Comparison of Machine Learning algorithms for venue presence with inclusion of neighbours

Comparison of Machine Learning algorithms for venue presence with inclusion of neighbours

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

Asif Raza
Department of Business Management and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
raza.asif@gmail.com

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— **User presence determination for being inside a venue, such that the user is provided with possible value-added services, is of high significance. It will get more prominent as we move to 5G and 6G networks' rollout as we'll get further means to have better aids. In this paper, machine learning (ML) algorithms computation results are obtained and analysed. Such algorithms would be candidate to be deployed for finding the confidence in decision making for a user's location with respect to a venue. Number of UEs (User Equipment) are simultaneously placed inside and outside a venue and kept over a longer duration. Data such as received reference signal received power for serving cells and neighbour candidate cells etc. data is collected by each UE. The different available neighbours' level in each data set is analysed. k-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) algorithms are used to find the accuracy based on neighbours' depth among the available info. Very convincing results are observed over different level of neighbours being included in each Machine Learning (ML) algorithms.**

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

I. INTRODUCTION

For a given venue, a user location determination is of high interest with respect to services that could be defined/restricted per venue. Different techniques have been deployed for positioning in the indoor facilities. Such techniques range from indoor satellite techniques [1] to mobile and WIFI networks-based techniques. A good survey about selected indoor positioning is presented in [2]. As we move towards 5G and 6G networks, we'll have more use cases for the indoor positioning. Along with the intrinsic features of the latest networks, the latest trends of ML and AI are giving us better chances for higher accuracy per needed use cases.

Indoor positioning is considered as bases towards indoor navigation. There is another aspect for positioning as well. It is finding a user's placement with respect to a venue that would offer venue owner driven configurable services as defined in [3]. It is different in nature from those cases where emphases are on locating a user against a pin point coordinates (with or without navigation involved).

Usually, in research caried out, it is observed that the same UE is moved around through a walk test or through some other means e.g., mobile robot etc to capture data at different location. While in our case we have collected original rich data set with the help of 20 android UEs that were placed inside and outside a venue for longer duration of time. All the UEs were simultaneously capturing all important info of available mobile network and WIFI by running an application on each UE. There was movement of people around and the data collection was done in natural way instead of a controlled environment to have real life case.

This work is part of larger plans where UE's request or the venue owner request will be incorporated with full ecosystem and the finding from this work will be utilised for real use cases.

The data collection setup was done at a conference hall inside Nokia office. The hall seating capacity is over hundred participants. Two zones were defined- Zone 1 and Zone 2 as depicted in Fig. 1. The inside venue UEs belong to Zone 1 and the rest of UEs that were placed outside the venue belong to Zone 2. All the UEs were placed in areas such that there could be probable movement of users in real life cases.

The part of Zone 2 that is on the left side of Zone 1 is a passage where people just pass by. While the right side of Zone 1 is area where some people could pass by or sit as they go out of the hall.

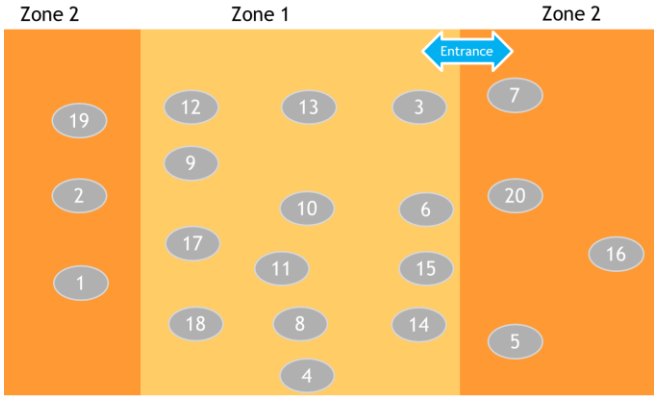

Fig. 1. The Venue and UEs distribution

On the collected data, machine learning algorithms are applied to see the confidence level for the decision accuracy for each UE to be inside or outside the venue per given zones defined. There are further iterations done based on data sizes and neighbour's depth combinations.

II. ANALYSIS:

The RSRP (Reference Signal Received Power) [4] values are mainly used that are obtained for different PCIs (Physical Cell Identity) of the mobile network using LTE network. In [5] only the serving cell PCIs were used while in this paper we are using neighbour cells' PCIs as well with different depths i.e. different counts of neighbours are considered. There are comprehensive studies available regarding Machine Learning for WIFI indoor positioning e.g., [6] and merging WIFI and BLEs signal power [7]. In this article using RSRPs of commercial LTE network and our studies case consist of indoor and outdoor UEs. Unlike work done in [7], our approach does not need dedicated additional devices e.g., BLEs. A good reference where real cellular networks are used is given in [8]. There two indoor positioning methods, based on Support Vector Machine learning algorithms and space-partitioning principle, are proposed. In our case on the collected data, different ML algorithms are iterated with increased neighbours' PCIs and its respective RSRPs along with the serving PCIs and its RSRP values. Our emphasis is to find the impact on the accuracy of a user presence to be inside a venue our outside as we increase the number of neighbouring cells in addition to serving cell with respective RSSP values.

A. Machine Learning description:

Difference ML algos are taken applied using Python. In our ML methodology we have used different Classifiers and used PCIs and RSRP as our input feature to identify the Zone binary class classification.

We are taking available neighbours' information depth, considering the number of neighbours' PCI's and its respective RSRPs, in addition to serving PCI and its RSRP. Hence the numbers of features are increased for each analysis set.

After the data cleansing, we label the data. Those labels are defined as Zone and grouped as Zone 1 and Zone 2. A zone represents the physical geographical space where phones are placed. Those zones are used as classes in our classifier.

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.

The ML algorithms used for classification are below:

- k-Nearest Neighbour (KNN)
- Logistic Regression (LR)
- Random Forest (RF)
- Decision Tree (DT)

These algorithms are used to compare the performance and test accuracy. The comparison results are presented in the Results section.

B. Neighbours' depth

We have iteration for ML algorithms on the original data based on neighbours' depth. In first such iteration, data set is obtained without any neighbour i.e. only serving cell PCI and RSRP are taken as features. For second iteration data set, only one neighbour (PCI of first neighbour from neighbours' list, and its respective RSRP) are added along with serving PCI and its RSRP as feature. While for third iteration data set is obtained with inclusion of two neighbours (first two neighbours from neighbours' list) PCIs and respective RSRPs along with the serving cell PCI and its RSRP. In each data sets we have additional features based on neighbours' layers inclusion.

C. Balance and non-balanced dataset per zone

The number of UEs placed in a zone and the captured data set per UE are not identical. Therefore, the data set sizes per zone are not symmetric.

The original data captured is non balanced for both zones and it is named as 'non-balanced data size'. The least zone data size is taken as a reference such that the other data set is randomly picked to make it equal to the size of lesser data set. Such data set is named as the 'balanced data set'. It has equal sizes of data from each zone.

After iteration on original non-balanced data, the neighbours' depth iterations are repeated with the balanced data set by increasing number of neighbours from 0 to 2.

D. Confusion matrices

The confusion matrix analysis is computed for all algorithms. The ML algorithm giving the better results is chosen for detailed confusion matrix details and they are presented at the end.

III. RESULTS AND DISCUSSION

A. Results and observations

The data is split into 70% for training the model and 30% for testing the model performance. The results shown below represents training and testing performance of the model.

The summary of the results for all four ML algorithms i.e. KNN, Logistic Regression (LR), Random Forest (RF) and Decision Tree (DT) are shown in TABLE I and TABLE II for non-balanced and balanced data sets respectively. It provides a comprehensive comparison among different ML algorithms and the count of neighbours in each computation.

TABLE I. ML ALGORITHMS RESULTS SUMMARY FOR NON-BALANCED DATA SET

non-balanced data sets			
Neighbours' depth	**ML algorithm**	**Training Accuracy**	**Test Accuracy**
0	KNN	0.962	0.960
	Logistic Regression	0.948	0.945
	Random Forest	0.962	0.960
	Decision Tree	0.962	0.960
1	KNN	0.976	0.976
	Logistic Regression	0.944	0.941
	Random Forest	0.969	0.967
	Decision Tree	0.971	0.972
2	KNN	0.993	0.992
	Logistic Regression	0.942	0.940
	Random Forest	0.985	0.984
	Decision Tree	0.985	0.986

For ‘balanced data set’ and ‘non-balanced data, the ML algorithms are iterated based on neighbours’ depth. Based on 3 neighbours’ depth and balancing logic for 4 algorithms we have 24 combinations of data sets and respective results.

From the results it is visible that KNN is the best performing ML algorithm. It becomes more prominent as more neighbours are taken into account. KNN effectiveness is confirmed in other [9] same domain researched work carried.

TABLE II. ML ALGORITHMS RESULTS SUMMARY FOR BALANCED DATA SET

Balanced data sets			
Neighbours’ depth	**ML algorithm**	**Training Accuracy**	**Test Accuracy**
0	KNN	0.940	0.940
	Logistic Regression	0.886	0.884
	Random Forest	0.940	0.940
	Decision Tree	0.940	0.94
1	KNN	0.977	0.975
	Logistic Regression	0.879	0.876
	Random Forest	0.976	0.974
	Decision Tree	0.977	0.976
2	KNN	0.991	0.985
	Logistic Regression	0.821	0.818
	Random Forest	0.984	0.983
	Decision Tree	0.986	0.986

The overall results for test accuracy is represented in Fig. 2.

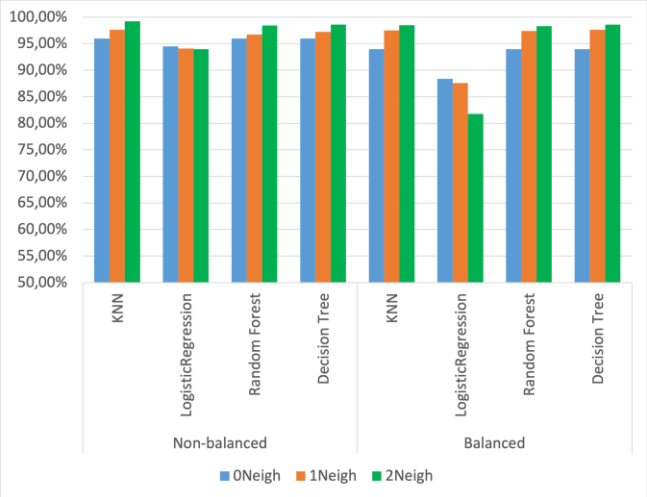

Fig. 2. Overall test_score comparison

It is observed that increase of more neighbours into our analysis we get better results. The inclusion of further neighbours for higher neighbours’ depth will take our algorithm results towards 100%.

As more neighbours are included, there will be higher computation power needs. The number of available neighbours might not be identical in all data points.

The LR performance is different. Firstly, it is the least performing algorithm. It is because of the nature of the data. Secondly, in Fig. 2, it is visible that with the increase in the neighbours’ count the results of the LR gets poorer. Logistic regression is not able to perform well to differentiate the multi-classes in high dimensional space as we have data increase by inclusion of more neighbouring cells. Other ML algorithms specially KNN is able to identify the class boundaries of neighbouring cells much better than others.

We see underfitting for LR. The LR assume linearly separable problem. Moreover, LR requires small or even no multicollinearity between the independent variables.

B. Confusion matrix

From previous sub section, it is mentioned that KNN has over all edge over other algorithms when we analyse overall results. Since we have many cases therefore Confusion matrices analysis is only given for KNN. Normalised confusion matrices are depicted in Fig. 3 to Fig. 8.

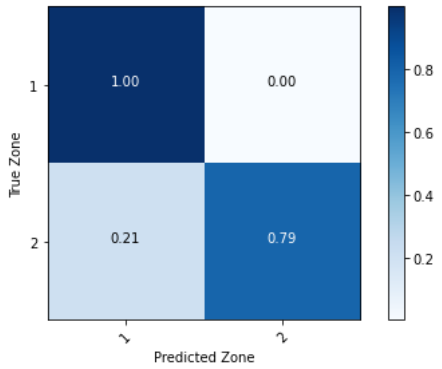

Fig. 3. Confusion Matrix, Neighbours’ Count=0, non-balanced data

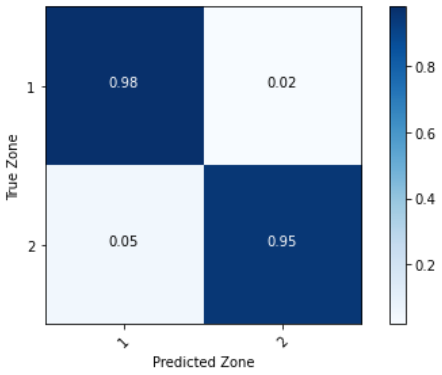

Fig. 4. Confusion Matrix, Neighbours’ Count=1, non-balanced data

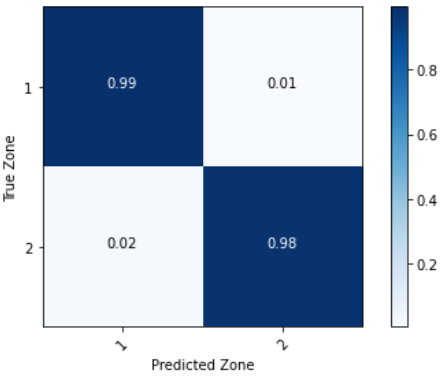

Fig. 5. Confusion Matrix, Neighbours’ Count=2, non-balanced data

Zone 1 has more data as compare to Zone 2. Moreover, the serving cells of Zone 1 have high RSSI values with consisting trend. Therefore, the Zone 1 accuracy is better than Zone 2. As we include the first neighbouring cells, being relatively away cells, the respective RSSIs are not that dominant therefore the Zone 1 accuracy decreases a bit but with higher data samples i.e. after we include the 2nd neighbouring cell, we get back to 99% accuracy for Zone 1.

The confusion matrices for balanced data depicted from Fig. 6 to Fig. 8, do not suggest much dominant accuracy in either zone. For both zones, the accuracy increases as we get more neighbouring cells added in to our computation.

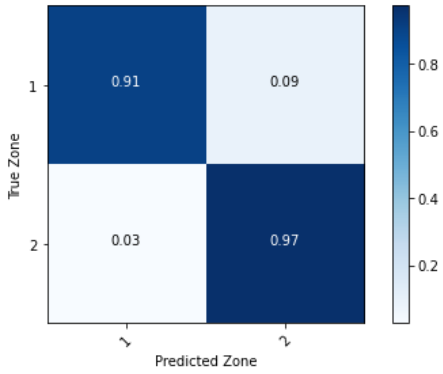

Fig. 6. Confusion Matrix, Neighbours' Count=0,balanced data

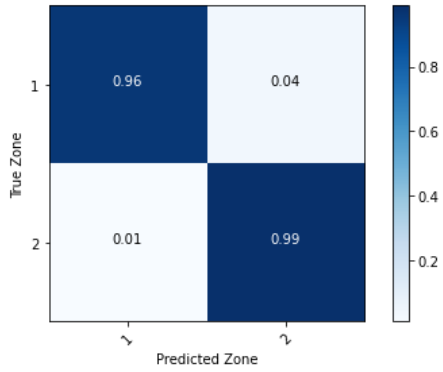

Fig. 7. Confusion Matrix, Neighbours' Count=1, balanced data

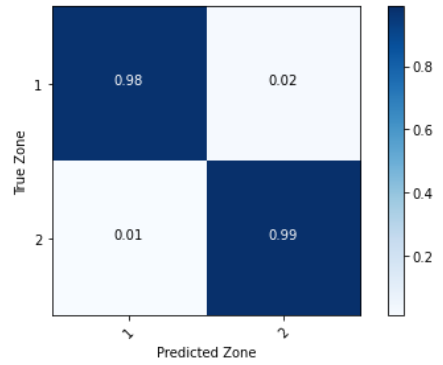

Fig. 8. Confusion Matrix, Neighbours' Count =2, balanced data

IV. Conclusion and Recommendations

The best ML algorithm that we conclude here is KNN. Its accuracy gets better as we have more neighbours considered and adding features based on neighbours' info.

Balanced data at hand is better option therefore in practical cases, the data collection should be done so that we obtain equal data sizes for each Zone and do the training based on such sets.

In our venue case, there were two possible sides for natural users' (i.e. UEs') movement. The same type of analysis could be carried for venues where we can have four sides with possible UEs. Such cases will be of importance when we take into account real world use cases.

The venue did not have its own cell inside the venue. The small cell inside venue or near to venue will contribute to better results.

It will be worth to include more features in addition to RSRPs and PCIs.

References

[1] H. Kuusniemi *et al*., "Utilizing pulsed pseudolites and high-sensitivity GNSS for ubiquitous outdoor/indoor satellite navigation," *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2012, pp. 1-7, doi: 10.1109/IPIN.2012.6418911.

[2] P. Davidson and R. Piché, "A Survey of Selected Indoor Positioning Methods for Smartphones," in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1347-1370, Secondquarter 2017, doi: 10.1109/COMST.2016.2637663.

[3] W. G. Khan and K. E. O. Nyman, "Venue owner-controllable per-venue service configuration," U.S. Patent 10 320 973 B2, Jun. 11, 2019.

[4] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements,* 3GPP TS 36.214 V14.4.0, 2017.

[5] W. Khan, M. Keskinen, A. Raza, H. Kuusniemi and M. Elmusrati, "Using Machine Learning for In-Out decision accuracy for venue owner definable services," *2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2021, pp. 1-6, doi: 10.1109/ICCSPA49915.2021.9385759.

[6] S. Bozkurt, G. Elibol, S. Gunal and U. Yayan, "A comparative study on machine learning algorithms for indoor positioning," *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 2015, pp. 1-8, doi: 10.1109/INISTA.2015.7276725.

[7] S. Tsuchida, T. Takahashi, S. Ibi and S. Sampei, "Machine Learning-Aided Indoor Positioning Based on Unified Fingerprints of Wi-Fi and BLE," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1468-1472, doi: 10.1109/APSIPAASC47483.2019.9023051.

[8] M. Petric, A. Neskovic, N. Neskovic and M. Borenovic, "Indoor Localization Using Multi-operator Public Land Mobile Networks and Support Vector Machine Learning Algorithms," *Wireless Personal Communications*, vol. 104, pp. 1573–1597, 2019, doi.org/10.1007/s11277-018-6099-1

[9] X. Han and Z. He, "A Wireless Fingerprint Location Method Based on Target Tracking," *2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)*, 2018, pp. 1-4, doi: 10.1109/ISAPE.2018.8634177.