ICCSPA20 1570628684

Using Machine Learning for In-Out decision accuracy for venue owner definable services

Wiqar Khan
NSW
Nokia
Espoo, Finland
wiqar.khan@nokia.com

Matti Keskinen
Mobile Networks
Nokia
Espoo, Finland
matti.keskinen@nokia.com

Asif Raza
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
razaasif@arcada.fi

Heidi Kuusniemi
Digital Economy
University of Vaasa
Vaasa, Finland
heidi.kuusniemi@uwasa.fi

Mohammed Elmusrati
School of Technology and Innovations
University of Vaasa
Vaasa, Finland
moel@uwasa.fi

Abstract— **Presence confirmation for being inside certain venue becomes matter of more importance when venue owner might have option to restrict or to provide value added contents for the user per its presence in a given venue during a given time window. In this paper, machine learning is applied to find the confidence of decision about a User Equipment (UE) presence inside a designated venue based on the accumulated data set used for learning. 20 UEs are used such that some are placed inside venue and other outside to collect data set to be used for ML algorithms. The outside locations are the possible human movement areas around the venue. The UEs works as reference data collection sources both from outside and inside. The received mobile network info by each UE is collected over extended time. Data is labeled based on the actual positions of the UEs. Using Python, Machine Learning is applied with very encouraging results to conclude the presence confirmation inside venue or the other way around. Hyper parameter tuning is applied for kNN ML algorithm.**

Keywords—Machine Learning algorithms, Android, LTE, RSRP, Python

I. INTRODUCTION

Venue owner might enrich or restrict services to a user when he/she is inside a given venue for a given time window. Though indoor positioning is widely discussed in different papers but this study of UE to be inside a venue or not is in principle different than the normal indoor positioning that are thoroughly described in [1] and [2]. Usually indoor positioning computes to pinpoint a UE based on the available techniques but here we would only be interested to know that whether the UE is present inside a given venue (could be part of larger indoor facility) or outside of it. There are many use cases where venue owner would prefer to have a say in restricting or enriching services for UEs that are inside the venue's premises [3]. The trigger of venue referenced services could be initiated by the UE and the presence with respect to the designated venue is required from the locality.

A. The refrenced indoor & outdoor:

In some cases, outside of venue might be fully outdoor, with respect to the venue of the interest. While in some cases the outside venue region set consists of partially outdoor and partially indoor space. In such cases the venue is next to inside a boundary of a bigger premises (as taken in this study). The signal propagation behavior of set of all UEs are different in such cases.

Moreover most of the studies for indoor cases are around WIFI [4] while in this article we use LTE (Long Term Evolution) RSRP (Reference Signal Received Power) values.

B. The setup:

A 108 seaters capacity auditorium was chosen as a venue of interest for data collection. The venue has one side towards open sky and half of the side wall consist of glass. The rest sides are indoor but with respect to venue, we consider it outside of the venue. 20 Android based LTE UEs were used to collect data through an app. The app scans for the network data and records it locally for later analysis. Figure 1 shows the layout of the UEs placement.

Some UEs were placed inside the venue (Zone1) and some were placed outside the venue. Among the outside space we have further two types: 'Zone 3' is situated on the right side of the venue- it is outside the venue but indoor; while 'Zone 2' is fully outdoor and situated on the left side of the venue. Towards 'Zone 3', we have passage, but user do not have direct passage towards 'Zone 2'.

978-1-7281-6535-6/20/$31.00 ©2020 IEEE

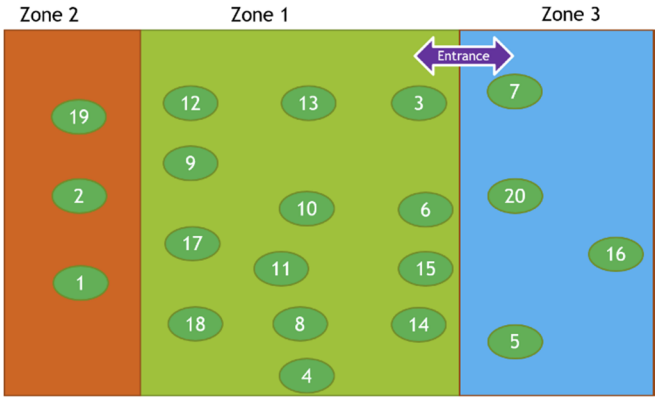

Figure 1 The Venue and UEs distribution

The auditorium view from inside is shown in Figure 2 . The phones were left there for extended period of around 8 hours such that we have longer term data as compared to snap drive test or walk test data collection (usually performed practice). The people movement around the venue and inside the venue was normal during the data collection time.

Figure 2 Inside view of the Venue with UEs

II. ANALYSIS:

The RSRP (Reference Signal Received Power) [5] values are used mainly that are obtained for different PCIs (Physical Cell Identity) of the mobile network using LTE network. There are comprehensive studies available regarding Machine Learning for WIFI indoor positioning e.g. [6] but we are using RSRPs of LTE network and our studied case consist of indoor and outdoor UEs.

ML (Machine Learning) approach is deployed using Python to find the IN/OUT answer with respect to the venue. Gaussian noise is assumed with different standard deviations to understand the noise tolerance. In result section, the analysis results are presented.

In our Machine learning approach, we use different Classifiers and use RSRP and PCI as an input feature to identify the Zone multi/binary class classification problem

Identifying the zone as classes is an ideal classification approach in our data set, input variables are described as features and labels (zones) are predicted as classes. The aim is to predict accurately class label in our data set.

We use ML for predicting the classes which are zones in our dataset using the supervised technique.

After the data cleansing, we label the data. Those labels are defined as Zones (Figure 1 and Figure 3). There are one to three zones. A zone represents the physical geographical space where phones are placed. Those zones are used as classes in our classifier.

A. Zoning:

As shown in Figure 1, the UEs used for data collection are divided into three zones per their locations: the right side of the venue; The left side of the venue; and the auditorium itself. It is considered as 3 zones concept. The front and back of the auditorium are not considered for UE placements as we believe that there less likelihood of users' movement. Alternatively, one can opt for 2 zones concept also, as shown in Figure 3, where geographical area inside the auditorium is called 'Zone 1' while the rest that is outside of the venue is grouped together, regardless of the side, is considered 'Zone 2'. The ML application, analysis and results are presented for '3 Zone' and '2 Zone' approaches separately.

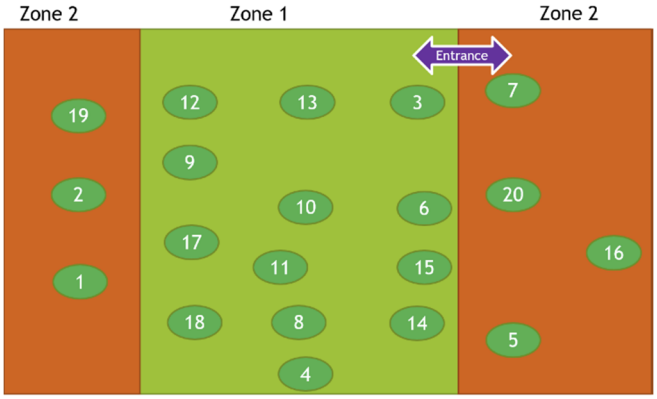

Figure 3 UEs location categorized into two zones

B. Balance and non-balanced dataset per zone

The UEs distribution in different regions and the captured data set per UE are not equal. Therefore, the data set sizes per zone (true for 2 zones and 3 zones) are not symmetric.

The original data captured is non balanced. It is named as 'non-balanced data size'. Depending upon the number of zones, the least zone data size is taken for the remaining zone(s) to have the balance data set that represent equal sizes of data from each zone. It is mentioned as 'balanced data size'.

Analysis are done for both set of data sizes.

C. ML approaches

The below ML algorithms are used for classification:

- kNN
- Logistic Regression (LR)
- Linear SVM (SVM)
- Random Forest (RF)

We use kNN as a benchmark classifier to classify the zone based on RSRP and PCI values. The other classification machine learning algorithms are used to compare the performance and test accuracy.

D. Per UE detailed analysis

The per UE data analysis is done based on the result's data available for the best performing algorithm analysis. It shows the details per UE for correctness of decisions made for a given UE based on the chosen algorithm. It represents the results as confusion matrix for each UE instead of a zone.

E. Noise impact

For training and testing the classifier, we first use the default parameters with all classifiers applied.

Gaussian noise with mean '0' and standard deviation '1.00' is added gradually to a chosen data set ('2 zones'- Balanced data set). Addition of such noise is done step wise to test data to check the noise impact. The impact on results accuracy is checked accordingly for each classifier.

For given data set with certain noise scenario the hyper parameters are applied for possible improvements to counter the impact on the results accuracy because of noise.

III. Results and Discussion

A. Captured data observations:

i. RSRP spread

RSRP values of serving PCIs captured over the data collection time are presented for each UEs with the help of box plot as shown in Figure 4. It shows the RSRP values spread regardless if the serving PCI for a given UE was changing during the data collection duration.

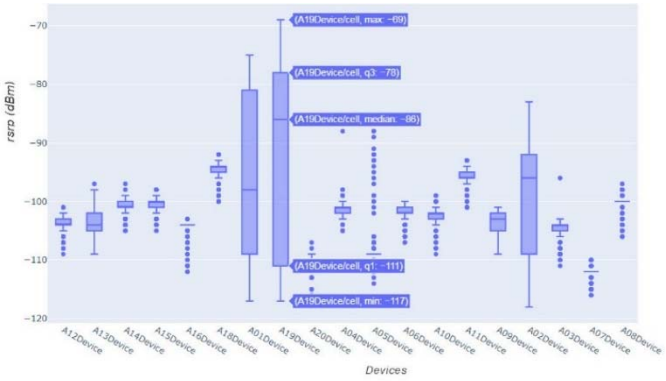

Figure 4 Box plot per UE for rsrp from serving PCIs

ii. PCIs counts

The data captured contains info about the PCIs serving a given UE per scan. Also, it gives the neighboring candidate PCIs that were available to serve a UE in case the UE need to switch to different PCI based on the network configuration.

All the PCIs appearance in the overall data set collected for all UEs is shown in Figure 5. This pie chart represents the appearance of any PCI, represented as percentage, in the collected data set regardless the PCI was serving any UE, or it was a candidate PCI for a UE.

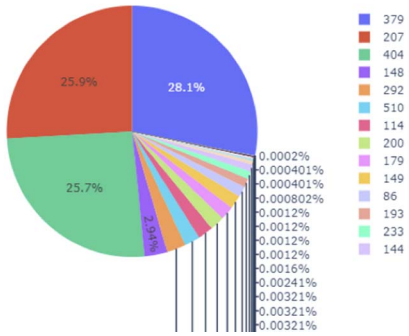

Figure 5 PCIs appearance, both serving and as neighboring

iii. Serving PCIs counts

This gives counts of PCIs appearance in the overall collected data only for the cases where the PCI was serving a UE. The count per serving PCI (without info of served UEs) is given in Figure 6 . The serving PCI's swap for some UEs was observed time by time for different UEs.

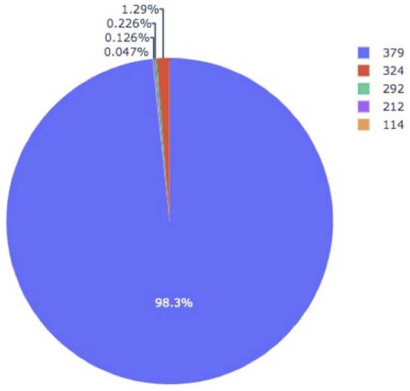

Figure 6 Serving PCIs counts in overall collected data

B. Results

i. ML summary:

The results of employing chosen ML algorithms for given data sets per zoning concepts are summarized in Table 1.

Table 1 ML algorithms results per data set

ML algorithm	Score	2 Zones		3 Zones	
		Non-balanced data set	Balanced data set	Non-balanced data set	Balanced data set
kNN	train_score	96,20 %	94,00 %	95,30 %	86,10 %
	test_score	96,00 %	94,20 %	95,00 %	86,30 %
Logistic Regression	train_score	94,80 %	85,40 %	93,60 %	82,20 %
	test_score	94,50 %	85,40 %	93,40 %	82,50 %
Random Forest	train_score	96,20 %	94,00 %	95,30 %	90,30 %
	test_score	96,00 %	94,00 %	95,00 %	90,20 %
Linear SVM	train_score	94,70 %	85,90 %	93,70 %	84,80 %
	test_score	94,50 %	86,00 %	93,40 %	84,40 %

The training data and testing data split was taken as 70% and 30% respectively. The results show that the 2 zone data sets provide better accuracy to determine correctly the UE presence in one of the zones based on the ML over the collected data. It shows that the decision making to predict the UE being 'In' or 'Out' with respect to a venue is easier as compare to predict the presence in one of the 3 zones. Practically, the 3-zone concept would help us to determine that UE is either 'In' (in Zone 1) or 'Out' (in Zone 2 or Zone 3). In case of 'Out', it will specify further that whether it is on the left side (in Zone 2) or on the right side of the venue (in Zone 3).

To determine any UE presence, based on the learning algorithms, in a given venue and time window '2 Zones' approach is sufficient.

The score of non-balanced data is higher and it could be because of the biasness that could arise from the more available data of a given zone than the other. The balancing of the data with respect to the zone that has lesser data gives us a balanced data set. From the results of Table 1, we can see that 2 zones non-balanced data set has higher scores as compared to balanced data set, but it is relatively very small difference hence the biasness in data sizes per zone is not dominant on accuracy results.

The data sets split for '3 Zones' non-balanced data among 'Zone 1', 'Zone 2' and 'Zone 3' is 81,5 %, 4,2 % and 14,3 % respectively. '3 Zone balanced data set' is achieved by balancing the data for each zone as 33,33% after taking data for remaining zones per Zone 2 data size being Zone 2 data size is the minimum one.

ii. Confusion matrices

Random forest approach is used for further analysis of the results for confusion matrices.

Normalized confusion matrices for 2 zone non-balanced and balanced data sets are given Figure 7 and Figure 8 respectively.

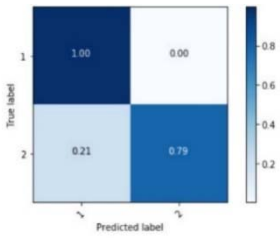

Figure 7 Confusion Matrix, normalized, Non-Balanced 2-Zone data

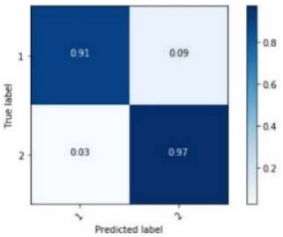

Figure 8 Confusion Matrix, normalized, Balanced 2-Zone data

The 3-zone data set confusion matrices are given in Figure 9 and Figure 10 respectively.

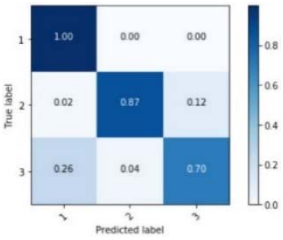

Figure 9 Confusion Matrix, normalized, Non-Balanced 3-Zone data

The non-balanced data set has little bit dominances on the results, and it is per expectations as the data set size of Zone 1 is higher.

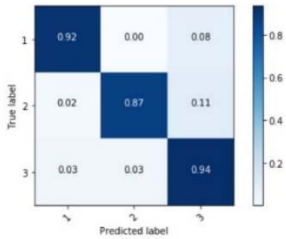

Figure 10 Confusion Matrix, normalized, Balanced 3-Zone data

Moreover, the False True normalized figures are not symmetric both for 2 Zone and 3 Zones (for balanced & and non-balanced data sets.

iii. Per UE perdiction detailed analysis

Confusion matrices give us good inside for relative errors that we have in our prediction at zone level. Per UE level analysis would show us that what UEs are contributing more to the non-successful predictions for a given zone.

The detailed analyses are also computed based on the random forest results.

In Figure 11, each UE is represented by 3 bar charts-the actual values (the number of counts), the correct predictions and the false predictions of the respective zone. Depending upon the UE own location, its correct or false predictions is represented by either Pred1 or Pred2 values. Pred1 for a UE would mean that how many times the UE was predicted in Zone 1. If the UE originally belong to Zone 1 then the Pred1 counts shows the correct predictions count.

The results show that UE3, UE 12 and UE 13 are more contributing to the non-correct predictions. Per layout of these UEs, given in Figure 3, it is expected so as these are near to the boundaries of the zone.

On the other hand, the UEs that are deep inside the auditorium have better predictions values e.g. UE18 having 100%.

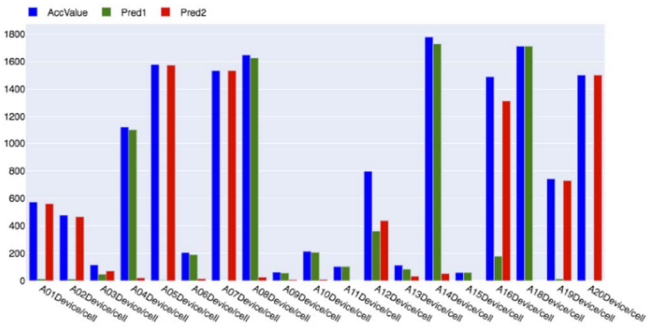

Figure 11 Per UE predictions analysis, Balanced 2-Zone data

In Figure 12, the '3-Zones balanced data' per UE predictions analysis under Random Forest algorithm is shown. It shows per UE prediction in the same way as above but for 3 zone case.

Per UE analysis for '3-Zones balanced data set' gives the same results as we have for '2-Zones balanced data set'. Additionally, for '3-Zones' case, UE 7 is also having relatively more false predictions.

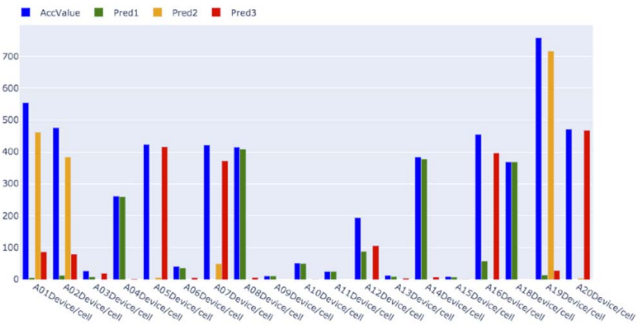

Figure 12 Per UE predictions analysis, Balanced 3-Zone data

iv. Impact of noise addition

Gaussian noise is added to see the impact on our results. Only '2-Zones balanced data set' is taken for noise addition impact analysis.

Gaussian noise with mean value of '0' and standard deviation of '1.00' is generated and added to 3%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%,70% and 100% of the "test data" to check accordingly the gradual noise impact on the classifier accuracy. The impact on accuracy results is shown in Figure 13. The horizontal axis shows the percentage of the test data used with noise. The vertical axis shows the accuracy of the 'In' and 'Out' results. It shows that as the noise added test data size increases the accuracy of the prediction of 'In' and 'Out' decision also decreases.

Next subsection describes the recovery of such degradation in accuracy by applying hyper parameters for given ML algorithm.

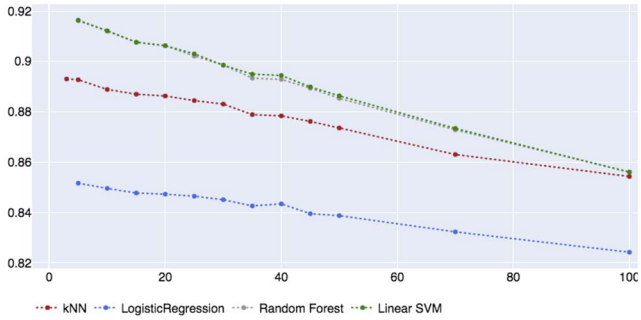

Figure 13 ML accuracy percentages for Gaussian Noise addition to given portion of test data

v. Correction efforts- Hyper parameters for classifier improvements

Hyper parameters are applied to improve the accuracy which was reduced by adding the gradual noise.

The '2-Zones balanced data set' that has Gaussian noise added with mean value '0' and standard deviation '1.00' to 50% of the test data is taken for correction efforts.

By introduction of the Gaussian noise on 50% of "test data" the test accuracy reduced by 1.87% in kNN, 1.53% in LR and 1.47% in RF.

We tried improving the test accuracy by using the hyperparameter of kNN and Linear SVM classifiers. The best result produced by kNN with hyperparameters tuning. The test accuracy was improved 2.12% (87.33% to 89.45%).

Results are summarized in Table 2.

Table 2 ML algorithms performance improvements using hyperparameters

ML algorithms	Test prediction score 2 Zones balance data			
	Without-Gaussian Noise	With-Gaussian Noise (0,1) & added to 50% test data	Using hyper Parameters	Improvement in predictions
kNN	94,20 %	87,33%	89.45%	+2.12%
Logistic Regression	85,40 %	83,87%	Hyper Parameters appraoch was not applied for these algorithms	
Random Forest	94,00 %	88.53%		
Linear SVM	86,00 %	88.61%	89.2%	+0.59%

IV. CONCLUSION AND RECOMMENDATIONS

We can conclude with higher confidence of presence of UE in 2 zone as compared to 3 zone case. It is in line with the expectation to find only the presence of a UE being either inside or outside of a venue as compared to situation of pointing a UE with more precise geographical location out of 3 zones.

kNN and Random forest turned out to be the most reliable algorithms.

Another important result was the kNN hyper parameter usage to counter the confidence degradation because of noise addition.

To achieve results with more confidence, we shall have balanced data for more reference UEs equally distributed in all the designated zones around the venue.

The given venue did not have any of its own small cell. All the PCIs were remotely away from the venue building. To achieve higher accuracy, local small cell shall be helpful.

Some UEs that were near the entrance or the boundary of the zone are having relatively higher non-correct predictions. To achieve the 'In' and 'Out' higher accuracy, directional antennas usage in the given venue shall be beneficial.

Including other features, e.g. neighbors' list and its rsrp or other RAT (Radio Access Technology) info shall be useful to achieve higher accuracy.

REFERENCES

[1] P. &. P. R. Davidson, "A Survey of Selected Indoor Positioning Methods for Smartphones," *IEEE Communications Surveys and Tutorials,* pp. 1347-1370, 2017.

[2] L. B. a. M. Tomic, "Overview of indoor positioning system technologies, , pp. 0473-0478.," in *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2018.

[3] W. G. Khan and K. E. O. Nyman, "VENUE OWNER - CONTROLLABLE PER - VENUE SERVICE CONFIGURATION". United States of America Patent US010320973B2, 11 June 2019.

[4] T. P. S. A.-L. a. R. P. V. Honkavirta, "A Comparative Survey of WLAN Location Fingerprinting Methods," in *Proc. of the 6th Workshop on Positioning, Navigation and Communication*, 2009.

[5] 3. G. P. Project, "3GPP TS 36.214 V14.4.0," 3GPP, 2017.

[6] S. Bozkurt, G. Elibol, S. Gunal and U. Yayan, "A comparative study on machine learning algorithms for indoor positioning," in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, Madrid, Spain, 2-4 Sept. 2015.