An Improved XGBoost Indoor Localization Algorithm

Wei Qiao, Xiaofei Kang, Mengmeng Li

ABSTRACT

This paper proposes an improved XGBoost Wi-Fi indoor positioning algorithm aiming at the accuracy problem caused by the change of environment. The method first uses Extreme Gradient Boosting (XGBoost) algorithm to establish indoor positioning model, which can achieve indoor positioning. When the environment changes, further combine error compensation (EC) method to improve the initial positioning. In addition, the positioning trajectory is compared with the actual trajectory and the unimproved positioning trajectory to verify the stability of the algorithm. The experimental results show that the 80-th percentile of the achieved accuracy is 1.11m after the change of environment, which is significantly better than the unimproved positioning algorithms based on support vector machine, random forest and extreme gradient promotion, and the obtained positioning trajectory tends to converge with the actual trajectory.

KEYWORDS
Indoor Positioning, Wi-Fi Fingerprint, Extreme Gradient Enhancement, Error Compensation, Environmental Change.

INTRODUCTION

With the rapid development of mobile Internet, location-based services are widely used, and the positioning technology that supports their core extends from outdoor to indoor, indoor positioning technology plays an important role in public safety and emergency response, indoor positioning and navigation, medical assistance, goods management, business promotion and information push[1-2]. However, in indoor environment, the non-line-of-sight (NLOS) propagation and multipath effect affect the performance of indoor positioning seriously[3-4]. Therefore, the design of high-precision, low-complexity, real-time, robust indoor positioning system has become a hot research in recent years.

At present, due to the advantages of low deployment cost and convenient signal acquisition, the fingerprint matching method of received signal strength (RSS) based on Wi-Fi has been widely used[5]. The method based on RSS fingerprint matching is divided into two stages: offline and online. In the offline phase, the RSS data and positioning coordinates of the reference points are collected and stored in the fingerprint database; in the online phase, the RSS received is matched with the data

Wei Qiao, Xiaofei Kang, Mengmeng Li, Xi’an University of Science and Technology, Xi’an 710054, China.
in the fingerprint database, and the positioning results are obtained by using the related algorithm. In recent years, with the development of machine learning algorithms[6], K Nearest Neighbor (KNN)[7], Random Forest (RF)[8], Gradient Boosting Decision Tree (GBDT)[9], Support Vector Machine (SVM)[10] have been applied to the RSS fingerprinting positioning technique, and have achieved better location results.

The positioning accuracy is largely determined by positioning fingerprint. When the structure and layout of the indoor environment change, the indoor wireless communication environment also changes, which leads to a large gap between the new environment and the established positioning fingerprint. Using the non updated fingerprint database to calculate the positioning results usually leads to a large positioning error, so it is necessary to collect RSS samples again to re-establish the positioning fingerprint or update the existing ones[11]. However, the establishment process of the fingerprint is very time-consuming and laborious. It is not economical or realistic to update all positioning fingerprints regularly and frequently, which will greatly improve the maintenance cost of RSS location fingerprinting system. Several methods to reduce the inaccuracies in location measurements are proposed in the literature[12], with the Kalman filter and particle filter being two of the most commonly used. Whereas the Kalman filter results in a linear filter which is optimal when the underlying data statistics are Gaussian, optimality vanishes when the statistics deviate from Gaussianity and the performance of the filter deteriorates.

In order to solve the non-linear relationship between signal and positioning in fingerprint database and improve the location accuracy caused by environmental changes, this paper proposes an indoor positioning algorithm XGBoost-EC based on XGBoost combined with the error compensation (EC) of polynomial regression[13]. The algorithm only needs to update fewer fingerprint database, and the error compensation technology can effectively solve the problem of positioning accuracy decline caused by environmental changes.

LOCALIZATION MODEL

In the offline stage of the algorithm, for the unupdated offline fingerprint, XGBoost is used to build a localization model to achieve the initial positioning of the location. Taking advantage of the XGBoost model's good generalization ability and parallel processing advantages, it can obtain higher positioning accuracy and operating efficiency[14]. When the environment changes, only a few fingerprint database samples need to be updated. The updated samples are used to establish the error fingerprint database, and the polynomial regression algorithm is used to build the error compensation model. Polynomial regression has the advantages of flexible handling of complex nonlinear, which can efficiently estimate the error and complete the error compensation. In the online phase, first, the XGBoost positioning model is used to obtain the initial estimated position. Then, the initial estimated position is used as an input variable to input the error compensation model that has been trained to obtain the positioning error. Finally, the positioning error is used to compensate the preliminary estimated position to obtain the final
The XGBoost-EC positioning algorithm model proposed is shown in Figure 1.

**The XGBoost Localization Mode**

For indoor environment, the fingerprint data set (the number of samples is N) \( D = \{(s_1, p_1), (s_2, p_2), (s_3, p_3), ..., (s_N, p_N)\} \), where \( s_i = [r_1^i, r_2^i, r_3^i, ..., r_K^i] \) is the vector of received \( k \)-dimensional Wi-Fi signal strength, and \( p_i = [x_i, y_i] \) is the coordinate position of the \( i \)-th reference point. The data set \( D \) can be decomposed into \( d_1 \) and \( d_2 \) for modeling respectively, namely \( d_1 = \{(s_1, x_i)\} \), \( d_2 = \{(s_1, y_i)\} \). For \( d_1 \) and \( d_2 \) data sets, the XGBoost algorithm is used to predict \( \hat{x}_i \) and \( \hat{y}_i \) respectively, so as to obtain the estimated value of physical location coordinate \( \hat{p}_i = [\hat{x}_i, \hat{y}_i] \). Taking the data set \( d_1 \) as an example, assuming that \( M \) trees have been trained, the predicted value for the \( i \)-th sample is:

\[
\hat{x}_i^{(M)} = \sum_{m=1}^{M} f_m(s_i), \quad f_m \in F
\]

Where \( f_m(s_i) \) is the predicted value of the \( m \)-th tree to the sample \( s_i \). The objective function can be modeled as:
\[ Obj = \sum_{i=1}^{n} L(x_i, \hat{x}_i^{(m)}) + \sum_{m=1}^{M} \Omega(f_m) \]  

(2)

Where \( L(.) \) represents the loss function, and \( \Omega(f_m) \) is the complexity of the m-th tree. According to the characteristics of Additive Training, the m-1-th tree is known when the m-th tree is trained, so the optimization problem of equation (2) can be expressed as follows:

\[ Obj^{(m)} = \sum_{i=1}^{n} L(x_i, \hat{x}_i^{(m-1)} + f_m(s_i)) + \Omega(f_m) \]  

(3)

Further using Taylor formula, as shown in equation (4):

\[ Obj^{(m)} = \sum_{i=1}^{n} g_i f_m(s_i) + \frac{1}{2} h_i f_m^2(s_i) + \Omega(f_m) \]  

(4)

\[ g_i = \frac{\partial L(x_i, \hat{x}_i^{(m-1)})}{\partial \hat{x}_i^{(m-1)}}, \quad h_i = \frac{\partial^2 L(x_i, \hat{x}_i^{(m-1)})}{\partial^2 \hat{x}_i^{(m-1)}}, \quad g_i, h_i \text{ are constant.} \]

In order to solve, \( f_m(s_i) \) needs to be expressed by parameterization. Therefore, we define: the value of the t-th leaf node of the m-th tree is \( \omega_t \), the number of all leaf nodes is \( T \), the leaf node of the sample \( s_i \) is \( q(s_i) \), the whole sample set of the T node of the m-th tree is \( I_t = \{ s_i | q(s_i) = t \} \). The complexity of a tree can be expressed:

\[ \Omega(f_m) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \| \omega_j \|^2 \]  

(5)

Where, \( \gamma \) and \( \lambda \) are super parameters to adjust complexity. The parameterized new objective function of formula (4) can be expressed as:

\[ Obj^{(m)} = \sum_{i=1}^{n} \left( g_i f_m(s_i) + \frac{1}{2} h_i f_m^2(s_i) \right) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \]

\[ = \sum_{j=1}^{T} \left[ \left( \sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \]  

(6)

For the fixed structure \( q(s_i) \), we can get the optimal weight of leaf node \( j \), that is, formula (7):

\[ \omega_j^* = \frac{\sum_{i \in I_j} g_i}{\left( \sum_{i \in I_j} h_i \right) + \lambda} \]  

(7)

Thus, the optimal value is obtained, as shown in formula (8):
In general, it is not possible to list all possible algorithm constructs. We use the greedy algorithm to traverse all feature partition points to find the best sharding point and tree structure, so as to complete the training of XGBoost optimal model and achieve the prediction of \( \hat{x} \). Similarly, modeling on d2 data set can achieve the prediction of \( \hat{y} \), thus completing the estimation of physical location coordinates \( \hat{p}_i = [\hat{x}_i, \hat{y}_i] \).

When the environment changes, the precision of indoor positioning model based on XGBoost will decrease. In order to improve the positioning accuracy, a polynomial regression error compensation algorithm is proposed, which only need update a few reference points, the positioning accuracy can be improved greatly.

**Error Compensation Model**

For a changing environment, a small amount of fingerprint data is sampled to build a new data set \( S = \{ (\mathbf{p}_i, \text{error}_i), (\mathbf{p}_j, \text{error}_j), (\mathbf{p}_k, \text{error}_k), \ldots, (\mathbf{p}_n, \text{error}_n) \} \), where \( \mathbf{p}_i = [\overline{x}_i, \overline{y}_i] \) is defined as the average of the estimated of the coordinates of the physical position of the i-th reference point through the XGBoost algorithm, namely:

\[
\overline{p}_i = \frac{1}{N} \sum_{t=1}^{T} \text{FP}(p_i, t)
\]  

\( \text{FP}(\bullet) \) indicates that the estimated position measured \( \hat{p}_i \) by the non updated XGBoost positioning model at time t under coordinates \( p_i = (x_i, y_i) \). T is the whole observation period, N is the number of times observed during this period. \( \overline{p}_i = (\overline{x}_i, \overline{y}_i) \) is the average of the estimated position abscissa over a period.

Where \( \text{error}_i = (\Delta x_i, \Delta y_i) \) is defined as the error vector between the average of the estimated value and the true value along the X-axis and Y-axis of the position of the i-th reference point, namely:

\[
\Delta x_i = \overline{x}_i - x_i
\]  

\[
\Delta y_i = \overline{y}_i - y_i
\]

Using the polynomial regression model, the error function can be modeled as:

\[
\Delta x_i \approx Q(\overline{x}_i, \overline{y}_i) = a_n \overline{x}_i^n + a_{n-1} \overline{x}_i^{n-1} + \ldots + a_1 \overline{x}_i + b_m \overline{y}_i^m + b_{m-1} \overline{y}_i^{m-1} + \ldots + b_1 \overline{y}_i
\]  

\( (8) \)
\[ \Delta y_i \approx R(\bar{x}_i, \bar{y}_i) = d_i \bar{x}_i' + d_{i-1} \bar{x}_{i-1}' + ... + d_1 \bar{x}_1' + e_i \bar{y}_i' + e_{i-1} \bar{y}_{i-1}' + ... + e_1 \bar{y}_1' \]

(13)

For data sets \( S_i = \{ (p_i, \Delta x_i) \} \) and \( S_i = \{ (p_i, \Delta y_i) \} \), polynomial regression model is used to estimate error \( \Delta \hat{x}_i \) and \( \Delta \hat{y}_i \) respectively, and further improve the estimation accuracy through error compensation, namely:

\[ \hat{x}_i' = \bar{x}_i - \Delta \hat{x}_i \]

(14)

\[ \hat{y}_i' = \bar{y}_i - \Delta \hat{y}_i \]

(15)

So far, after the XGBoost algorithm and the error compensation (EC) algorithm of polynomial regression, the estimated value of the i-th physical location coordinate \( p_i \) in the indoor plane is \( \hat{p}_i' = [\hat{x}_i', \hat{y}_i'] \).

EXPERIMENT

This experiment was conducted in the underground parking, as shown in Figure 2. The indoor positioning area includes the lane (40m × 5.5m) and some parking spaces (12m × 4m). The reference points are arranged at a distance of 0.5m, and a total of 972 points are evenly distributed. At the experimental environment, we held a mobile phone, collected fingerprint signals multiple times at each reference point, and then averaged them to reduce noise interference. The application for collecting fingerprint data is shown in Figure 3. The data collected from each point is stored in the document as a training set \( D \).

Figure 2. Indoor positioning experiment environment.
Simulation Results

After a lapse of two months, some facilities were added to the underground parking. The localization of the underground parking was tested again, and the positioning accuracy has dropped dramatically. In the previous positioning area, a part of the reference points are evenly selected as the new fingerprint training database. The positioning error is obtained through the XGBoost positioning model, a database S of coordinates and errors is established. Randomly select test points for data collection, and randomly walk a path to test the accuracy of the positioning model.

Through data simulation, it is concluded that the location of different algorithms of the unupdated fingerprint database is compared with the positioning accuracy of XGBoost-EC, as shown in Figure 4. It can be seen from the figure that the positioning accuracy of the previous fingerprint database has decreased significantly under the circumstances of environmental changes. But the localization algorithm based on XGBoost-EC has good localization performance, and the obtained positioning accuracy can be effectively controlled within 1.11m at the 80-th percentile, which is respectively higher than KNN of 4.35m and SVM of 3.22m, RF of 2.61m, GBDT of 2.23m, XGBoost of 1.83m positioning accuracy performance increased by 74.48%, 65.53%, 57.47%, 50.22%, 39.34%. Therefore, the revised algorithm is significantly more accurate than the uncorrected algorithm. In addition, the running time is also an important indicator to evaluate the positioning algorithm. Table I shows the average positioning accuracy of several positioning algorithms based on machine learning and the comparison of the running time of each algorithm. It can be seen that the average accuracy of location algorithm based on XGBoos-EC is 0.79 m, which is 75.2%, 67.2%, 54.3%, 51.8%, 42.7% higher than that based on KNN (3.19 m), SVM (2.41 m), RF (1.73 m), GBDT (1.64 m) and XGBoost (1.38 m). Combined with the running time and average positioning accuracy of each algorithm, it can be seen that the XGBoos-EC positioning algorithm performs well in real time.
TABLE I. COMPARISON OF ACCURACY AND RUNNING TIME OF SEVERAL POSITIONING ALGORITHMS.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average Positioning Accuracy</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>3.19m</td>
<td>16ms</td>
</tr>
<tr>
<td>SVM</td>
<td>2.41m</td>
<td>209ms</td>
</tr>
<tr>
<td>RF</td>
<td>1.73m</td>
<td>266ms</td>
</tr>
<tr>
<td>GBDT</td>
<td>1.64m</td>
<td>157ms</td>
</tr>
<tr>
<td>XGBoost</td>
<td>1.38m</td>
<td>142ms</td>
</tr>
<tr>
<td>XGBoost-EC</td>
<td>0.79m</td>
<td>148ms</td>
</tr>
</tbody>
</table>

Figure 4. Comparison of the positioning accuracy of different algorithms.

Figure 5 shows a comparison of the trajectories walking in the positioning area, where Real Trace is the real trajectory route and the XGBoost is the unupdated fingerprint database, using the XGBoost model for the trajectory; XGBoost-EC is a trajectory route based on XGBoost using an error compensation model. After calculation, in this simulation environment, the uncorrected coordinate error is 1.66m, and the corrected coordinate error is 1.03m. Through comparison, it is found that the polynomial regression model can find the nonlinear mapping relationship between the positioning coordinates and the positioning errors well, and has a certain correction effect on the positioning coordinates.
In order to verify the stability of the XGBoost-EC algorithm, after the environment changes, Gaussian white noise with a variance of 2 is added to 2, 6, and 10 APs respectively, and the positioning performance is shown in Figure 6. It can be clearly seen from the figure that XGBoost-EC can still improve the positioning accuracy under the condition that different APs increase noise, and the environmental change has little effect on the accuracy and is very robust.

Figure 6. Comparison of changes in noise positioning accuracy by the different AP.
CONCLUSION

In this paper, we presented an improvement on the XGBoost algorithm, which not only enhances the accuracy of indoor positioning system, but also solves the accuracy problems caused by environmental changes. The proposed technique is batch processing method but with a fixed complexity. And independent of the underlying statistics of the observed processes and measurements, to the extent that they are superior to other methods of rebuilding a fingerprint database. Moreover, the experimental results show that the proposed algorithm still achieves a good positioning effect even in the case of environmental changes, which is significantly better than other positioning algorithms; the obtained positioning path is closer to the actual path; and it shows good robustness.

ACKNOWLEDGMENTS

Corresponding author: Wei Qiao
The work was supported by The Natural Science Foundation Program of China: (Grant: 61801372).

REFERENCES