Position Quantization Approach with Multi-class Classification for Wi-Fi Indoor Positioning System

Werayuth Charoenruengkit, Sunisa Saejun, Ramunya Jongfungfeuang, Kewali Multhonggad
Department of Computer Science, Faculty of Science
Srinakharinwirot University
Bangkok, Thailand

Abstract—**Indoor positioning system is a challenging problem due to the variety of environment and unreliable of data that are used for a prediction of the position. For Wi-Fi based indoor positioning system, signal intensity used to predict the co-ordinate of the device are known to fluctuate greatly despite being measured at the same position. Therefore, significant errors are often found when solving this problem with regression algorithms. A quantization of co-ordinate data into position IDs can mitigate the fluctuated noises in the data and is able to reformulate the problem into a multi-class classification problem. The error in positioning can then be computed from the distance between the true co-ordinate and the predicted co-ordinate. The experiment shows that Random forest classification can predict the position with the error in positing at 5.65 meters on average when the quantization is applied with threshold setting to 1 meter.**

Keywords—K-Nearest neighbor; Gaussian Naive Bays; Support vector machine; Random forest; Indoor positioning, Quantization

I. Introduction

Indoor positioning system is an application that attempt to locate a person or a device in an indoor environment. Unlike, the GPS technologies used in the outdoor environment, the indoor positioning problem that requires more sophisticated and reliable techniques to accurately locate a person or a mobile device within a building. Much research has been investigated on using different sensor technologies and a combination of sensor technologies. However, there is still no single solution to this problem that can work reliably in all indoor environment [1]. Most Indoor positioning applications are based on a measurement of the power of the signal between transmitters and receivers, which is called the Received Signal Strength Indicator (RSSI). The popular approaches are Bluetooth [2], ZigBee [3], and Wi-Fi [4][5]. Recent developments have much interest on the indoor positioning system based on a measurement of RSSI from iBeacons and WIFI from mobile devices due to a wide availability of sensors on the mobile phones. However, the reading of the RSSI is often fluctuated greatly despite being measured at the same position and from the save mobile phones or access points.

Early methods for indoor positioning problems utilize the trilateration techniques that have been maturely developed for the outdoor positioning method. An estimated position is calculated using a basic mathematic calculation based on the angle intersection of three spheres or four spheres. This approach estimate the position based on the known positions of at least 3 transmitters (or receivers). Although these technique does not achieve much accurate result due to the unreliable of measured signal [6], a combination of mobile device sensors can help improve its positioning estimate [7].

Another popular approach is the Fingerprinting techniques, which is based on the unique characteristic of the measured signals at each indoor position. For this approach, the referenced RSSI of each position are collected into a database during an offline phase (training data). These RSSI are then compared with the RSSI measured during the online phase. The position is determined corresponding to the position with the smallest distance between the RSSIs. Nevertheless, the fingerprint method suffers the similar challenge because of the large variance of RSSI being measured; resulting in the low accuracy [5]. In addition, a large storage is required to build a fingerprint database from all reference data.

This paper attempts to apply the fingerprinting techniques armed with a quantization approach to the co-ordinates of the indoor positions. Each position is assigned by a position id, corresponding to classes in the multi-class classification. The dataset used for the experiment is the Wi-Fi data collected from 520 access points, which have been collected for research use by Jaume I University. These data are publicly distributed as UJIIndoorLoc Data Set [8]. The next section presents a quantization technique and describes the evaluation methods used in the experiments. Section III describes the data preparation process and illustrates the results from the experiments. The summary is discussed in IV.

II. Methods

This paper formulates the indoor positioning into a multi-class classification by a quantization of co-ordinates. Each class represents a set of position in the buildings that has been created based on the quantized co-ordinates. The process includes data cleaning, data transformation, dimension reduction. The experiment are conducted by evaluating the data with four classification algorithms. The following subsections describe the quantization technique and classification algorithms.

A. Co-ordinate Quantization

The idea of the proposed co-ordinate quantization is similar to that of vector quantization that has been used in communication and signal processing to represent the dataset in p-dimensional space into K disjoint regions such that the centroid of each K region represent the all its member data [9][10]. The proposed quantization in this paper does not iterate through the centroid finding, but rather use the sorted distance as an approach to select the centroid. The algorithm is explained below:

1) Find distance d_i of co-ordinate i from a reference co-ordinate for all i in the N-record dataset.

2) Sort the d_i in a accendance order (d_0, d_1,…, d_N)

3) Assign a position id to record with distance d_0.

4) If $d_1 - d_0 > t_0$, assign a new position id to the record 1st, otherwise, assign the position id and distance d_0 to record 1st.

5) Repeat 4) for the record 2nd and for all N records.

This algorithm reduces the inevitable noises in the processing RSSI. The defined threshold, t_0, determines the maximum quantization error. If t_0 is at a small value, the quantization error will also be small. However, reducing the t_0 will increase the number of classes and classification will be more challenge. Experiments are conducted to determine heuristically the value of t_0 that achieves the best position prediction accuracy. The Euclidean distance can be used to compute the distance:

$$d_i = \sqrt{(lat_i - lat_0)^2 + (lgn_i - lgn_0)^2 + (\beta * fl_i)^2} \quad (1)$$

, where lat_i, lgn_i, and fl_i are the latitude, longitude, and floor of the data record I, respectively. The lat_0 and lgn_0 are the latitude and longitude of the reference point, which can be the mean average of the dataset or any selected reference point. β is the height from ground of each floor in the building, which is set to 5 meters in the experiment.

B. K-Nearest neighbor's classification

K-nearest neighbor (KNN) classification is based on fingerprinting techniques, where the training data are used as database of reference values for each position [11]. During the evaluation, predicted position is the smallest distance of the evaluating features to those reference values in the database. Given a feature vector of an evaluating signal intensity $v_a = (v_1, v_2, \dots, v_n)$, the predicted position is where the smallest Euclidean distance $d_a(v_a, v_i)$ among all i in the fingerprint database:

$$d_a(v_a, v_i) = \sqrt{\sum_{j=0}^{n} (v_a - v_{ij})^2} \quad (2)$$

,where v_i is the feature vector of reference values in the database and v_{ij} is the value of element j of the vector. If only the smallest distance is desire, the algorithm is called 1-KNN, which is K=1.

The algorithm can implement where k is greater than 1, in which case, a voting mechanism determines the final prediction among the K smallest distances. This 1-KNN is chosen for our result comparison because it has been used as the baseline in [8].

C. Gaussian Naive Bays classification

Gaussian Naive Bays (GNB) classification is based on the Naive Bayes [12], which has been successfully used in many classification problem due to its simplicity and low computation effort. It models each feature of dataset as a random variable that has been independently generated from a certain distribution as follows:

$$p(v_a|C) = \prod_{j=0}^{n} p(v_j|C) \quad (3)$$

,where the value $p(v_a|C)$ is the probability of a feature vector of an evaluating signal intensity given that it is at the location C.

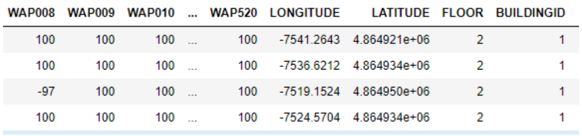
The distribution of the continuous feature data are commonly assumed to be a normally distributed process. However, other distribution has also been applied to Naive Bay classification as well [13]. Although the dataset used in this paper are not strictly normal, a transformation has been applied to the data so that normally distributed assumption can be used.

D. Support vector classification

Support vector classification (SVC) is another technique commonly used to divide dataset into two classes using hyperplane, which defines decision boundaries between the classes. Much research have shown that support vector machine (SVM) classifiers often have superior accuracy in comparison to other classification methods [14]. The hyperplane is created from the training data that has the largest distance between the closest dataset to the hyperplane. The SVC has also been applied to multi-class classification in many pattern recognition application [15]. In this paper, the one-vs-one scheme is used for the multi-class classification. The implementation is based on the libsvm used in the scikit-learn machine learning libraries [18].

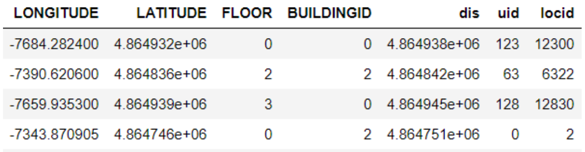
E. Random forest classification

Random forest classification is an improvement of Decision tree based classification, which is a popular technique for many classification problems. The decision tree technique splits dataset recursively into two partitions of homogeneous or near-homogeneous terminal nodes [16]. The technique often suffers from overfitting problems due to suboptimal decision for the ends of splitting. Random forest (RF) classification mitigates the overfitting due to having only a single decision tree by creating multiple decision trees. The output categories are determined by the equal-weight voting of decision tree classification results [17]. The technique uses two random selection processes. The first process is by random selection of training dataset to build each decision tree. The second process is the random selection of the characteristics attributes of the sample. The purpose of this two random processes is to reduce the variance of the estimate similar to the ensemble average of multiple estimates.



WAP008	WAP009	WAP010	...	WAP520	LONGITUDE	LATITUDE	FLOOR	BUILDINGID
100	100	100	...	100	-7541.2643	4.864921e+06	2	1
100	100	100	...	100	-7536.6212	4.864934e+06	2	1
-97	100	100	...	100	-7519.1524	4.864950e+06	2	1
100	100	100	...	100	-7524.5704	4.864934e+06	2	1

Fig. 1. Example of the dataset



LONGITUDE	LATITUDE	FLOOR	BUILDINGID	dis	uid	locid
-7684.282400	4.864932e+06	0	0	4.864938e+06	123	12300
-7390.620600	4.864836e+06	2	2	4.864842e+06	63	6322
-7659.935300	4.864939e+06	3	0	4.864945e+06	128	12830
-7343.870905	4.864746e+06	0	2	4.864751e+06	0	2

Fig. 2. The new prediction target locid that is created from vector distance, floor, and building id.

III. Experiment Results

The experiments consist of data preparation process and classification process. Both steps are conducted using python 3 and scikit-learn machine learning libraries [18]. The evaluation of classification results are compared among four classification algorithms: KNN, GNB, SVC, and RF. For GNB, SVC (radial basis function kernel), and 1-KNN, the default parameters provided from the sklearn's library is used for the classification. The default parameters are also used for RF, but the number of trees (n_estimators) is set to 50.

A. Data preparation

The dataset used in this paper is the UJIIndoorLoc Data Set, which are collected from three buildings of Jaume I University. Each building has four or five floors covering 108703 square meters. The data were collected from more than 20 different users and 25 different models of mobile devices. The database contains approximately 21,000 records. Each record has 529 attributes consisting of the RSSI from 520 wireless access points (WAPs), the coordinates of mobile devices, and other useful information [8]. An example of dataset is shown in Fig. 1. The WAPyyy shows the RSSI measured at the access point number yyy, in which the value of 100 indicates that the record does not detected by the WAP and value of -97 indicates that the RSSI measured at -97dB. The longitude and latitude coordinates have a unit in meters with UTM from WGS84 [8].

The goal of this paper is to predict the position based on the co-ordinate that has been quantized. The mulit-class classification techniques are used to determine the position, representing by an ID. The number of classes are varied depending on the threshold t_0 mentioned in IIII.A. The experiment are investigated for $t_0 = 0.5$, 0.75, and 1. The average quantization errors and number of classes as shown in TABLE I. A trade-off between the number of classes and the quantization error must be determined since the consequence of the small quantization error is the higher number of classes, leading to more difficult for the classification. From the table, the average quantization errors from the dataset are 0.33 meter when setting the t_0 to 1 meter. This setting creates dataset with 584 classes.

TABLE I. Average Quantization Errors and the Corresponding Number of Classes for Each Threshold

Threshold (t_0)	Average quantization error (m)	Number of classes
0.5	0.13	738
0.75	0.23	652
1	0.33	584

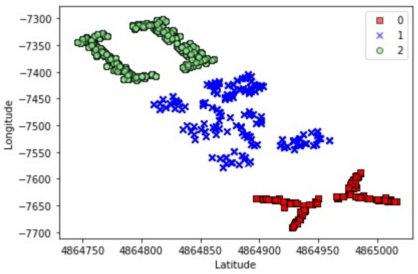


Fig. 3. The positions of the measurement points from the dataset

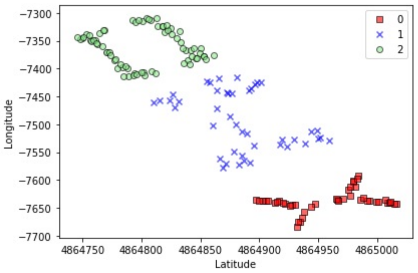


Fig. 4. The 584 quantized positions of the measurement points obtained from setting $t_0 = 1$

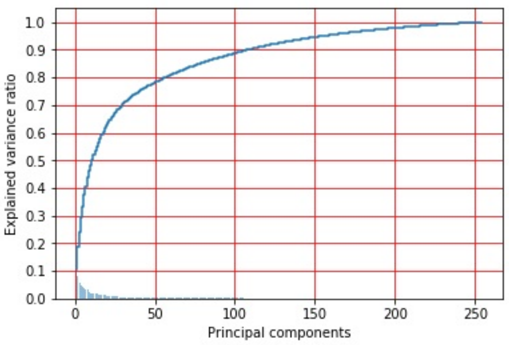
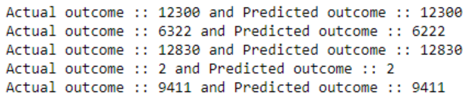


Fig. 5. Explained variance for eigenvalue

For each position with a distance of t_0 or less in radius, the same uid is assigned by the algorithm introduced in IIII.A. Without the loss of generality, the distance is computed based on (1) using the origin (0,0) as the reference point for convenience. To make the data more interpretable, the floor id and building id is concatenated to the uid forming the locid. For example, locid 12300 is the location of the uid at 123, floor 0, and building 0. Fig. 2 shows the new target feature called locid, which will be used as a class label for the classification algorithms. Fig. 3 depicts the positions of the measurement points from the dataset whereas Fig. 4 displays the resulting 584 quantized positions of the measurement points obtained from setting $t_0 = 1$.

Further data preparation process follows the investigation done by [19] in which the box-cox transform is applied to the dataset in order to obtain normal distributed characteristic of the dataset [20]. This transform is necessary for the Gaussian Naive Bays classification. The Principal Component Analysis (PCA) is applied to reduce the dimension of the features that are highly correlated [21]. As displayed in Fig. 5, using PCA, 95% of the variance can be explained by the top 150 eigenvector, resulting in reduction of 520 features to 150 features to be used for subsequent analysis.

Preliminary classification experiments have been investigated with different value of t_0. For t_0 less than 1 using the four classification algorithms, all algorithms achieve the accuracy merely less than 50%. As a result, the final dataset presented in the next section are those obtained from setting t_0 to be 1. The data are, then, divided into two parts for evaluations, which are 17,874 training records and 1,987 testing records. Each record consists of 150 features (WAPs) and 1 target labels (locid). The 584 classes (584 unique locid) are to be predicted based on the dataset.



```
Actual outcome :: 12300 and Predicted outcome :: 12300
Actual outcome :: 6322 and Predicted outcome :: 6222
Actual outcome :: 12830 and Predicted outcome :: 12830
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 9411 and Predicted outcome :: 9411
```

Fig. 6. An example of classification results obtained from RF classification

TABLE II. THE CLASSIFICATION ACCURACY IN PERCENT AND AVERAGE ERROR IN POSITIONING IN METER FOR EACH CLASSIFICATION ALGORITHMS

Classification Techniques	**Classification Accuracy**	**Average Error in positioning (m)**
1-KNN	67.49%	6.18
GNB	57.57%	9.24
SVC	62.71%	6.85
RF	68.50%	5.65

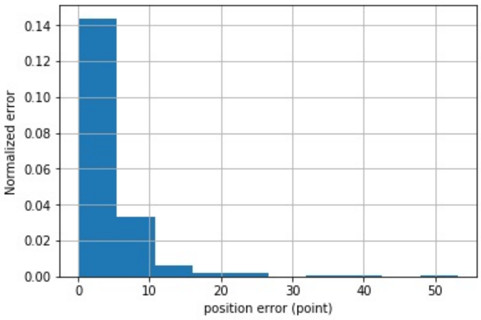


Fig. 7. Position error distribution computed from the difference between the true uid and predicted uid

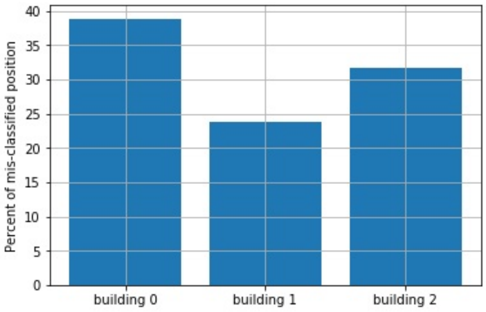


Fig. 8. The number of mis-classified records for each building

B. Classification results

Before an attempt to classify the data from the 584 classes, a trivial problem of predicting floors and buildings are conducted. This experiment is to justify whether the classification program is written correctly and whether the dataset is prepared relatively with no error. The results show that all classification techniques can predict the position with an average accuracy more than 95%. The RF achieves 97% accuracy for floor prediction and 100% building prediction. Nevertheless, the goal is to predict the locid, which can be translated it into an error in position. The error in position is calculated from the Euclidean distance between the co-ordinate (Longitude and Latitude) of the actual locid and the co-ordinate of the predicted locid.

Fig. 6 illustrates some examples of the classification results obtained from RF classification. The results show one misclassified, which is position 6322 (uid=63, floor=2, building=2) is predicted as position 6222 (uid=62, floor=2, building=2). They are essentially, the next-point positions in the same building and at same floor. TABLE II. tabulates the classification accuracy and average error in positioning obtained from four classification algorithms. The RF classification achieves the highest accuracy of 68.50%. The corresponding average error in positioning is 5.65 meters. The error, after adding the 1 meter quantization error, is comparatively better than that of the baseline of 7.9 meters, which is implemented with 1-KNN [8]. An observation of low classification accuracy, but reasonable average error in positioning leads to additional investigation on the relationship between the errors.

Further investigation is conducted to understand the sources of errors with an aim to further improve classification accuracy. The different between the true locid and predicted locid, is then visualized as an error distribution as shown in Fig. 7. The incorrect classification is mainly from a distance error of 0 to 11 points (from 0 to 12 meters). Only a fraction of classification errors is between 12 to 60 points. This investigation suggests that error itself may be difficult to solve due to the nature of the Wi-Fi data that commonly have unstable intensity, causing the classification to predict at its best for only near-by location. The error at greater than 12 points requires an outliner removal in the dataset or further investigation. The mis-classification records are extracted and visualized to show the records with errors in percent for each building and are shown in Fig. 8. Based on this result, a substantial in-depth data exploration may be needed for the building-0 and building-2 data. While further investigation on data are needed for the succeeding research, several hyper-parameter tunings have also been explored to seek a possibility to improve the accuracy of the models. However, little accuracy improvement are gained. For example, increasing number of trees in RF to 500, the model achieves 70.71% accuracy. However, the improvement to the average error in positioning is less than 10%. The results from the hyper-parameter tunings are, therefore, not presented in this paper.

IV. SUMMARY AND DISCUSSION

This paper attempts to predict the position of Wi-Fi equipped devices using the RSSI data collected at 520 WAPs using a proposed quantization algorithm to quantize co-ordinate into a position ID, which can construct the indoor positing problem into a classification problem. Experiments have been conducted on variation of quantization thresholds and evaluated by different classification algorithms. The best quantization threshold is at 1 meter proven by the classification results for this dataset. The results show that Random forest classification achieves the best classification accuracy leading to the least average error in positioning. Further improvement can be achieved through outliner removals and possibly with other advanced classification algorithms.

Acknowledgment

The authors gratefully acknowledge financial support from the Graduate School of Srinakharinwirot University.

References

[1] D. Lymberopoulos, J. Liu, X. Yang, R. R. Choudhury, V. Handziski, and S. Sen, "A Realistic Evaluation and Comparison of Indoor Location Technologies: Experiences and Lessons Learned," *Proc. ACM/IEEE IPSN*, no. April, pp. 178–189, 2015.

[2] A. De Blas and D. López-de-Ipiña, "Improving trilateration for indoors localization using BLE beacons," *Int. Multidiscip. Conf. Comput. Energy Sci.*, vol. 4, pp. 1–6, 2017.

[3] M. Sugano, T. Kawazoe, Y. Ohta, and M. Murata, "Indoor localization system using rssi measurement of wireless sensor network based on zigbee standard," *Proc. IASTED Int. Conf. Wirel. Sens. Networks*, vol. 7, pp. 54–69, 2006.

[4] J. Torres-Sospedra *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," *IPIN 2014 - 2014 Int. Conf. Indoor Position. Indoor Navig.*, no. October, pp. 261–270, 2014.

[5] S. Xia, Y. Liu, G. Yuan, M. Zhu, and Z. Wang, "Indoor Fingerprint Positioning Based on Wi-Fi: An Overview," *ISPRS Int. J. Geo-Information*, vol. 6, no. 5, p. 135, 2017.

[6] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 37, no. 6, pp. 1067–1080, 2007.

[7] Z. Chen, Q. Zhu, H. Jiang, and Y. C. Soh, "Indoor localization using smartphone sensors and iBeacons," *2015 IEEE 10th Conf. Ind. Electron. Appl.*, pp. 1723–1728, 2015.

[8] J. Torres-Sospedra *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," *IPIN 2014 - 2014 Int. Conf. Indoor Position. Indoor Navig.*, no. October, pp. 261–270, 2014.

[9] P. F. Swaszek, "Uniform Spherical Coordinate Quantization of Spherically Symmetric Sources," *IEEE Trans. Commun.*, vol. 33, no. 6, pp. 518–521, 1985.

[10] R. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, no. 2, pp. 4–29, 1984.

[11] J. Ma, X. Li, X. Tao, and J. Lu, "Cluster filtered KNN: A WLAN-based indoor positioning scheme," *2008 IEEE Int. Symp. A World Wireless, Mob. Multimed. Networks, WoWMoM2008*, 2008.

[12] I. Rish, "An empirical study of the naive Bayes classifier," *Empir. methods Artif. Intell. Work. IJCAI*, vol. 22230, no. JANUARY 2001, pp. 41–46, 2001.

[13] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *CoRR*, vol. abs/1302.4964, 2013.

[14] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[15] G. Madzarov, D. Gjorgjevikj, and I. Chorbev, "A Multi-class SVM Classifier Utilizing Binary Decision Tree Support vector machines for pattern recognition," *Informatica*, vol. 33, pp. 233–241, 2009.

[16] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[18] Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python ," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[19] S. Naribole, "A Machine Learning Approach to Wi-Fi Fingerprint based Localization." [Online]. Available: https://github.com/sharan-naribole/wlan_localization.

[20] G. E. P. Box and D. R. Cox, "An analysis of transformations," *J. R. Stat. Soc. Ser. B (Methodological*, pp. 211–252, 1964.

[21] J. Shlens, "A Tutorial on Principal Component Analysis," 2014.