of the b-th tree. These ignored samples in the training process group to a testing dataset, which is called OOB data. We use OOB error to define the misclassification probability on testing dataset$\{s'_b, \omega'_b\}$, by using the ensemble model trained by the samples that did not have s'_b in their bootstrap process.

The study of OOB error estimates for bagged classifiers in Breiman [30] gives empirical evidence to prove that the OOB estimate is as accurate as using a test set of the same size as the training set. Therefore, using the OOB error estimate can avoid the need for an independent cross-validation dataset.

D. FEATURE SELECTION

Feature selection is also called variable selection or attribute selection. It can be used to identify and remove unneeded, irrelevant, and redundant variables from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. The selected variables reduce the complexity of the ensemble model and find a trade-off between efficiency and accuracy. The features are selected based on variable importance(VI) in this study. The VI is measured as follows: 1.

1) In every tree grown in the ensemble model put down the OOB cases and count the number of votes cast for the correct class as R_{oob};
2) For the $b-th$ predictor variable, randomly permute the values of variable m in the OOB cases and put these cases down the tree and count the correct predictions as R_{perm};
3) Subtract the number of votes for the correct class in the variable-m-permuted OOB data from the number of votes for the correct class in the untouched OOB data. The average of this number over all trees in the forest is the raw importance score(RIS) for variable m,

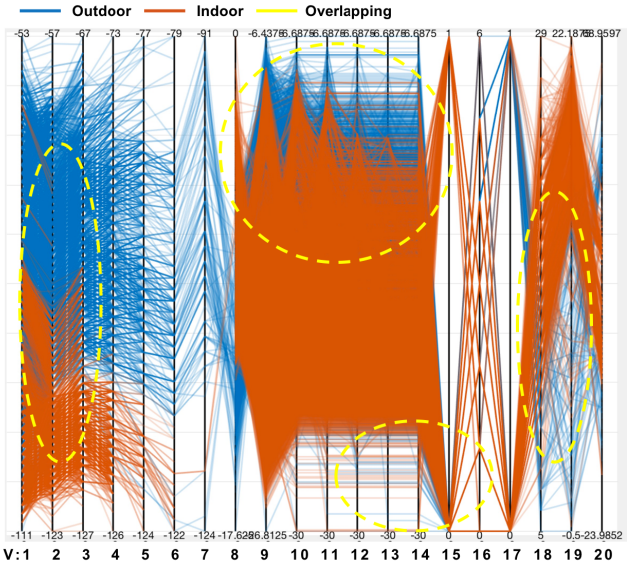

FIGURE 4. Parallel coordinates plot of the input data.

given by:

$$RIS = \frac{\sum_{b=1}^{B} R_{oob} - R_{perm}}{B} \quad (4)$$

If the RIS from tree to tree are independent, then the standard error can be computed by a standard computation. The correlations of these scores between trees have been computed for a number of data sets and proved to be quite low, therefore we compute standard errors in a classical way, divide the raw score by its standard error to get a z-score, and assign a significance level to the z-score assuming normality. When the number of variables is very large, ensemble modeling can filter out the most important variables based on the RIS, to achieve an efficient and precise prediction.

IV. MODELING & EVALUATION

The basic idea of Machine Learning is to study pattern recognition, make predictions, improve predictions based on examples or data. A classification modeling task usually involves training and test sets which consist of data samples. Each sample in the training set contains one target value (class label) and several variables (features). The goal of an ensemble learner is to produce an ensemble model able to predict target values of data samples in the testing set, for which only the variables are known. Without loss of generality, the IO classification problem can be viewed as a binary problem in which one's objective is to separate the two classes by a function induced from available examples. The goal is to produce a classifier that generalizes well, i.e. that works well on unseen/unlabeled examples (OOB data).

A. TRAINING

As the pre-process of the training, we clean and convert the cellular big data as described in Section II-B. Then, the features are extracted to variables based on the radio propagation knowledge in Section III-B. To better understand, the multivariate data are visualized by the parallel coordinated plot. As shown in Fig. 4, each variable is normalized on its own axis and all the axes are placed in parallel to each other. Overall, the normalized indoor variables and outdoor variables have large overlapping regions, which is difficult to classify it directly. In this case, we can use machine learning algorithms to do the classification. The training process is configured as presented in Table. 2. 40 CARTs are applied to the training data with 11676*21 size. The bootstrapped input data with a known label at a time is prepared through each decision tree and split by Gini Index without pruning. Fig. 5 gives the first CART as an example to show the growth of each tree. Eventually, all bootstrapped input data train an ensemble model with 40 decision trees for further prediction.

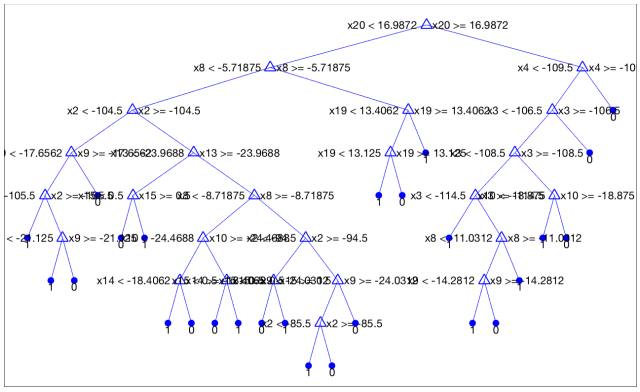

FIGURE 5. The first CART in ensemble.

TABLE 2. Training configurations.

Item	Description
Algorithm	Ensemble trees
Method	Classification
No. of Trees	40
Data size	11676 x 21
No. of variables	20
No. of labels	2
Platform	Matlab R2017b
Labeling period	5.6 s
Training period	8.28 s
CPU	2.4 GHz Intel Core i5
Memory	4 GB 1600 MHz DDR3
System	macOS High Sierra 10.13.6

B. PREDICTION

As discussed in Sec. III-C, about 37% of samples are left out from the bootstrap in the training step. These left-out examples can be used to form an accurate estimate of important quantities, and then evaluate the ensemble model independently. Look at the confusion matrix in Fig. 6, which summarize how many OOB samples are classified to the true class. On the principal diagonal we can see the cases which are predicted well from the ensemble. Moreover, the confusion

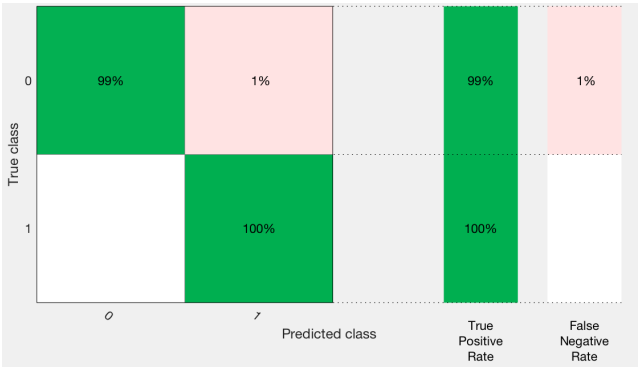

FIGURE 6. Confusion matrix of the prediction.

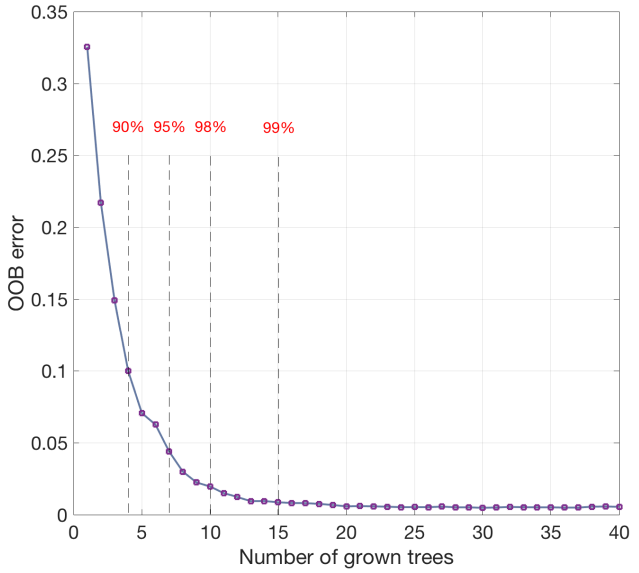

FIGURE 7. Out of bag error vs. number of grown trees.

matrix shows the false negative rate is as low as 1%. Indeed, it corresponds to the OOB error. This is equivalent to saying that the accuracy of our ensemble model is 99%. The OOB error is complementary to the accuracy, and it's calculated as the ratio and presented along with the number of grown trees in Fig 7. The OOB error decreases distinctly with the number of growing trees.

C. ANALYSIS & RECONFIGURATION

In general, the ensemble learning scheme constructed by more weak learners can obtain better predictive performance. As Fig 7 shows, the ensemble model have more than 4 CARTs can obtain the OOB error lower than 10%. As the number of trees increases to 15, the OOB error is stabilized around 1%. Therefore, an ensemble scheme consists of 15 CARTs is rational to take in full consideration on the precision and time-consuming.

The variable importance in classification prediction is another key aspect to optimize the ensemble model. As shown in Fig. 8, the variables extract based on the radio propagation knowledge contribute to the classification task at different levels. Note the contribution of the 7th variable is almost nothing, due to the missing data or undetectable RSRP of the 6th neighboring cell in MRs. Also, this is the reason for the gradual decline of contributions from the 2nd to the 6th variables. Moreover, there is very few indoor cells are deployed in the current 4G network, compared with outdoor cells, which causes the contribution of the 17th variable is very limited. Oppositely, the serving cell's RSRP& RSRQ& cell class (variables 1,8,15), the 1st neighboring cell's RSRP(variable 2), as well as the difference between measured and estimated receiving power strength (variable 20), are the most important features to finish the classification tasks. In this case, these four variables are chosen to be the prominent variables for feature selection.

Through the analysis of the above two facets, we can reconfigure the training process to obtain a more efficient ensemble model, i.e., extracting 4 prominent variables from the training data to generate an ensemble learner consist of 15 weak learners.

TABLE 3. Performance comparison.

Classifier	Accuracy	TP (I/O)	FN (I/O)	Time
Initial Config	20 variables & 40 trees			
Single Tree	97.6%	98%/98%	2%/2%	1.8s
Bagged Trees	99.5%	99%/99%	1%/1%	19.5s
Reconfig	4 variables & 15 trees			
Bagged Trees	99.3%	99%/99%	1%/1%	4.8s
Single Tree	97.7%	98%/98%	2%/2%	0.8s
Boosted Trees	98.4%	97%/99%	3%/1%	5.3s
SVM	97.6%	98%/97%	2%/3%	33.8s
KNN	98.1%	98%/98%	2%/2%	1.6s
W-KNN	98.1%	98%/99%	2%/1%	1.5s
LR	97.6%	98%/97%	2%/3%	3.8s

V. COMPARISON

To evaluate the performance, the comparisons are conducted in this section in two aspects:

- Self-comparison: the reconfigured ensemble model is compared with the initial model based on more variables and weak learners.
- Cross-comparison: the reconfigured ensemble model is compared with models trained by other popular machine learning algorithms with five-fold cross-validation, include: Single decision tree, boosted trees, linear support vector machine (SVM), k-nearest neighbors(KNN), weighted KNN, logistic regression (LR).

Table. 3 gives the comparisons of accuracy, true positive rate (TP), false negative (FN), and compute time-consuming between different configurations and various machine learning algorithms. On the one hand, through the self-comparisons of tree-based classifications, the four prominent variables with fewer trees can achieve very accurate IO classifications, which is very close to the performance obtained by more variables and weak learners. Meanwhile, it

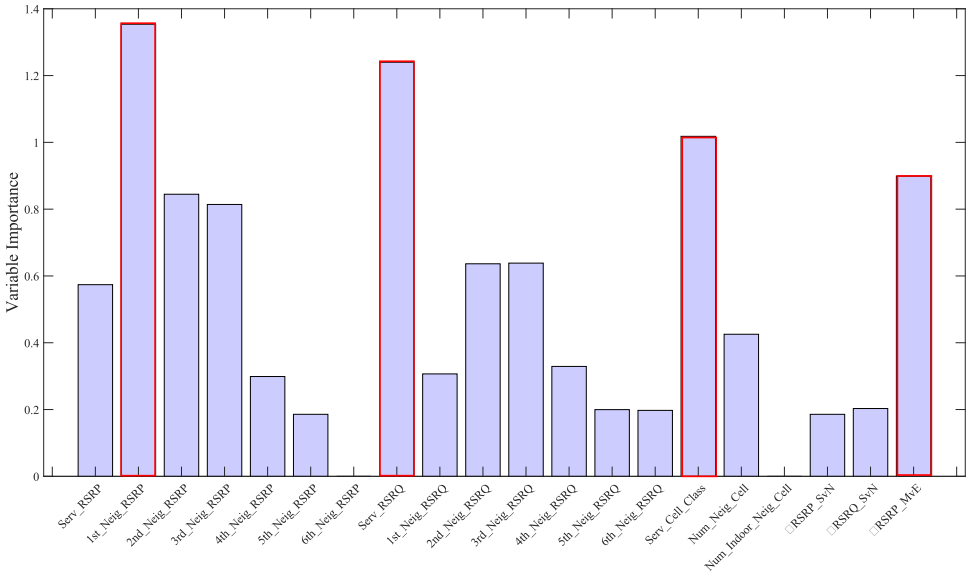

FIGURE 8. The variable importance in classification prediction.

decreases 75% compute time-consuming. On the other hand, the ensemble algorithms include bagged and boosted trees obtain higher accuracies than other popular machine learning methods. Especially, the bagged trees achieve the highest accuracy and relatively short time-consuming.

Note that even a single decision tree has good accuracy in very short compute time, over-fitting is the most practical difficulty limits the application on a large amount of samples in complex propagation environments. In addition, popular machine learning algorithms like SVM or KNN, they are sensitive to the missed samples and pre-set parameters(e.g., k in KNN). Ensemble learning method just overcomes these disadvantages. It can process the vast data with a large number of variables and/or a large proportion of the data are missing. Also, the capabilities of dealing with unlabeled data promote the ensemble scheme proposed in this study to be an optimal choice for IO classification based on cellular big data.

VI. CONCLUSION

In this paper, we present a radio propagation knowledge assisted ensemble scheme for IO classification based on cellular big data. The measurement conducted in the realistic urban area provides the first-hand physical context information for network operators. By evaluating the importance of the variables, the ensemble scheme can be simplified with predominant variables and less weak learners. The reconfigurable ensemble approach gives extremely high accuracy($> 99\%$) and low latency($< 5s$) for offering further personalized services, compared with other classical machine learning algorithms. The proposed approach is insensitive to the data missing, thus can be easily extended to more refined and dynamic context sensing tasks in a future smart environment.

REFERENCES

[1] N. Zhang, P. Yang, J. Ren, D. Chen, Y. Li, and X. Shen, "Synergy of big data and 5G wireless networks: Opportunities, approaches, and challenges," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 12–18, Feb. 2018.

[2] Y. Yang, K. Wang, G. Zhang, X. Chen, X. Luo, and M. Zhou, "MEETS: Maximal energy efficient task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4076–4087, Oct. 2018. doi: 10.1109/JIOT.2018.2846644.

[3] R. Li *et al.*, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[4] P. Grifoni, A. D'Ulizia, and F. Ferri, "Context-awareness in location based services in the big data era," in *Mobile Big Data* (Lecture Notes on Data Engineering and Communications Technologies), vol. 10. G. Skourletopoulos, G. Mastorakis, C. Mavromoustakis, C. Dobre, and E. Pallis, Eds. Cham, Switzerland: Springer, 2018.

[5] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.

[6] J. Lee, K. Lee, E. Jeong, J. Jo, and N. B. Shroff, "CAS: Context-aware background application scheduling in interactive mobile systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1013–1029, May 2017.

[7] C. Chen, "Large-scale scene classification with machine learning techniques," Ph.D. dissertation, Viterbi School Eng., Univ. Southern California, Los Angeles, CA, USA, 2017.

[8] B. Lee, C. Lim, and K. Lee, "Classification of indoor-outdoor location using combined global positioning system (GPS) and temperature data for personal exposure assessment," *Environ. Health Preventive Med.*, vol. 22, p. 29, Apr. 2017.

[9] Q. Ni, A. B. G. Hernando, and I. De la Cruz, "The Elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors*, vol. 15, no. 5, pp. 11312–11362, May 2015.

[10] Y. Kim, S. Lee, S. Lee, and H. Cha, "A GPS sensing strategy for accurate and energy-efficient outdoor-to-indoor handover in seamless localization systems," *Mobile Inf. Syst.*, vol. 8, no. 4, pp. 315–332, 2014.

[11] M. Okamoto and C. Chen, "Improving GPS-based indoor-outdoor detection with moving direction information from smartphone," in *Proc. ACM Int. Symp. Wearable Comput.*, 2015, pp. 257–260.

[12] W. Guan, X. Chen, M. Huang, Z. Liu, Y. Wu, and Y. Chen, "High-speed robust dynamic positioning and tracking method based on visual visible light communication using optical flow detection and Bayesian forecast," *IEEE Photon. J.*, vol. 10, no. 3, Jun. 2018, Art. no. 7904722.

[13] M. Yasir, S.-W. Ho, and B. N. Vellambi, "Indoor positioning system using visible light and accelerometer," *J. Lightw. Technol.*, vol. 32, no. 19, pp. 3306–3316, Oct. 1, 2014.
[14] B. Kim and S.-H. Kong, "A novel indoor positioning technique using magnetic fingerprint difference," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2035–2045, Sep. 2016.
[15] Q. Zeng, J. Wang, Q. Meng, X. Zhang, and S. Zeng, "Seamless pedestrian navigation methodology optimized for indoor/outdoor detection," *IEEE Sensors J.*, vol. 18, no. 1, pp. 363–374, Jan. 2018.
[16] I. Sharp and K. Yu, *Wireless Positioning: Principles and Practice*. Singapore: Springer, 2019. doi: 10.1007/978-981-10-8791-2.
[17] I. Rodriguez, H. C. Nguyen, I. Z. Kovács, T. B. Sørensen, and P. Mogensen, "An empirical outdoor-to-indoor path loss model from below 6 GHz to cm-Wave frequency bands," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 1329–1332, 2016.
[18] A. Ö. Kaya and D. Calin, "On the wireless channel characteristics of outdoor-to-indoor LTE small cells," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5453–5466, Aug. 2016.
[19] L. Zhang *et al.*, "Propagation modeling for outdoor-to-indoor and indoor-to-indoor wireless links in high-speed train," *Measurement*, vol. 110, pp. 43–52, Nov. 2017.
[20] M. Li, P. Zhou, Y. Zheng, Z. Li, and G. Shen, "IODetector: A generic service for indoor/outdoor detection," *ACM Trans. Sensor Netw.*, vol. 11, no. 2, 2015, Art. no. 28.
[21] M. Ali, T. ElBatt, and M. Youssef, "SenseIO: Realistic ubiquitous indoor outdoor detection system using smartphones," *IEEE Sensors J.*, vol. 18, no. 9, pp. 3684–3693, May 2018.
[22] V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina, "Poster: Am I indoor or outdoor?" in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2014, pp. 401–404.
[23] A. Karandikar, N. Akhtar, and M. Mehta, "Mobility management in LTE networks," in *Mobility Management in LTE Heterogeneous Networks*. Singapore: Springer, 2017.
[24] V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina, "A semi-supervised learning approach for robust indoor-outdoor detection with smartphones," in *Proc. 12th ACM Conf. Embedded Netw. Sensor Syst.*, 2014, pp. 280–294.
[25] X. Chen, Y. Xu, Q. Li, J. Tang, and C. Shen, "Improving ultrasonic-based seamless navigation for indoor mobile robots utilizing EKF and LS-SVM," *Measurement*, vol. 92, pp. 243–251, Oct. 2016.
[26] W. Wang, Q. Chang, Q. Li, Z. Shi, and W. Chen, "Indoor-outdoor detection using a smart phone sensor," *Sensors*, vol. 16, no. 10, pp. 1563–1577, Oct. 2016.
[27] O. Canovas, P. E. Lopez-de-Teruel, and A. Ruiz, "Detecting indoor/outdoor places using WiFi signals and AdaBoost," *IEEE Sensors J.*, vol. 17, no. 5, pp. 1443–1453, Mar. 2017.
[28] R. J. Lewis, "An introduction to classification and regression tree (CART) analysis," in *Proc. Annual Meeting Soc. Academic Emergency Med.*, San Francisco, CA, USA, 2000, pp. 1–14.
[29] *Physical Layer Procedures (Release 8)*, document 3GPP TS 36.213 V8.6.0, 3GPP, Mar. 2009.
[30] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

LEI ZHANG received the B.S. degree in communication engineering from Anhui University, Hefei, China, in 2009, and the M.S. and Ph.D. degrees in telecommunications from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2013 and 2016, respectively. In 2016, he was a Visiting Scholar with the University of South Carolina, Columbia, SC, USA. From 2016 to 2017, he was a Research Assistant Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science. Since 2018, he has been an Assistant Professor with the College of Information Science and Technology, Donghua University, Shanghai, China. His research interests include wireless channel sounding and modeling, mobile positioning, the Internet of Vehicles, and machine learning applications. Dr. Zhang was a recipient of the Extraordinary Doctoral Award (Premio Extraordinario de Doctorado) from UPM.

QIN NI received the M.S. and Ph.D. degrees in software and system from the Universidad Politécnica de Madrid, in 2013 and 2016, respectively. Since 2017, she has been an Assistant Professor with the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China. Her research interests include pervasive computing, data mining, smart environment, activity recognition, assistive robot design, and STEM education.

MENGLIN ZHAI received the B.S. degree in information science and engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in electromagnetic fields and microwave techniques from Shanghai Jiao Tong University, Shanghai, China, in 2016. Since 2016, she has been a Lecturer with the College of Information Science and Technology, Donghua University. Her current research interests include computational electromagnetics and applications, especially the FDTD method, EMI/EMC analysis, wireless communication simulation, and nanotechnology.

JUAN MORENO received the M.Sc. and Ph.D. degrees from the Universidad Politécnica de Madrid, in 2007 and 2015. He is currently a Rolling Stock Engineer with the Engineering and Research Department, Metro de Madrid, and also a part-time Professor with the Technical University of Madrid. He has been with the railway sector, since 2007. He has participated in many railway-related research projects. He has authored over 25 papers on railway communications. His research interests include channel measurement and modeling, railway communications systems, and software-defined radio.

CESAR BRISO was born in Valladolid, Spain, in 1968. He received the Ph.D. degree in mobile communications from the Universidad Politécnica de Madrid, in 1998. In 2000, he became an Associate Professor with the Technical University of Madrid, where he has been a Permanent Professor, since 2008. He is an Expert on the design of high-frequency circuits and systems and in propagation measurements and modeling in complex environments, such as tunnels, stations, UAVs. He has made 20 industrial and research projects with national and international companies and institutions. He has authored 40 publications on international SCI journals and more than 60 international congress. He was a recipient of the National Awards for the Best Ph.D. of the Spanish Association of Telecommunications Engineers.