Clustering-assisted Stacked Generalization Approach for Fingerprint Crowdsourcing-based Indoor Localization

Changhyun Lee, Yonghun Kim, and Kiseon Kim
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, KOREA
{chlee1016, mizpah, kskim}@gist.ac.kr

Abstract

Currently, crowdsource based Wi-Fi fingerprinting indoor localization is getting more attention, because the system does not require time-consuming and labor-intensive site survey. However, a performance degradation problem caused by RSS variation between training and test data is still a challenging issue of crowdsource based Wi-Fi fingerprinting technique. In this paper, we propose a clustering-assisted stacked generalization approach (C-StackLoc), combining clustering and ensemble technique for mitigating the effect of the RSS variation. We compare the proposed C-StackLoc with StackLoc in which clustering is not included, to confirm that clustering can improve the localization accuracy. In addition, we see that the C-StackLoc can achieve better localization accuracy compared to existing algorithms.

I. INTRODUCTION

Nowadays, Wi-Fi fingerprint crowdsourcing can easily collect the fingerprint data from crowds with reducing the burden of labor and time for site survey [1]. Since the crowdsource based data is collected from different time of the day, placement and direction of the device, RSS variation problem between training and test data tends to be larger than the data from the site survey [2]. Therefore, mitigating the impact of RSS variation is important for designing an accurate indoor localization system, when using crowdsource based fingerprinting data.

Recently, many researches have dealt with the variation of training and test data by using ensemble and clustering [3]. B. Akram et al. used the soft clustering features to train each prediction model separately, and they applied voting ensemble to gather the prediction values [4]. However, if the fingerprint data does not belong to multiple clusters, the ensemble strategy does not be applied. Therefore, the clustering features need to be used in a different way.

In this paper, we propose a clustering-assisted stacked generalization ensemble approach for indoor localization (C-StackLoc) which uses the clustering results as additional features of the ensemble. The simulation results show that the proposed localization algorithm can achieve better localization accuracy compared to the existing algorithms.

II. PROPOSED ALGORITHM

Because the performance degradation is caused by the RSS variation problem between training and test data, model generalization is a major issue to prevent the model from overfitting to training dataset. This is a key motivation of our approach for using stacked generalization ensemble. In addition, we apply K-means clustering and use the clustering features as additional features, to improve HybLoc [4]. In this chapter, we briefly describe the existing algorithm and introduce a clustering-assisted stacked generalization ensemble algorithm.

1. Brief description of existing algorithm

Figure 1 describes the system design of HybLoc [4]. This algorithm uses soft clustering to determine which clusters the fingerprint data belongs to. After that, each prediction model is trained separately, by using the soft clustering features. However, the ensemble cannot be applied when the fingerprint data does not belong to multiple clusters.

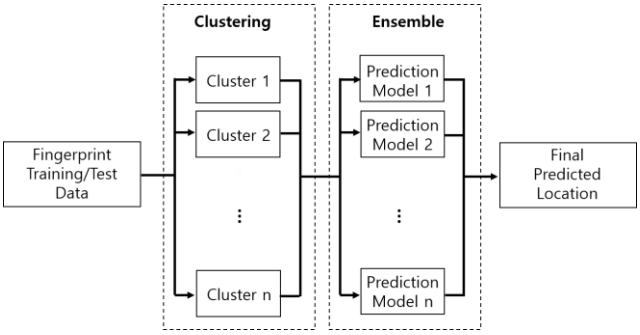

Figure 1. Illustration of HybLoc [4]

2. C-StackLoc

The illustration of C-StackLoc is shown in Figure 2. Unlike HybLoc [4], we combined the ensemble and clustering in parallel to avoid the situations where the ensemble does not be applied. To make this possible, we feed the results of ensemble prediction and additional clustering features into Generalizer, with generating final predicted location. Another reason for using the clustering features in parallel with ensembles is that the variation of the RSS can be moderated by additional clustering features because the fingerprint data classified into the same cluster are regarded as identical in clustering features, even if their fingerprint of RSS vector is little different. Therefore, more generalized prediction can be made mitigating the effect of performance degradation caused from the RSS variation.

The difference between StackLoc and C-StackLoc is the usage of clustering features. Intuitively, the closer the

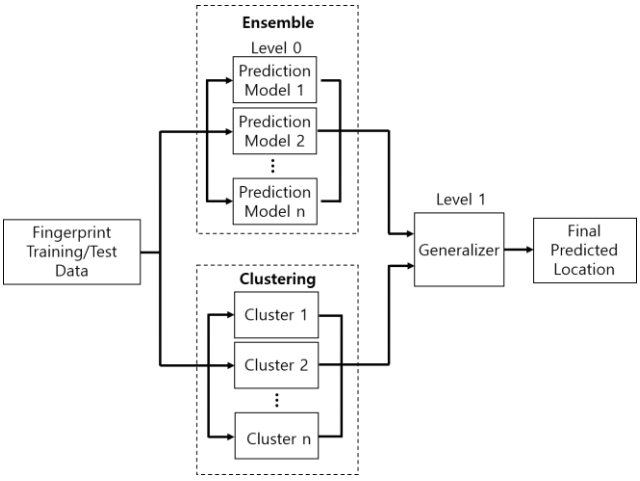

Figure 2. Illustration of proposed C-StackLoc

Algorithm	Mean Absolute Error (m)		Standard Deviation of Error (m)		Time (sec)	
	Positive	*Exponential*	*Positive*	*Exponential*	Training	Testing
HybLoc [4]	8.95	9.05	7.73	7.84	34.69	0.13
StackLoc [7]	8.33	8.52	7.25	7.74	177.66	130.07
C-StackLoc (Proposed)	**7.99**	**8.24**	**7.13**	**7.69**	179.23	130.93

Table 1. Performance comparisons between proposed C-StackLoc and existing algorithms on EU ZENODO database.

data samples are, the more likely it is to be classified into the same cluster because of the RSS vectors similarity. Therefore, the clustering features can give the Generalizer useful constraints for predicting location. The C-StackLoc is expected to improve the localization accuracy without additional location information, even if it requires more computation time.

The ensemble technique with Generalizer is called stacked generalization [5]. Stacked generalization is known as effective at correcting the error of each level 0 prediction model. The most prominent features of stacked generalization is the usage of Generalizer, trying to learn which prediction models are the reliable ones.

K-means clustering is used to generate additional clustering features. The objective of k-means clustering is to find centric points that minimize the sum of squares between each of cluster component points and the centric points represented as (1).

$$arg\min_{c}\sum_{n}\sum_{X_i \in c_k}\|X_i-\mu_k\|^2 \qquad (1)$$

where $\boldsymbol{X}$ is a set of 3D coordinates and **c** is a set of clusters, $\boldsymbol{\mu}$ is the mean of the points in the clusters **c** and **n** is the number of clusters.

III. SIMULATION RESULTS

We make use of the openly available EU ZENODO database, which is a Wi-Fi crowdsource based fingerprint database [6]. This dataset is composed of 697 training data and 3951 test data collected with 21 devices in a university building in Tampere, Finland. We used 5-fold cross validation for making meta data, and the missing value was replaced with -110dBm. K-NN and XGBDT were utilized as components of stacked generalization.

Figure 3 represents the CDF error comparison of proposed C-StackLoc with HybLoc [4] and stacked generalization based localization (StackLoc) algorithm [7]. The CDF error represents the probability that the error is less than a certain value. In the figure, C-StackLoc showed higher probability that the error is less than a certain value in all ranges compared to StackLoc. In addition, the most noticeable performance difference between C-StackLoc and HybLoc can be seen in the range below 10 meters. Specifically, C-StackLoc recorded about 40% probability that the error is less than 5 meters, while HybLoc does not reach to 40% probability.

Table 1 shows the performance comparisons of MAE, SDE and computation time. As we can find in Table 1, the proposed C-StackLoc showed the best performance with recording 7.99m MAE and 7.13m SDE in positive data representation. The comparison between StackLoc and C-StackLoc indicates that clustering features can improve the performance by giving Generalizer useful constraints. In case of the computation time, training and testing C-StackLoc takes longer than HybLoc. It is because the time for training and testing C-StackLoc depends on the number of cross validation folds and training and testing time of models which are components of the C-StackLoc.

IV. CONCLUSION

In the Wi-Fi fingerprint based indoor localization, fingerprint crowdsourcing technique has replaced the time-consuming and labor-intensive site survey. This paper proposed a clustering-assisted stacked generalization approach to improve localization performance by reducing the effect of RSS variation between training and test data. This paper showed that proposed C-StackLoc improved localization accuracy in MAE and SDE. This result suggests that combining clustering and stacked generalization ensemble can improve the performance of the recent indoor localization algorithm.

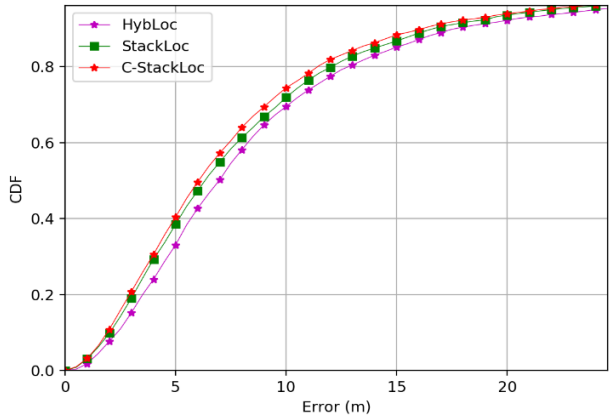

Figure 3. The CDF error of existing and proposed localization algorithms in positive data representation

ACKNOWLEDGEMENT

This research was a part of the project titled 'Development of Automatic Identification Monitoring System for Fishing Gears', funded by the Ministry of Oceans and Fisheries, Korea.

REFERENCES

[1] B. Wang, Q. Chen, L. Yang, and H. Chao, "Indoor smartphone localization via fingerprint crowdsourcing: Challenges and approaches," *IEEE Wireless Commun.*, vol. 23, 2016.

[2] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y.Chen, and F.Ye, "Push the limit of WiFi based localization for smartphones," in *Proc. ACM MobiCom*, 2012.

[3] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han, "A Graph-Based Consensus Maximization Approach for Combining Multiple Supervised and Unsupervised Models," *IEEE Trans. Knowl. Data Eng.*, 2013.

[4] B. Akram, A. Akbar, and O. Shafiq, "HybLoc: Hybrid Indoor Wi-Fi Localization Using Soft Clustering-Based Random Decision Forest Ensembles," *IEEE Access*, vol. 6, 2018.

[5] D. Wolperk, "Stacked generalization", *Neural Networks*, vol. 5, 1992.

[6] E. Lohan, J. Torres-Sospedra, H. Leppakoski, P. Richter, Z. Peng, and J. Huerta, "Wi-Fi crowdsourced fingerprinting dataset for indoor positioning," *Data*, vol. 2, 2017.

[7] C. Lee, Y. Kim, and K. Kim, "StackLoc : Stacked Generalization Approach for Wi-Fi Fingerprint-Based Indoor Localization," *KICS*, 2019.