A Comparison of Decision Tree Based Techniques for Indoor Positioning System

Lummanee Chanama
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
60606067@kmitl.ac.th

Olarn Wongwirat
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
olarn@kmitl.ac.th

***Abstract*— Currently, an indoor positioning system based on a fingerprint technique for wireless networks under IEEE 802.11 standard uses a method to collect a received signal strengths (RSS) to create a radio map in an offline phase. Then, it detects the RSS in an online phase to compare and find the position. However, while detecting and gathering the RSS, there are some variations of the RSS that affect the accuracy of position estimation. Therefore, there are several methods used to estimate the position in order to improve the accuracy, but the one focused in this paper is a decision tree based classification. The decision tree based classification method is found that it can provide better improvement in accuracy than the others, e.g., K-Nearest Neighbor (K-NN), Bayesian, and Neural networks. However, the techniques used to construct the decision tree are varied depending on the algorithm used to implement. Therefore, this paper is a comparison of decision tree based techniques using typical decision tree (DT) and Gradient boosted tree algorithms for estimating the position indoor. In the study, the RSSs collected from access points in the experimental area are used as the training and testing data. The decision tree models are created by using typical DT and Gradient boosted algorithms based on the training data obtained. There are two factors to consider in the comparative study, i.e., the number of training data and the number of reference radio signals. The testing results from the experiment showed that the decision tree based on Gradient boosted algorithm yielded more accurate results than typical DT, where the amount of 19 reference radio signals and 50 samples of training data gave the best result.**

Keywords—indoor positioning system; fingerprint technique; decision tree; classification method; WLAN

I. INTRODUCTION

Nowadays, a fingerprint technique used for indoor positioning system is interested among various researchers, due to its accuracy in position estimation [1]-[5]. There are two main processes used to estimate the position by the fingerprint technique, i.e., offline and online phases. The offline phase is first gathering received signal strengths (RSSs) in the environment to construct a radio map database by using their mean value. This mean value is used to represent each reference point in the area. In the online phase, the RSSs are detected at the point to identify the position on the run. They are compared with the mean values of radio map database by using a Euclidean distance (ED). The value that has the nearest distance to the point in database is used as the estimated position. In both phases, the RSS is mainly used as the variable to estimate the position. Using the RSS faces a challenge in propagation effects, especially for indoor environment. The propagation effects result in inaccuracy of position estimation. Therefore, there are several methods implemented to improve the accuracy of position estimation with minimum error.

The research works focusing on the methods used for position estimation can be found in [6]-[9]. Their work aimed to minimize the estimation error by deploying clustering and classification methods from data mining, which are our interest in this paper. In [6]-[7], the authors used a k-means clustering method to increase the K-Nearest Neighbor (K-NN) efficiency and to reduce the size of database. The K-NN algorithm compared the RSSs with database. If the compared result was the same value, or too close to each other, the algorithm could not locate the test subject, which meant the result exceeded the margin of error. The clustering method has also a drawback in computational complexity causing battery drain out easily. Thus, it might not be suitable to use in mobile devices. The research work in [8] used a decision tree based classification method to reduce computational complexity and to give high accuracy for position estimation. The multiple weighted decision tree technique was proposed. In this technique, a C4.5 algorithm was used to create a decision tree base model. Then, the adaptive boosting algorithm was employed to build the ensemble model from the decision tree created. The results expressed that the proposed method gave the high accuracy in position estimation at 2.1 m, when comparing with the 1-NN, C4.5, and bagging C4.5 algorithms. However, the proposed technique acquired the RSS data not only from the access point, but also the compass sensor, which might not be able to acquire from every device. The research work in [9] introduced the construct decision tree technique to estimate the position and to compare the result with the 1-NN and Bayesian methods. The result depicted that the presented decision tree technique gave the higher accuracy in position estimation than the other techniques at around 2.3 m for 80 sample data.

The results from [8]-[9] confirm that the decision tree based classification method provides the position estimation more accurate than clustering method. However, there are different algorithms used for constructing the decision tree model. Also, there are questions on the number of training data and reference radio signals used differently. Therefore, in this

paper, we present the comparison study for decision tree techniques using typical decision tree (DT) and Gradient boosted tree algorithms, which represent the base algorithms used in classification. The number of training data and reference radio signals are used as the factors for comparison in the study. The results will be considered to use in our future implementation.

The rest of paper is arranged as follows; Section II explains the decision tree technique used for indoor positioning system; Section III expresses the experimental set to obtain the result; Section IV gives the result analysis; Section V provides the conclusion and future work.

II. Indoor Positioning Technique By Decision Tree

A. Fingerprint Technique

There are several techniques used to locate the position indoor based on three main methods, i.e., triangulation, scene analysis, and proximity [10]. Among those methods, the RSS-based fingerprint technique under scene analysis method is found commonly used [1]-[5]. The fingerprint technique has 2 phases of operation, i.e., offline and online phases. In the offline phase, the RSSs are collected from several access points by a device in the environment. They are used to create a radio map, e.g., position (x, y) at (0, 0), (6, 0), (12, 0), …, (m, n), as in Fig. 1. In the online phase, the RSSs at the point to verify the position on the run are detected and used to compare with the database. The ED is used to find the nearest point in the database for position estimation. For instance, the nearest point at (x, y) equal to (6, 6) is estimated as the current position, as in Fig. 2.

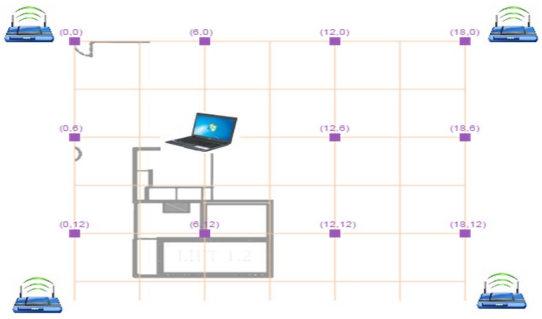


Fig. 1. Radio map creation in the offline phase.

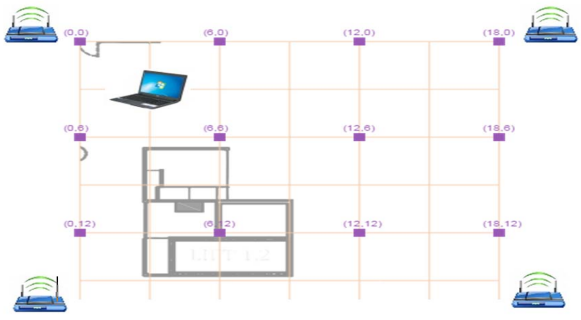


Fig. 2. Estimated position in the online phase.

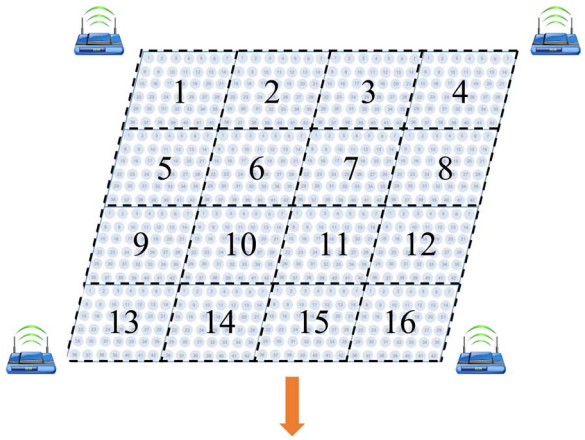
The fingerprint technique based on the ED has limitation on estimation error, depending on a number of reference points in the database. Also, the RSSs quality on the run affects the accuracy of estimation due to the propagation effects indoor. Therefore, the decision tree technique from classification method in data mining is used instead, since it provides higher accuracy and less complexity as mentioned in [8]-[9].

B. Decision Tree Based Techniques

The decision tree based techniques follow the 2 phases of operation, but use different algorithm, as follows;

1) Offline phase: The radio map is created by using the decision tree model for calssification instead. There are two main processes, as in the following sections.

a) Collecting RSSs: The RSSs are collected at the points in predefined zones. For instance, there are 16 zones and each zone contains several samples of collected RSSs, as in Fig. 3. The reference ID (Ref. ID) is used to specify each point at particular (x, y) position containing the collected RSSs, e.g., $RSSI_1$, $RSSI_2$, …, $RSSI_n$. The collected RSSs are used as the reference radio signals from access points in the area of predefined zones. These sampling points are grouped into a corresponding zone as predefined, as in Fig. 4. They are used as a training data to create the decision tree model.



Ref. ID	X	Y	$RSSI_1$	$RSSI_2$	…	$RSSI_n$
1	0	0	-76.5	-80	-55.98	-60.44
2	3	0	-80.65	-82.33	-60.96	-56.55
…	…	…	…	…	…	…
16	9	9	-70.88	-93	-50.66	-66.99

Fig. 3. Collected RSSs at each sampling point.

X	Y	$RSSI_1$	$RSSI_2$	…	$RSSI_n$	ZONE
0	0	-76	-80	-55	-60	1
1	0	-80	-82	-60	-56	1
2	0	-70	-93	-50	-66	1
…	…	…	…	…	…	…
1	9	-66	-78	-55	-80	16
2	9	-88	-56	-70	-82	16
3	9	-90	-56	-89	-70	16

Fig. 4. Grouped RSSs into a corresponding zone.

b) Create a decision tree model: There are four steps to create the decision tree model. First, arrange the training data into attributes and classified zone, i.e., $RSSI_1$- $RSSI_6$, and Zone, respectively, as in Table I.

TABLE I. TRAINING DATA

$RSSI_1$	$RSSI_2$	$RSSI_3$	$RSSI_4$	$RSSI_5$	$RSSI_6$	Zone
-81	-49	-58	-80	-84	-77	G1
-79	-49	-59	-80	-81	-74	G1
-79	-49	-58	-81	-80	-74	G1
-76	-51	-60	-86	-80	-74	G1
-80	-55	-53	-76	-90	-79	G1
-75	-52	-52	-75	-80	-73	G2
-75	-51	-55	-78	-76	-85	G2
-71	-59	-51	-77	-78	-70	G2
-77	-57	-59	-81	-83	-85	G3
-70	-56	-52	-79	-77	-77	G3
-68	-54	-50	-83	-82	-78	G3
-65	-52	-49	-76	-85	-75	G3
-78	-58	-59	-82	-83	-83	G4
-77	-59	-62	-82	-90	-79	G4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
-90	-68	-79	-84	-73	-85	G16

Second, find the root node by using information gain. The information gain is used to measure diversity of data. The root node can be found by comparing the information gain from each attribute to the entropy of classified zones [11], by using Eq. (1)-(3),

$$Gain(S, A) = E(S) - I(S, A) \tag{1}$$

$$E(S) = -\sum_i p_i \log_2(p_{i)} \tag{2}$$

$$I(S, A) = \sum_i \frac{|S_i|}{|S|} E(S_i) \tag{3}$$

where *E(S)* is the entropy of *S* data set, p_i is a relative frequency of class *i* in *S*, and *I(S, A)* is the information gain at each attribute.

The attribute where the *Gain (S, A)* is highest is chosen as the root node. For instance, suppose we are going to classify the data into two zones, e.g., G1 and G2. Choose the $RSSI_1$ as the first attribute to examine by sorting the data from the lowest to the highest in accordance with the zone, as in Table II. Consider the first data at -81 dB_m and find how many data from $RSSI_1$ that are greater than -81 dB_m and less than or equal to -81 dB_m. Then, calculate the entropy for the data at $RSSI_1$ > -81 and <= -81 by using Eq. (2). As the result, the entropy of data in $RSSI_1$ > -81 dB_m is 0.985 and $RSSI_1$ <= -81 dB_m is 0. Thus, the total entropy for $RSSI_1$ that is greater than -81 dB_m is equal to $(0.985\ \frac{7}{8}) + (0\ \frac{1}{8}) = 0.8620$. Perform this calculation recursively until the data at -71 dB_m is reached.

TABLE II. SORTING DATA FROM THE FIRST ATTRIBUTE.

$RSSI_1$	Zone
-81	G1
-80	G1
-79	G1
-79	G1
-76	G1
-75	G2
-75	G2
-71	G2

Find the two highest values from the entropy of data in $RSSI_1$, as follows,

$$E(S_1 > -76) = 0$$

$$E(S_1 > -79) = 0.4056$$

$$E(S_1 > -80) = 0.7500$$

$$E(S_1 > -75) = 0.7552$$

$$E(S_1 > -81) = 0.8620$$

$$E(S_1 > -71) = 0.9544$$

The two highest entropy from the data in $RSSI_1$ are at $E(S_1 > -81)$ and $E(S_1 > -71)$, respectively. Then, average these two values to find the information gain representative of $RSSI_1$, which is equal to -76 dB_m. Repeat the same process as $RSSI_1$ for the rest of attributes, i.e., from $RSSI_2$ to $RSSI_6$.

Then, find the average *I(S, A)* at each attribute by using Eq. (3). Once the average *I(S, A)* are found from each attribute, the root node can be obtained by using Eq. (1). As in our example, the result of attribute where the *Gain (S, A)* is highest comes from $RSSI_1$. Therefore, the $RSSI_1$ attribute will be chosen as the root node.

Third, find the split point at the root node to create the decision tree, where the split point is the average of the two highest values from the attribute, i.e., -76 dB_m in this case. The split point is found in the same manner for every node created, if any.

Finally, the decision tree model can be constructed, as in Fig. 5. If the decision tree model created from the final step cannot distinguish all the data, the steps from the second to the final will be repeated to create the child nodes recursively until all the data can be classified, or it reaches its depth limit. This will increase the depth of decision tree model.

RSSI1	Group
-81	G1
-80	G1
-79	G1
-79	G1
-76	G1
-75	G2
-75	G2
-71	G2

Fig. 5. Decision tree model for two zones classification.

2) Online phase: The online phase is used to estimate the position by using the decision tree model obtained from the offline phase. In this phase, the RSSs from access points at specified position are detected by the device, i.e., $RSSI_1$, $RSSI_2$, …, and $RSSI_6$, as in our example. The estimated position can be found by taking only the value from $RSSI_1$ and verifying whether it is greater than -76 dB_m or less than or equal to -76 dB_m. If the value is greater than -76 dB_m, the specified position is estimated to be at G2, or zone 2. Otherwise, it will be at G1, or zone 1.

By using the decision tree based technique for indoor positioning system, the computational cost can be reduced. The computational complexity depends on the depth of decision tree model, as in Fig. 6.

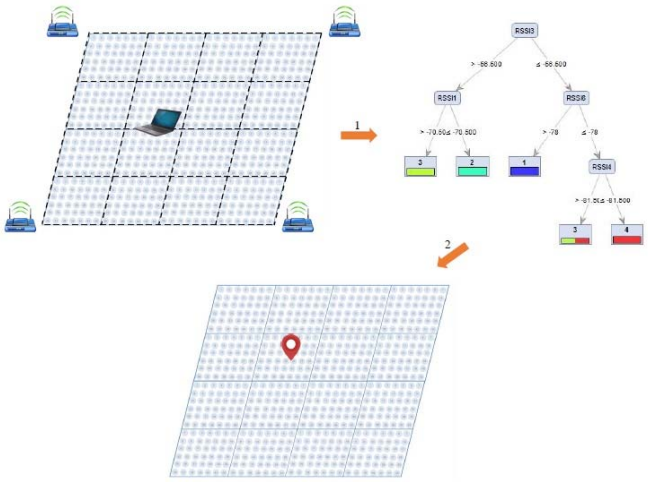


Fig. 6. Position estimation based on decision tree model.

III. Experimental Design

This section expresses the accuracy verification of decision tree model expressed in Section II. As mentioned in Section I, there are questions on which algorithm suitable to create the decision tree model is and what the number of training data and reference radio signals are. Therefore, the experiment is set in accordance with the questions that we focus. There are two algorithms used to verify in this work, i.e., typical DT and Gradient boosted algorithms. They are the main base algorithms used as in literatures reviewed. The experiment is set, as follows;

A. Experimental Environment

The area used to perform the experiment is on the ground floor of Faculty of Information Technology building at King Mongkut's Institute of Technology Ladkrabang. The size of experimental area is 18 m × 18 m, or 324 m^2 in total, as the color marked in Fig. 7. The experimental area is divided into 9 zones for classification, where each zone covers the area of 2 m × 2 m.

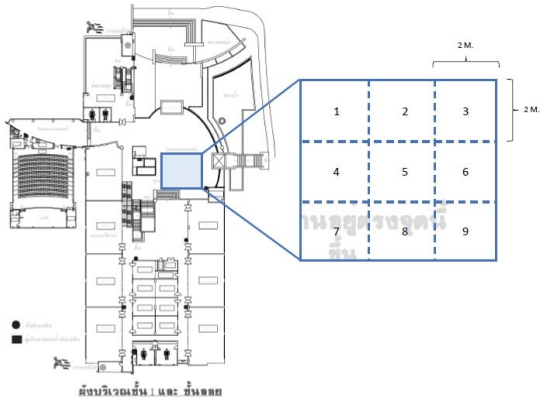


Fig. 7. Area of experimental environment.

B. Verification Factors

The number of RSSs used as the reference radio signals from various access points in the area is divided into 5, 10, and 19 reference access points, respectively. The number of RSSs used as the samples in each zone is divided into 10, 30, and 50 samples per zone. These numbers are examined and compared in order to find the proper value for future implementation.

C. Decition Tree Model Construction

To create the decision tree models, the number of RSSs from 19 reference access points is collected randomly in the 9 zones at 50 samples per zone. Therefore, there are 450 samples in the area used as the training data to create the decision tree models, where each sample contains 19 reference radio signals. To accelerate our work, RapidMiner software is used as a tool to create the decision tree models based on typical DT and Gradient boosted algorithms. Figure 8 and 9 present the decision tree models generated by RapidMiner for typical DT and Gradient boosted algorithms, respectively.

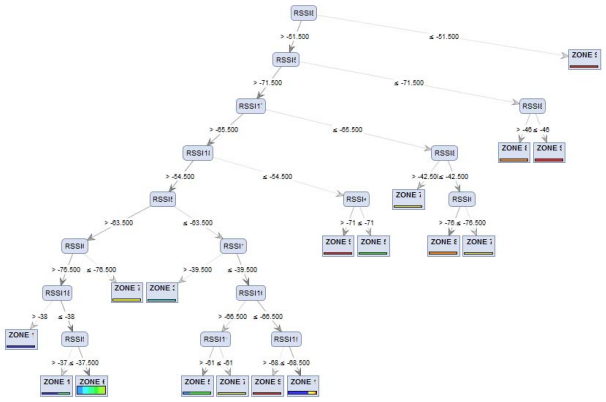


Fig. 8. Decision tree model based on typical DT algorithm.

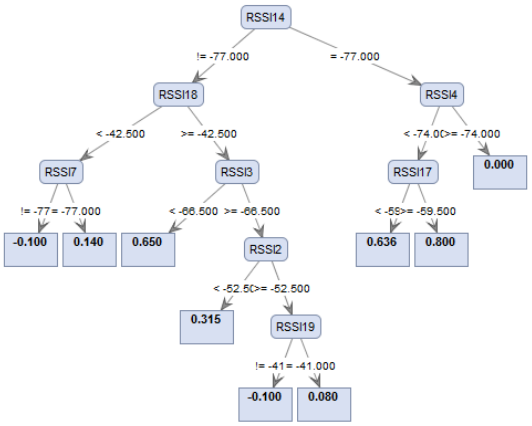


Fig. 9. Decision tree model based on Gradient boosted algorithm.

IV. Result Verification and Analysis

The verification is done by taking the RSSs randomly at 90 points in the experimental area as a testing data. Each point contains 19 radio signals from reference access points. The testing data are passed into the decision tree models to verify the accuracy of position estimation. We also construct the model to verify the results from decision tree based DT and Gradient boosted algorithms by RapidMiner. The results are examined in accordance with the verification factors explained in Section III. Figure 10 shows the verification model created by using RapidMiner.

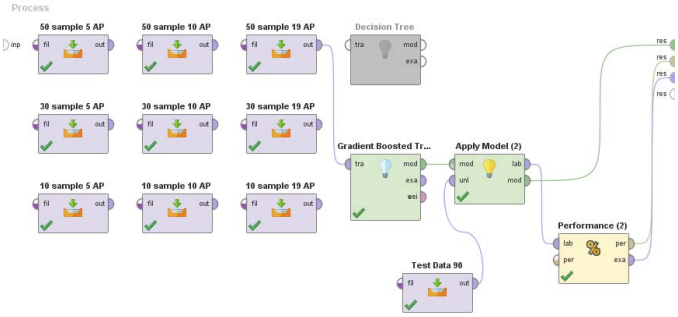


Fig. 10. Verification model for decision tree based on ID3 and Gradient boosted algorithms.

The accuracy report results taking from RapidMiner are shown in Fig. 11 and 12 for typical DT and Gradient boosted algorithms, respectively.

accuracy: 55.56%

	true ZONE 1	true ZONE 2	true ZONE 3	true ZONE 4	true ZONE 5	true ZONE 6	true ZONE 7	true ZONE 8	true ZONE 9	class precis...
pred. ZONE 1	4	0	0	0	0	1	2	0	0	57.14%
pred. ZONE 2	1	4	1	1	0	0	1	0	0	50.00%
pred. ZONE 3	1	0	0	0	0	0	0	0	0	0.00%
pred. ZONE 4	1	4	8	9	5	3	1	0	1	28.12%
pred. ZONE 5	0	0	0	0	4	0	0	0	0	100.00%
pred. ZONE 6	0	2	0	0	0	5	0	0	0	71.43%
pred. ZONE 7	3	0	0	0	1	0	5	0	0	55.56%
pred. ZONE 8	0	0	0	0	0	1	1	10	0	83.33%
pred. ZONE 9	0	0	1	0	0	0	0	0	9	90.00%
class recall	40.00%	40.00%	0.00%	90.00%	40.00%	50.00%	50.00%	100.00%	90.00%	

Fig. 11. Accuracy report for decision tree based on typical DT algorithm.

accuracy: 73.33%

	true ZONE 1	true ZONE 2	true ZONE 3	true ZONE 4	true ZONE 5	true ZONE 6	true ZONE 7	true ZONE 8	true ZONE 9	class precis...
pred. ZONE 1	5	0	0	0	0	0	1	0	0	83.33%
pred. ZONE 2	3	7	0	0	0	0	0	0	0	70.00%
pred. ZONE 3	1	0	7	0	1	1	1	0	0	63.64%
pred. ZONE 4	0	1	0	7	1	0	0	0	0	77.78%
pred. ZONE 5	0	0	3	3	7	1	1	0	0	46.67%
pred. ZONE 6	0	1	0	0	1	8	0	0	0	80.00%
pred. ZONE 7	1	0	0	0	0	0	5	0	0	83.33%
pred. ZONE 8	0	0	0	0	0	0	1	10	0	90.91%
pred. ZONE 9	0	1	0	0	0	0	1	0	10	83.33%
class recall	50.00%	70.00%	70.00%	70.00%	70.00%	80.00%	50.00%	100.00%	100.00%	

Fig. 12. Accuracy report for decision tree based on Gradient boosted algorithm.

The results in Fig. 11 and 12 express that the decision tree model based on Gradient boosted algorithm yields accuracy at 73.33%, which is higher than typical DT algorithm at 55.56%. This can be implied that the Gradient boosted algorithm provides better result for position estimation.

When taking the estimated positions obtained from the results in each verification factor to find the errors of position estimation, the results are shown in Fig. 13 and 14 as error distances for typical DT and Gradient boosted algorithms, respectively.

As in Fig. 13, when comparing the results for the number of reference radio signals at 5, 10, and 19 access points, respectively, the least error distance is found at 0.754 m for 19 access points by using Gradient boosted algorithm. For typical DT based decision tree algorithm, the error distance is found higher at 2.408 m. The error distance is increased as the number of reference signals decreased in both decision tree models, i.e., 10 and 5 access points, respectively.

As in Fig. 14, the compared result between the two algorithms for the number of samples used at each point shows that the Gradient boosted algorithm yields the least error distance at 0.754 m for 50 samples data. It is lower than the typical DT lgorithm that gives the error distance at 2.408 m. The error distance is also increased when the number of sampling data is decreased in both cases, i.e., at 30 and 10 samples, respectively.

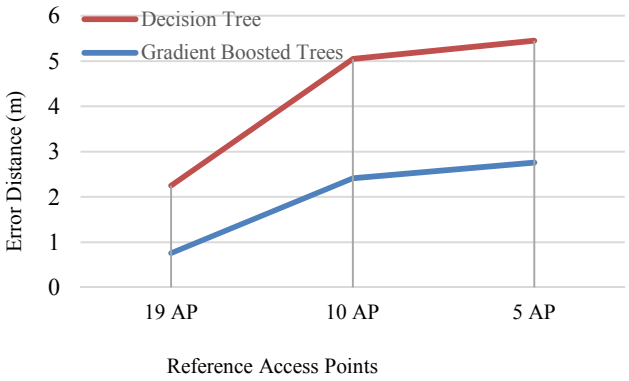


Fig. 13. Comparing result for different number of reference access points.

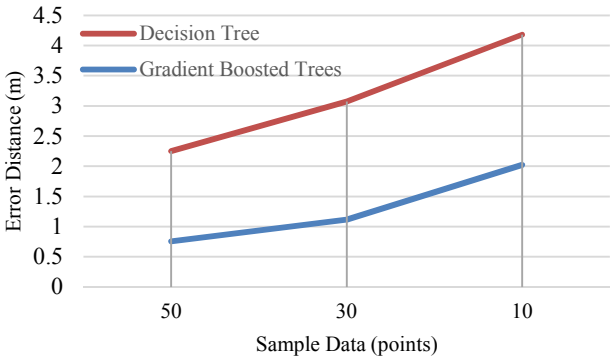


Fig. 14. Comparing result for different number of samples.

When comparing the results of decisions tree models based classification method in this work with the K-NN clustering method in [7], the k-mean clustering method provides the estimation accuracy at around 2.2 m that is lower than decision tree based models at 0.754 m. The results come out in similar fashion to the work in [9]. However, the results obtained from the two methods are verified under different environments. Thus, they cannot be clearly justified and are used to compare to notify the overall result differently.

V. Conclusion

This paper presented the comparison study of decision tree based techniques used for position estimation in the indoor environment. Typical DT and Gradient boosted algorithms were used to compare in this work in order to find the suitable algorithm for future implementation. The two algorithms were also the base common algorithms in the literatures searched. There were two factors used to verify and compare in this work, i.e., the number of reference radio signals and the number of samples of training data.

The paper explained the method used to construct the decision tree model for position estimation indoor. The experiment was set to examine the decision tree models created by typical DT and Gradient boosted algorithms. The software RapidMiner was used as a tool for construction and verification of decision tree models created. The experimental area was set and divided into 9 zones covering 324 m^2 in total. The area in each zone covered the size of 2 m × 2 m. The reference radio signals were the RSSs obtained from access points in the area of experiment. The number of reference signals used came from 5, 10, and 19 access points, respectively. The number of RSSs used in each zone was set at 10, 30, and 50 samples per zone, respectively.

The RSSs were collected in the experimental area in accordance with the number of verification factors as specified. They were used as the training and testing data. The experimental results expressed that the decision tree model based on Gradient boosted algorithm provided the estimation accuracy of 73.33%, which was higher than typical DT at 55.56%. The results also showed that, when comparing with the typical DT algorithm, the decision tree based Gradient boosted algorithm yielded the least estimation error at 0.754 m for 19 reference radio signals at 50 samples per zone. The accuracy of position estimation tended to be decreased in accordance with the number of reference radio signals and training data.

In the future work, the decision tree model based on Gradient boosted algorithm will be implemented by taking the values from the experiment in this work to be considered. The implemented model will be used in actual mobile devices to verify the performance in real. It is expected that the decision tree based technique will provide better solution for indoor positioning systems in the future.

References

[1] A. Varshavsky, D. Pankratov, J. Krumm, and E. Lara, "Calibree: Calibration-Free Localization Using Relative Distance Estimations," Proc. of 6th Int. Conf. on Pervasive Computing, pp. 146-161, 2008.

[2] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piché, "A Comparative Survey of WLAN Location Fingerprinting Methods," Proc. of 6th Work. on Positioning, Navigation and Communication, pp.243-251, 2009.

[3] H. Lemelson, S. Schnaufer, and W. Effelsberg, "Automatic Identification of Fingerprint Regions for Quick and Reliable Location Estimation," Proc. of 8th IEEE Int. Conf. on Pervasive Computing and Communications Workshops, pp. 540-545, 2010.

[4] Truc D. Le, Hung M. Le, Nhu Q. T. Nguyen, Dinh Tran, Nam T. Nguyen, "Convert Wi-Fi Signals for Fingerprint Localization Algorithm," Proc. of 7th Int. Conf. on Wireless Communications, Networking and Mobile Computing, 2011.

[5] L. Chanama, S. Yongyos, and O. Wongwirat, "A Study on Approach for Improving Indoor Positioning Method," Proc. of 8th Nat. Conf. on Information Technology, 2016.

[6] B. Altintas and T. Serif, "Improving RSS-Based Indoor Positioning Algorithm via K-Means Clustering," Proc. of 17th European Wireless, Sustanable Wireless Technologies, pp. 681-685, 2011.

[7] Ali H. Saeed and Dia M. Ali, "Indoor Wi-Fi Positioning System Based On K-means Cluster Analysis," Int. Jour. of Emerging Technology and Advanced Engineering, vol.6, 2016, pp. 248-257.

[8] S. David, "A Low Complexity System Based on Multiple Weighted Decision Trees for Indoor Localization," Sensors, vol. 15, 2015, pp. 14809-14829.

[9] Y. Jaegeol, "Introducing a Decision Tree-based Indoor Positioning Technique," Expert Systems with Applications, vol. 34, 2008, pp. 1296–1302.

[10] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of Wireless Indoor Positioning Techniques and Systems," IEEE Tran. of Systems, Man, and Cybernetics, Part C, vol. 37, no.6, 2007, pp. 1067-1080.

[11] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd Ed., Morgan Kaufmann, NY, 2012.