Enhanced Conformal Predictors for Indoor Localisation Based on Fingerprinting Method

Khuong An Nguyen and Zhiyuan Luo

Dept. of Computer Science, Royal Holloway, University of London
Egham, Surrey TW20 0EX, United Kingdom
khuong@cantab.net, zhiyuan@cs.rhul.ac.uk

Abstract. We proposed the first Conformal Prediction (CP) algorithm for indoor localisation with a classification approach. The algorithm can provide a region of predicted locations, and a reliability measurement for each prediction. However, one of the shortcomings of the former approach was the individual treatment of each dimension. In reality, the training database usually contains multiple signal readings at each location, which can be used to improve the prediction accuracy. In this paper, we enhance our former CP with the Kullback-Leibler divergence, and propose two new classification CPs. The empirical studies show that our new CPs performed slightly better than the previous CP when the resolution and density of the training database are high. However, the new CPs performs much better than the old CP when the resolution and density are low.

Keywords: indoor localisation, fingerprinting, conformal prediction.

1 Introduction

The purpose of indoor localisation is to identify and observe a user inside a building. Global Positioning System (GPS) has long been an optimal solution for outdoor localisation, yet the indoor counterpart remains an open research problem, because of the harsh and complex indoor building structure. Current indoor localisation systems remain either too expensive or not accurate enough [4]. In our previous work [2], we proposed the first Conformal Predictor (CP) for indoor localisation based on classification with the weighted K-nearest neighbours algorithm, which performed well in our test sets. However, in reality, the prediction accuracy depends on the resolution, and the density of the training database. In this paper, we enhance our former CP with the Kullback-Leibler divergence, which is a better way to compare two signal strength distributions. We propose two new conformal predictors for classification. The empirical studies show that our new CPs perform slightly better than the previous CP when the resolution and density of the training database are high. However, the new CPs performs much better than the previous CPs when the the resolution and density are low.

The paper begins with a brief introduction of the indoor localisation problem, and the concept of location fingerprinting. The next section describes our

H. Papadopoulos et al. (Eds.): AIAI 2013, IFIP AICT 412, pp. 411–420, 2013.

implementation of the conformal prediction with two new nonconformity measures. The performance of our implementations is evaluated in Section 4. We summarise our contributions and future work in Section 5.

2 The Indoor Localisation Problem

An indoor localisation system is a network of devices used to wirelessly locate objects or people inside a building. A user can be coarsely identified at room-level or precisely localised at sub-room level. Different approaches have been developed in recent years, however, most precise sub-room level tracking systems are expensive, while most affordable tracking systems are not accurate enough [1,7]. Many attempts to improve the location accuracy might work in one environment, but did not work in others because of the signal attenuation. We consider a popular indoor localisation method known as Location Fingerprinting and discuss why CP is a suitable approach for such problem.

The idea of Location Fingerprinting is to use a pre-surveyed database containing a mapping of wireless signal properties to 3-dimensional physical location. The signal properties can be the signal strength (RSSI) or the link quality (LQ). Based on the fact that the wireless signal attenuates and gets weaker as it travels in the air, with many wireless transmitter sources, each location in the tracking zone can have a unique combination of signal properties. To predict the physical location of a known signal properties, the properties are compared with each entry in the database to find a closest match, which will predict possible physical location. The wireless signal properties are regarded as the object set, while the physical location is regarded as the label set.

The main challenge is that two distinct locations might not have a linear relationship in terms of RSSI/LQ and the distance between them. This phenomenon is caused by the human movements, humidity, furniture re-arrangement, as well as the multi-path fading of the indoor environment [3]. In this paper, we show how to deal with such problem using the Conformal Prediction algorithm.

The Location Fingerprinting method can be mathematically formulated as follows. A single location L is modelled in a 3-dimensional space $L = (d^x, d^y, d^z)$. The signal strength RSSI between the user at a location L and all N stations is modelled as an N-tuple $RSSI_L = (s_1, \ldots, s_N)$, where s_i is the signal strength received from the station i $(1 \leq i \leq N)$. The task is that, given an RSSI tuple $RSSI_{unknown} = (s_1, \ldots, s_N)$ of an unknown location inside the tracking zone, the system uses the database B to predict the co-ordinate (d^x, d^y, d^z) of the unknown location. This is a multi-label problem.

3 Conformal Prediction for Indoor Localisation

In our previous work, we proposed the first Conformal Predictor for the indoor localisation based on the weighted K-nearest neighbours and the location fingerprinting method [2]. The algorithm provided better location accuracy than existing methods with a similar approach. Further, our method added a confidence parameter for each prediction to solve the uncertainty problem of the indoor localisation.

3.1 Conformal Prediction

Conformal Prediction (CP) is a relatively new machine learning framework, which uses experiences in the past to confidently and precisely predict the outcome of a new sample [5,6]. It has been mathematically proved that the prediction region generated by CP is valid in the on-line setting [6]. However, CP demands a weak assumption that the training database and the new sample to be classified are generated from the same distribution independently.

The most important component of CP is the 'nonconformity measure', which is a real-valued function $A(B, z)$ measuring how different a sample z is to the training database B. Ideally, for a wrong test label, we would prefer this sample to be completely different from all training samples in B. Therefore, the sequence's randomness can be maintained. This is also the core idea of CP, which can be seen as a test of randomness. Whenever a new sample needs to be classified, we exhaustedly test every label recorded in the training data. A p-value function compares the non-conformity score α_{l+1} of the new sample to all remaining α_i of each example in the training data. If the new sample's label is wrong, the returned p-value is small because, α_{l+1} will be bigger than most α_i. The assumed label of the new sample is accepted if p-value is greater than the significance level ε we choose. Regardless of the chosen nonconformity measure, the set of locations predicted by CP is always valid in the on-line setting.

3.2 Enhancement of Classification Indoor Localisation

In this paper, we enhance our work with two key improvements. First, the Kullback-Leibler divergence (KL) approach will emphasize the similarity between the 2 distributions, rather than calculating the difference in distance in the Euclidean space, as we used in the previous approach. Second, we propose 2 new CPs to include more information with the x, y and z co-ordinates for our nonconformity measure. The new CP algorithm is summarised as follows.

Giving a training database B mapping the signal strength to the physical location co-ordinate (d^x, d^y, d^z), and a signal strength fingerprint at an unknown location, we will predict a set of possible locations in the database, which likely matches this new signal fingerprint. The task can be formulated as a classification problem, because we divide the locations into grid points, and the label set is finite. The measured signal strengths at these grid points are regarded as the object set $\mathbf{X}$, and the physical locations are regarded as the label set $\mathbf{Y}$. We will apply CP using both the old examples - the training database $B = (z_1, z_2, \ldots, z_l)$, and the signal fingerprint of the unknown location (as a new object of z_{l+1}). Each example z_i is a combination of the signal strength $RSSI_i = (s_1^i, s_2^i, \ldots, s_N^i)$ and the co-ordinate $L_i = (d_i^x, d_i^y, d_i^z)$. A prediction region of K examples is $R^{\varepsilon}(L_1, L_2, \ldots, L_K) \subset \mathbf{Y}$.

To calculate the similarity between 2 signal strength distributions P_X and P_Y, we use the symmetrised KL formula, with M is the number of bins in the histogram, and N is the number of stations.

$$Sym\_D_{KL}(P_X, P_Y) = D_{KL}(P_X||P_Y) + D_{KL}(P_Y||P_X) \tag{1}$$

where

$$D_{KL}(P_X \mid\mid P_Y) = \sum_{j=1}^{N} \sum_{i=1}^{M} P_X^j[i] \; log_2 \frac{P_X^j[i]}{P_Y^j[i]} \tag{2}$$

According to the CP algorithm, we will assign each of the labels in the training database to the new sample, and calculate the nonconformity measure for such label. We propose 2 new nonconformity measures based on the above KL divergence, one version is simple and easy to implement, while the other includes more information of the dimensions, but requires more computation.

Using the nonconformity measure, we calculate the nonconformity score α_i, with $i = 1, \ldots, l$, for every example in the database B. We then count the number of α_i which is larger than or equal to the nonconformity score α_{l+1} of the new sample, and divide the total number of examples in the database B to have the p-value for a possible label $\hat{y}$.

$$p(\hat{y}) = \frac{\#\{i = 1, \ldots, l+1 : \alpha_i \geq \alpha_{l+1}\}}{l+1} \quad . \tag{3}$$

Given a significance level ε beforehand (such as $\varepsilon = 0.05$), the current assumed co-ordinate label is accepted as the label for the new sample, if and only if the p-value $> \varepsilon$. All accepted locations form a prediction region, which guarantees to contain the correct position 95% of the time (when $\varepsilon = 0.05$) in the on-line setting. We explain our 2 new CPs below.

3.3 The First Nonconformity Measure

For the first nonconformity measure, we first find K nearest examples in the training database with **different location labels** (d_i^x, d_i^y, d_i^z) to the label of the new sample $(d_{l+1}^x, d_{l+1}^y, d_{l+1}^z)$. We applied the KL divergence $D_{KL}(P_{l+1}, P_i)$, with $i = 1, \ldots, l$, to compare two signal strength distributions, as presented in Equation (2).

Once we obtain a set of K examples, we combine them into one weighted average example with the label $L_{diff} = (d_{diff}^x, d_{diff}^y, d_{diff}^z)$, with ϵ is a small constant to prevent division by zero. Our assumption is that the majority of the similar signal strengths are close to each others. By considering the KL weights, we penalise the isolated neighbours, which appear because of the indoor attenuation problem.

$$d_{diff}^{co\_ord} = \frac{\displaystyle\sum_{i=1}^{K} \frac{1}{D_{KL}(P_{l+1}, P_i) + \epsilon} \; d_{diff}^{co\_ord}}{\displaystyle\sum_{i=1}^{K} \frac{1}{D_{KL}(P_{l+1}, P_i) + \epsilon}}, co\_ord = x, y, z. \tag{4}$$

We then find another set of K nearest examples, this time with the **same location labels** $(d^x_{l+1}, d^y_{l+1}, d^z_{l+1})$ with the new sample. Another weighted average example $L_{same} = (d^x_{same}, d^y_{same}, d^z_{same})$ is calculated as follows.

$$d^{co\_ord}_{same} = \frac{\sum_{i=1}^{K} \frac{1}{D_{KL}(P_{l+1}, P_i) + \epsilon} d^{co\_ord}_{same}}{\sum_{i=1}^{K} \frac{1}{D_{KL}(P_{l+1}, P_i) + \epsilon}}, co\_ord = x, y, z. \quad (5)$$

The nonconformity measure is defined as the distance between these two locations L_{same} and L_{diff}. We use an Euclidean approach to calculate such difference.

$$A_1 = \sqrt{(d^x_{same} - d^x_{diff})^2 + (d^y_{same} - d^y_{diff})^2 + (d^z_{same} - d^z_{diff})^2} \ . \quad (6)$$

3.4 The Second Nonconformity Measure

For the second nonconformity measure, we implement a multi-label approach. Since we have 3 different dimensions (d^x, d^y, d^z) for each location, there are 8 possibilities for any 2 locations (d^x_1, d^y_1, d^z_1) and (d^x_2, d^y_2, d^z_2).

1. $(d^x_1 = d^x_2), (d^y_1 = d^y_2), (d^z_1 = d^z_2)$
2. $(d^x_1 \neq d^x_2), (d^y_1 = d^y_2), (d^z_1 = d^z_2)$
3. $(d^x_1 = d^x_2), (d^y_1 \neq d^y_2), (d^z_1 = d^z_2)$
4. $(d^x_1 = d^x_2), (d^y_1 = d^y_2), (d^z_1 \neq d^z_2)$
5. $(d^x_1 \neq d^x_2), (d^y_1 \neq d^y_2), (d^z_1 = d^z_2)$
6. $(d^x_1 \neq d^x_2), (d^y_1 = d^y_2), (d^z_1 \neq d^z_2)$
7. $(d^x_1 = d^x_2), (d^y_1 \neq d^y_2), (d^z_1 \neq d^z_2)$
8. $(d^x_1 \neq d^x_2), (d^y_1 \neq d^y_2), (d^z_1 \neq d^z_2)$

We find a set of K nearest examples for each of the 8 possibilities, using the Equation (2). For each of the 8 sets, we combine all K examples into one weighted average location, $L_i = (d^x_i, d^y_i, d^z_i)$, with $i = 1, \ldots, 8$, using the Equation (5), as similar to how we did in the previous CP above.

Our nonconformity measure is the difference between a combination of the first 7 cases (where at least one co-ordinate is similar), and the 8th case (where all co-ordinates of the label are different). We combine the first 7 cases into one average location $L_{one\_same} = (d^x_{one\_same}, d^y_{one\_same}, d^z_{one\_same})$.

$$L_{one\_same} = (\frac{\sum_{i=1}^{7} d^x_i}{7}, \frac{\sum_{i=1}^{7} d^y_i}{7}, \frac{\sum_{i=1}^{7} d^z_i}{7}) \ . \quad (7)$$

The nonconformity measure is defined as follows

$$A_2 = \sqrt{(d^x_{one\_same} - d^x_8)^2 + (d^y_{one\_same} - d^y_8)^2 + (d^z_{one\_same} - d^z_8)^2} \ . \quad (8)$$

The algorithm outline is presented in Algorithm 1.

Algorithm 1. Classification Conformal Predictor for Indoor Localisation

Input: Training database $B = \{z_1, \dots, z_l\}$, significance level ε, new example $z_{l+1} = (P_{l+1})$.
Output: Prediction region R.
Function D_{KL} is defined in Section 3.2

```
function WEIGHTED(KNN$_{set}$, $z$)
    for $coordinate = \{x, y, z\}$ do
        for $i = 1 \rightarrow K$ do
            $weight1 = 1/(D_{KL}(P_{l+1}, P_i) + \epsilon) *$ KNN$_{set}(d_i^{coordinate})$
            $weight2 = 1/(D_{KL}(P_{l+1}, P_i) + \epsilon)$
            $d_{weighted}^{coordinate} = weight1/weight2$
        end for
    end for
    return $d_{weighted}$
end function

function NONCONFORMITY_1($B$, $z$)
    for $i = 1 \rightarrow L$ do
        if $(d_i^{\{x,y,z\}} \neq d_z^{\{x,y,z\}}) \& D_{KL}(B_i, z)$ is the smallest then
            KNN$_{diff}$ = KNN$_{diff} + \{B_i\}$
        end if
        if $(d_i^{\{x,y,z\}} = d_z^{\{x,y,z\}}) \& D_{KL}(B_i, z)$ is the smallest then
            KNN$_{same}$ = KNN$_{same} + \{B_i\}$
        end if
    end for
    $L_{diff} = weighted($KNN$_{diff}, z)$ ; $L_{same} = weighted($KNN$_{same}, z)$
    $\alpha_1 = sqrt(L_{diff} - L_{same})$
    return $\alpha_1$
end function

function NONCONFORMITY_2($B$, $z$)
    for $i = 1 \rightarrow L$ do
        if $(d_i^x = d_z^x \& d_i^y = d_z^y \& d_i^z = d_z^z) or (d_i^x \neq d_z^x \& d_i^y = d_z^y \& d_i^z = d_z^z) or (d_i^x = d_z^x \& d_i^y \neq d_z^y \& d_i^z = d_z^z) or (d_i^x = d_z^x \& d_i^y = d_z^y \& d_i^z \neq d_z^z) or (d_i^x \neq d_z^x \& d_i^y \neq d_z^y \& d_i^z = d_z^z) or (d_i^x \neq d_z^x \& d_i^y = d_z^y \& d_i^z \neq d_z^z) or (d_i^x = d_z^x \& d_i^y \neq d_z^y \& d_i^z \neq d_z^z) and D_{KL}(B_i, z)$ is the smallest then
            KNN$_{one\_same}$ = KNN$_{one\_same} + \{B_i\}$
        end if
        if $(d_i^{\{x,y,z\}} \neq d_z^{\{x,y,z\}}) \& D_{KL}(B_i, z)$ is the smallest then
            KNN$_8$ = KNN$_8 + \{B_i\}$
        end if
    end for
    $L_{one\_same} = WEIGHTED($KNN$_{one\_same}, z)$ ; $L_8 = WEIGHTED($KNN$_8, z)$
    $\alpha_2 = sqrt(L_{one\_same} - L_8)$
    return $\alpha_2$
end function
```

Algorithm 1. (*Continued.*)

```
for each possible label $y \in Y$ do
    $z_{l+1} = (P_{l+1}, y)$
    $\alpha_{l+1} = NONCONFORMITY\_1/\_2(B, z_{l+1})$
    for $i = 1 \rightarrow L$ do
        $\alpha_i = NONCONFORMITY\_1/\_2(B - \{z_i\} + \{z_{l+1}\}, z_i)$
        if $(\alpha_i \geq \alpha_{l+1})$ then $count = count + 1$
        end if
    end for
    $p = count/(l + 1)$
    Predictive set $R = \{y : p(y) > \varepsilon\}$
end for
```

4 Performance Evaluation

Having explained our enhancements, we apply the new algorithms for our two Bluetooth fingerprinting testbeds presented in [4]. At each location, we collected the signals in 8 orientations (North, Northest, East, Southest, South, Southwest, West and Northwest). The signal readings are measured multiple times for each orientation.

Testbed 1 is just a single room of 15 square metres (5m x 3m). The resolution and the density of the training dataset is high, with 30-40 signal readings for each orientation at every location. There are 64,000 samples for the training set, and 1,000 samples for the test set.

Testbed 2 has lower resolution and signal density across a 170 square metre (17m x 10m) corridor, with over 48,000 samples for the training set, and 500 samples for the test set. There are just 10-15 signal readings for each of the 8 orientations.

4.1 Performance of the First Nonconformity Measure

Tables 1 and 2 show the performance of the first CP on our 2 data sets, compared to our previous CP. The CP error rate is the percentage in which CP does not produce a prediction region containing the exact location. At the significance level $\varepsilon = 0.15$ and $K = 3$, the new CP has fewer than 1.65m location error for testbed 1, and fewer than 2.3m error for testbed 2.

Table 1. Average system accuracy with 1st nonconformity measure for Testbed 1

Confidence level	Significance level ε	Pred. error	Pred. size	CP error rate	Old Pred. error	Old Pred. size
90%	0.10	≥2m	61	8.3%	≥2m	62
85%	0.15	1.65m	27	13.7%	1.7m	29
70%	0.30	≥1.8m	8	28.2%	≥1.8m	8

Table 2. Average system accuracy with 1st nonconformity measure for Testbed 2

Confidence level	Significance level ε	Pred. error	Pred. size	CP error rate	Old Pred. error	Old Pred. size
90%	0.10	≥2.9m	39	9.7%	≥3.1m	44
85%	0.15	2.3m	17	13.8%	2.5m	19
70%	0.30	≥3m	13	29%	≥3m	13

We observed that the nonconformity measures did not improve the accuracy much on Testbed 1, compared to our previous CP. Our assumption is because of the high resolution and signal density in this test bed, two close locations may have a very different signal readings because of the signal fluctuation and attenuation, which introduced many errors in the training data. We did observe a slight improvement on Testbed 2, compared to our old CP. Testbed 2 has lower resolution than Testbed 1 with few signal readings at each orientation.

4.2 Performance of the Second Nonconformity Measure

Tables 3 and 4 show the performance of the second CP on our 2 data sets, compared to our previous CP. The CP error rate is the percentage in which CP does not produce a prediction region containing the exact location. At a significance level $\varepsilon = 0.15$ and $K = 3$, the new CP has fewer than 1.6m location error for testbed 1, and fewer than 1.9m error for testbed 2.

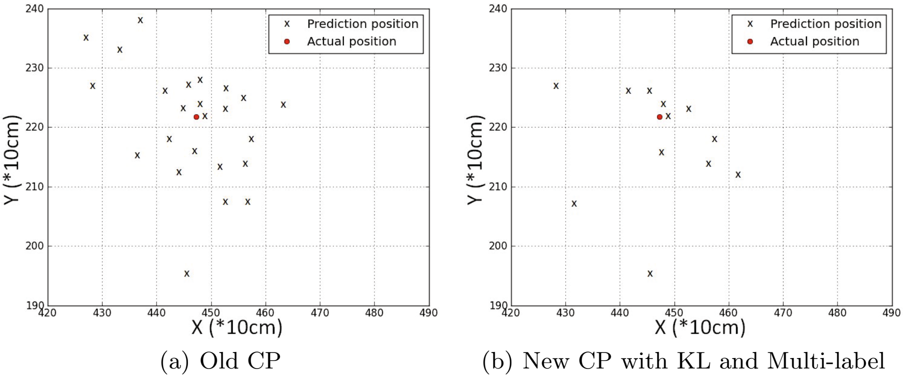
We observed that the nonconformity measure did not improve the accuracy much on Testbed 1, compared to our old CP. However, we did observe a better performance accuracy on Testbed 2, compared to our old CP. Not only does the new CP produce a more accurate prediction, the prediction region is also smaller than that in our old CP for Testbed 2. This finding implies that when the tracking zone is large, and the observed locations are spread out as in our

Table 3. Average system accuracy with 2nd nonconformity measure for Testbed 1

Confidence level	Significance level ε	Pred. error	Pred. size	CP error rate	Old Pred. error	Old Pred. size
90%	0.10	≥1.9m	58	8.3%	≥2m	62
85%	0.15	1.6m	24	13.7%	1.7m	29
70%	0.30	≥1.75m	8	28.2%	≥1.8m	8

Table 4. Average system accuracy with 2nd nonconformity measure for Testbed 2

Confidence level	Significance level ε	Pred. error	Pred. size	CP error rate	Old Pred. error	Old Pred. size
90%	0.10	≥2.6m	37	9.7%	≥3.1m	44
85%	0.15	1.9m	13	13.8%	2.5m	19
70%	0.30	≥2.4m	12	29%	≥3m	13



(a) Old CP (b) New CP with KL and Multi-label

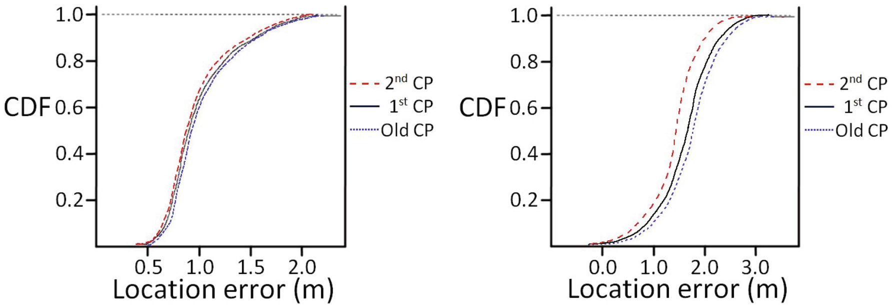
Fig. 1. Prediction Region of new CP with Multi-label Approach on Testbed 2

Testbed 2, the correlation among the co-ordinates has an impact on the prediction accuracy (Figure 1).

4.3 Overall Performance Discussion

Compared to our first classification CP with weighted K-nearest neighbours, our second nonconformity measure with the Kullback-Leibler divergence and the multi-label approach has reduced the prediction error by roughly 25% for Testbed 2 - from 2.5m to 1.9m at 85% confidence level. We still observed at least 16% improvement in different confidence levels.

Unfortunately, we did not see much improvement on Testbed 1 with both nonconformity measures. Our assumption is because of the high resolution and signal density in Testbed 1, two close locations may have a very different signal readings because of the signal fluctuation and attenuation, which introduced many errors in the training data. We use a Cumulative Distribution Function



(a) Testbed 1 (b) Testbed2

Fig. 2. Performance Gain of 2 new CPs with KL and Multi-label

(CDF) plot to compare the performance of both new CPs with the KL divergence to our former CP with the Euclidean approach in 2 test sets. (Figure 2).

5 Conclusion and Future Work

We have proposed 2 new CPs based on the Kullback-Leibler divergence, and the multi-label approach. Empirical studies showed that the performance accuracy has been enhanced up from 16%-25% for our Testbed 2, where the training data's resolution and the signal density are low. We did not observe much improvement in our Testbed 1, where the resolution and signal density are high.

The indoor localisation problem is far from solved, especially in the case with movement tracking, where the uncertainty between two consecutive readings are high. We are researching how to apply Conformal Prediction for such problem.

Acknowledgements. We are grateful to the Computer Science Department, Royal Holloway, University of London for continued support and funding. This work has also been supported by a grant from the National Natural Science Foundation of China (No. 61128003) and EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots").

References

1. Chen, Y., Lymberopoulos, D., Liu, J., Priyantha, B.: Fm-based indoor localization. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, pp. 169–182. ACM (2012)
2. Nguyen, K., Luo, Z.: Conformal prediction for indoor localisation with fingerprinting method. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S. (eds.) AIAI 2012 Workshops, Part II. IFIP AICT, vol. 382, pp. 214–223. Springer, Heidelberg (2012)
3. Nguyen, K., Luo, Z.: Evaluation of bluetooth properties for indoor localisation. In: Progress in Location-Based Services, pp. 127–149. Springer (2013)
4. Nguyen, K.A.: Robot-based evaluation of bluetooth fingerprinting. Master's thesis, Computer Lab, University of Cambridge (2011)
5. Shafer, G., Vovk, V.: A tutorial on conformal prediction. The Journal of Machine Learning Research 9, 371–421 (2008)
6. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer (2005)
7. Wang, H., Sen, S., Elgohary, A., Farid, M., Youssef, M., Choudhury, R.R.: No need to war-drive: Unsupervised indoor localization. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, pp. 197–210. ACM (2012)