KKU ENGINEERING JOURNAL 2016;43(S2):338-341 Research Article

KKU Engineering Journal

https://www.tci-thaijo.org/index.php/kkuenj/index

Filter random forest for indoor Wi-Fi positioning

Shutchon Premchaisawatt and Nararat Ruangchaijatupon*

Department of Electrical Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand.

Received April 2016
Accepted June 2016

Abstract

This paper proposes the method called the filter random forest (FRF), which can enhance an accuracy of indoor positioning based on fingerprinting by employing the random forest (RF) algorithms and informative access point (AP) selection. FRF selects the informative APs from all APs. This process reduces noise data and complexity of FRF's learning. FRF is compared with the machine learning classifiers; i.e. RF, decision tree (DT), bagging (BAG) and boosting (BOOST), by exploiting the signal strength from the real measurement. The performance comparison is done in terms of accuracy of classification of positions and computational complexity of algorithms. The result of this study shows that FRF's accuracy is very similar to BAG's accuracy which is more accurate than DT and RF. Besides that, the computational complexity of FRF is the lowest among the others due to the effect of AP reducing.

Keywords: Indoor positioning, Wi-Fi, Machine learning, Ensemble

1. Introduction

The Wi-Fi based positioning technology is a very important application because various electronic devices; i.e. smartphones, tablets, laptops, are widespread in daily life [1]. Those devices are equipped with Wi-Fi receiver. The fingerprinting is the one of prevalent Wi-Fi positioning due to cost-effectiveness [2]. Therefore, researchers attempt to develop Wi-Fi based positioning, which is more robust, accurate and cost-effective. The traditional fingerprinting uses standalone machine learning such as artificial neural network [3], K-nearest neighbor (K-NN) [4], and etc. Meanwhile, the group of machine learning working together called ensemble [5], is the techniques that used to increase the performance of standalone algorithms. There are two phrases in fingerprinting. In the offline phrase, received signal strength (RSS) data are collected and used them for algorithm training. The RSS data contains relation between RSSs of APs and positions. The online phrase uses a sample of RSS data to determine the position from the trained algorithm from the offline phrase. Usually, the ensemble algorithms have high computational complexity in the training process. Therefore, the proposed FRF reduces computational complexity while providing high accuracy of positioning.

2. Filter random forest

2.1 Informative AP filtering

Filter [6] is the algorithm for selecting features; i.e. selecting access points, before process with machine learning algorithms. Filter relies on information gain theory [7], which is used in a decision tree to measure good features for decision making. Filter can determine the informative access points, which are the access points that provide useful information for positioning. Therefore, the informative access points lead to correct predictions. Let D be the set of all samples that obtained from the measurement. These samples contain relation between RSS from all access points and each position m from all M positions. The number of samples measured at each position m is equal. The information gain of each access point ($gain\ (ap_i)$) can be calculated by using equation (1). Let *APs* be the set of all access points whose RSS can be measured and ap_i is an access point in the set *APs*. Let V be the set of non-duplicated RSS values measured from ap_i and v_j is each value in the set V. D_{v_j} is the subset of D, in which the RSS obtained from ap_i equals v_j and P_m is the probability of a position m obtained from the access point ap_i. P_m is calculated by dividing the number of samples in subset D_{v_j} which associated to position m by the number of all samples in D_{v_j}.

After information gain of all access points in the set *APs* is obtained, these access points are sorted by their values of information gain. The access points with high information gain illustrate that they are significant to predict the correct positions. These access points are called the informative access points.

$$gain(ap_i) = -\sum_{j=1}^{|V|} \frac{|D_{v_j}|}{|D|} (-\sum_{m=1}^{|M|} p_m \log_2 (p_m)) \qquad (1)$$

*Corresponding author. Tel.: +66 4334 7031
Email address: nararat@kku.ac.th

doi: 10.14456/kkuenj.2016.124

2.2 Random forest (RF)

The random forests [8] is one of the ensemble techniques in machine learning. It relies on the bagging method which produce a randomly sampled set of training data to each of the weak learner (decision tree in RF). These weak learners have different knowledge. The majority of answers of weak learners are the answer of RF.

2.3 Filter random forest (FRF)

Filter random forest is the random forest, built with RSS data from informative filtered APs. During this process, only some informative APs are selected to omit samples from APs that are not informative. Then, the random forest is built by data D_i from data $D_{filtered}$ which comes from $AP_{filtered}$. In classification process, the major answer of the weak learners in FRF is an answer of FRF. The pseudo code of FRF is shown in Figure 1. In the experiment, the proposed method is compared with other ensemble algorithms, which combine the predicting ability of many weak learners to predict the answer.

```
Pseudo code for the filter random forest (FRF) algorithm
#AP filtering
Find and sort information gain of each AP from the dataset $D$
The $AP_{filtered}$, The APs have information gain more than (threshold * the largest information gain) or called the informative APs, are selected to create the dataset $D_{filtered}$
#Building random forest
Create $n$ decision tree:
for $i$ = 1 to $n$
    Randomly, APs are selected from $AP_{filtered}$ and resample from dataset $D_{filtered}$ to create $D_i$
    Build decision tree with $D_i$
end for
#Classifying
the sample $d$ is classified by $n$ decision tree in the random forest the major answers of $n$ decision trees are the answer of the filter random forest
```

Figure 1 Pseudo code for filter random forest (FRF) algorithm

3. Experiment

The experiment is set up in the 30x10 m^2-sized area with the ceiling height of 2.8 m as illustrated in Figure 2. The distances between measured points, i.e. mark points, are 4 square-meter grids with 90 reference points (90 classes for classification) from 156 APs whose RSS can be discovered in the experimental area. The discovered APs are the facility of the university and the department. The RSS data was measured by a smartphone, Motorola Moto G with WLAN (font) Wi-Fi 802.11 b/g/n. For each reference point, RSS is measured 20 times with 1-second delay. The measuring process is repeated 3 days of the workday (there are people who do the usual activity in the building). On each day, the measurement can be divided into 4 time periods, i.e. 8:00 – 10:00, 10:00 – 12:00, 13:00 – 15:00, and 15:00 – 17:00. Hence, this created 90*20*3*4 = 21,600 samples for measured data. These data are provided to the machine learning algorithms to train and test positioning accuracy by using 10 folds-cross validation.

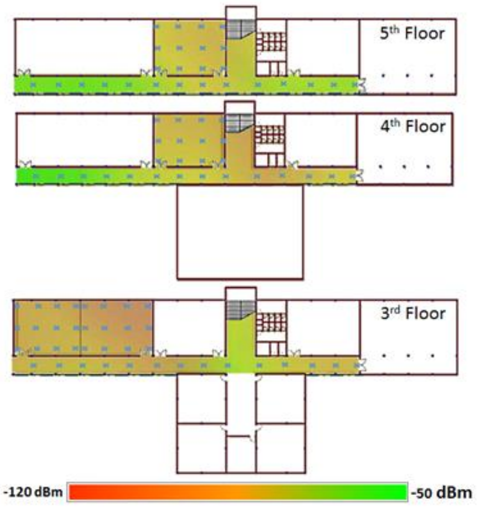

Figure 2 The experimental area

The accuracy of the proposed method is compared with those of other four algorithms. First, the decision tree is the standalone model which is used in traditional fingerprinting positioning. The standalone decision tree is the baseline to determine the performance of positioning. Second, the traditional random forest has the detail as aforementioned. Third, bagging consists of many identical algorithms, but each one in bagging is trained by different samples. Then, it chooses a majority of all answers [9]. The last one, boosting tries to increase weights of the incorrect samples. After that, boosting relearns from the adjust weight sample to create new knowledge [10].

The configuration of these ensemble algorithms is the number of weak learners. Each ensemble algorithm is built with 3 and 7 weaker learners because the effect of the less number of the weak learners is considered. Then the accuracy of each algorithms are compared.

4. Result and discussion

For filter random forest, the appropriate number of informative APs have to be determined.

The threshold in the FRF is selected from the proportion that provides high accuracy of DT prediction during APs are more omitted. The information gains are calculated for each AP and arrange APs by the ascending information gains. In Figure 3, it shows the relation between the accuracy of DT and proportion of AP omitting; omitting starts from the lowest information gain to the highest information gain in the APs arrangement. 20% is the suitable proportion of AP omitting because it decreases the number of APs in calculation while the accuracy is still high. Therefore, the threshold of FRF calculation is 0.2.

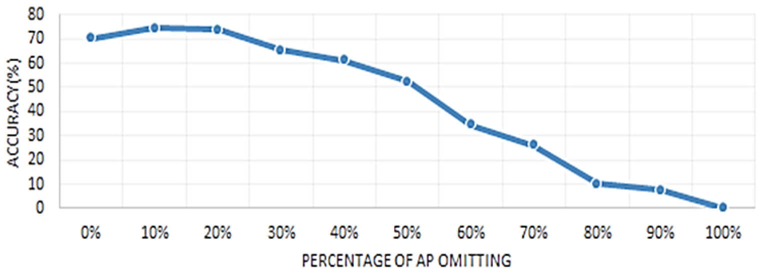

Figure 3 The relation of the AP omitting proportion and accuracy of DT

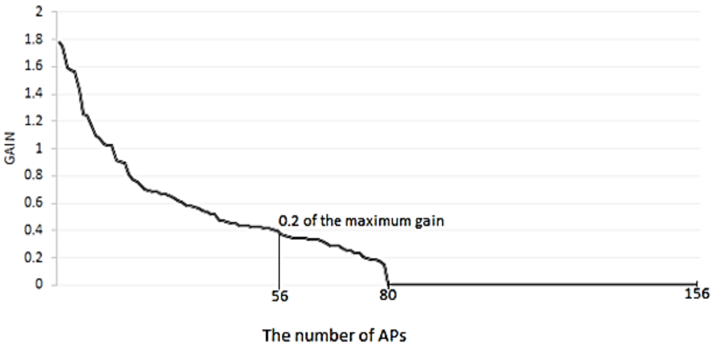

Figure 4 The ordered information gain from all AP

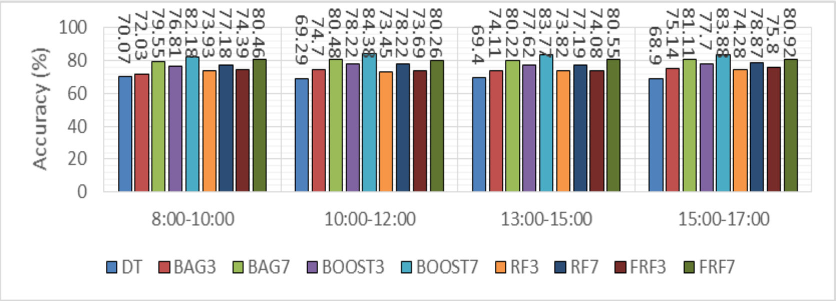

Figure 5 The mean accuracy of each algorithm in the part of the days

Table 1 Computational complexity and cost of each algorithm

	DT	*BAG*	*BOOST*	*RF*	*FRF*
big-O	$O(N*F^2)$	$O(M*N*F^2)$	$O(M*N*F^2)$	$O(M*N*F^2)$	$O(N*F)+O(M*N*\hat{F}^2)$
Cost	525,657,600	3,679,603,200	3,679,603,200	3,679,603,200	477,532,800

Figure 4 shows the ordered values of information gain calculated from equation (1) from 156 APs. According to Figure 4, the largest information gain is 1.8. We filter out the APs with low information gain by keeping the APs, whose their information gain is higher than 0.2*1.8 (threshold*the largest gain). Hence, only 56 APs are kept. Therefore 56 is the number of informative APs which is used for FRF's AP resampling.

Figure 5 shows percent of accuracy of algorithms which is averaged from 3 repetitions. From the result, time of the measurement does not affect the prediction of each algorithms. The accuracy of DT, the baseline, is 69 percent. The ensemble algorithms have accuracy higher than baseline. The proposed FRF and BAG has comparable accuracy which is higher than RF in both of 3 and 7 weak learners. The reason, FRF's accuracy is higher than RF, is filtered APs. There are smaller AP data but more useful information. However, the proposed model's accuracy is lower than BOOST, which gives the highest accuracy. Besides that, the result shows that more number of weak learners, the higher accuracies. However, the number of weak learners results in more computation complexity in the learning process.

From Table 1, the computational complexity of ensemble models is the number of weak learners (M) multiplies with the runtime of weak learners which is decision tree runtime. N is the number of sample and F is the number of AP. For computational cost calculation, parameters from real experiment are applied (N = 21,600, F = 156, $\hat{F}$ = 56, M = 7). By reducing the number of APs, the computational cost is reduced. From Table 1, the computation cost of the proposed FRF is lower than those of other algorithms. Therefore, FRF has decent accuracy while it has less complexity. FRF is suitable for applications or devices which require less complexity of algorithms.

5. Conclusions

This paper proposes the Filter Random Forest algorithms (FRF) for indoor location positioning. FRF filters APs in the area for informative APs before learning. The accuracy of FRF is better than decision tree or traditional random forest. FRF's accuracy is very similar to BAG's accuracy, but that of boosting is still higher than them. However, the computational complexity of FRF is less than those of other algorithms.

6. References

[1] Liu H, Darabi H, Banerjee P, Liu J. Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 2007;37(6):1067-1080.

[2] Mautz R. Overview of current indoor positioning systems. Geodezija ir kartogra_ja 2009;35(1):18-22.

[3] Chen RC, Lin YC, Lin YS. Indoor position location based on cascade correlation networks. 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2011 Oct 9-12; Alaska, USA. New Jersey: IEEE; 2011. p. 2295-2300.

[4] Yim J. Introducing a decision tree-based indoor positioning technique. Expert Systems with Applications 2008;34(2):1296-1302.

[5] Polikar R. Ensemble learning. In: Zhang C, Ma Y, editors. Ensemble machine learning. Berlin: Springer; 2012. p. 1-34.

[6] Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003;3:1157-1182.

[7] Quinlan JR. Induction of decision trees. Machine learning 1986;1(1):81-106.

[8] Breiman L. Random forests. Machine learning 2001; 45(1):5-32.

[9] Breiman L. Bagging predictors. Machine learning 1996; 24(2):123-140.

[10] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 1997;55(1):119-139.