Hernández Parra, N., Alonso Moral, J.M. & Ocaña Miguel, M. 2017, "Fuzzy classifier ensembles for hierarchical WiFi-based semantic indoor localization", Expert Systems with Applications, vol. 90, pp. 394-404.

Fuzzy classifier ensembles for hierarchical WiFi-based semantic indoor localization

Noelia Hernández[a,*], Jose M. Alonso[b], Manuel Ocaña[a]

[a] *Universidad de Alcalá, 28871 Alcalá de Henares (Madrid), Spain*
[b] *Universidade de Santiago de Compostela, E-15782, Santiago de Compostela (Galicia), Spain*

Abstract

The number of applications for smartphones and tablets is growing exponentially in the last years. Many of these applications are supported by the so-called Location Based Services, which are expected to provide reliable real-time localization anytime and anywhere, no matter either outdoors or indoors. Even though outdoors world-wide localization has been successfully developed through the well-known Global Navigation Satellite System technology, its counterpart large-scale deployment indoors is not available yet. In previous work, we have already introduced a novel technology for indoor localization supported by a WiFi fingerprint approach. In this paper, we describe how to enhance such approach through the combination of hierarchical localization and fuzzy classifier ensembles. It has been tested and validated at the University of Edinburgh, yielding promising results.

Keywords: Indoor Localization, WiFi, Fingerprinting, Fuzzy Logic, Classification, Ensembles
PACS: 02.50.Tt, 07.05.Mh
2000 MSC: 03B52, 68T27, 68T30, 68T35, 68T37, 94D05

*Corresponding author
Email addresses: noelia.hernandez@edu.uah.es (Noelia Hernández), josemaria.alonso.moral@usc.es (Jose M. Alonso), mocana@depeca.uah.es (Manuel Ocaña)

1. Introduction

Location Based Services (LBS) (Schiller & Voisard, 2004) for applications running on smartphones and tablets represent a flourishing and profitable business. As part of its annual Mobile Life study of 2012 (TNS, Kantar Group, 2012b), TNS remarked the fact that there were more than 6 billion mobile users world-wide. Among them, 19% of users made normally use of LBS while 62% of users were willing to do it in the near future. Moreover, users identified the following main reasons for using LBS (TNS, Kantar Group, 2012a): (1) to navigate directions; (2) to find nearby points of interest; and (3) to find restaurants and entertainment venues nearby. Illustrative examples of popular applications are assistance and guidance for disabled people (Benavente-Peces et al., 2009; Hammadi et al., 2012).

LBS applications running in indoor environments (Werner, 2014) represent a challenging hot research topic for both academy and industry (Neves et al., 2014). It is worthy to note that users demand their LBS applications work properly everywhere and at anytime, respecting privacy protection (Tang et al., 2015) and reducing the energy consumption (Bisio et al., 2016). Unfortunately, these kinds of applications do not work properly indoors yet. This is mainly due to the fact that the well-known Global Navigation Satellite System technology (GNSS), which is the *de-facto* standard outdoors, fails indoors.

Different technologies (magnetic field (Torres-Sospedra et al., 2016), RFID (Tesoriero et al., 2010), WiFi (Camposa et al., 2014)(Stella et al., 2014), etc.) have been used to provide indoor localization. Among them, WiFi is likely to be the most common choice due to its outstanding advantages. Firstly, there are WiFi Access Points (APs) in most buildings. Secondly, measuring WiFi signal is free of charge even for private networks. Of course, the treatment of WiFi signal for localization is not straightforward. APs are usually deployed with the aim of optimizing connectivity but disregarding localization purposes. Moreover, most of the WiFi devices works at 2.4 GHz which is a free band in the frequency spectrum. In consequence, the Received Signal Strength (RSS) becomes extremely noisy and applications must deal with lots of undesired effects: the so-called co-channel interferences (Cardieri & Rappaport, 2001), the multipath effect (Elnahrawy et al., 2004), the small-scale effect (Youssef & Agrawala, 2003), the absorption of part of the RSS by people and objects (Garcia-Villalonga & Perez-Navarro, 2015), and so on.

Indoor WiFi localization systems usually take a map as reference. In robotics, localization is mainly based on a continuous map and the combination of action and propagation models (Malagon-Soldara et al., 2015; Thrun et al., 2005). On the contrary, in case of applications dealing with humans, a fingerprint approach (Campuzano et al., 2015)(Yim, 2008)(Bisio et al., 2014), where the map is made up of a set of discrete semantic locations (Kelley, 2014), is commonly used. RADAR was the pioneer development of a WiFi-based fingerprint localization system with semantic representation (Bahl & Padmanabhan, 2000), providing promising results.

The Fuzzy Sets Theory, introduced by Zadeh in 1965 (Zadeh, 1965), is ready to cope with imprecision and uncertainty which are inherent to the treatment of WiFi RSS. Astrain et al. (Astrain et al., 2006) first introduced how to apply fuzzy sets and systems to deal with imprecise location based on WiFi trilateration estimations. Dharne et al. (Dharne et al., 2006) addressed the problem of localization in mobile sensor networks taken as reference a grid-based map and applying fuzzy rules for local position tracking. Fuzzy sets and systems have also been used for people localization in the context of ambient intelligence (García-Valverde et al., 2013), regarding human activity recognition (Alvarez-Alvarez et al., 2013), or proposing location-aware services (Chen, 2016). Alonso et al. (Alonso et al., 2009) proved how it becomes natural to address the WiFi-based fingerprint localization problem with fuzzy rule-based classifiers. In addition,

they successfully tackled with RSS small-scale variations by means of fuzzy classifiers (Alonso et al., 2011). Recently, the use of fuzzy classifier ensembles is gaining attention for addressing WiFi-based semantic fingerprint localization in large environments (Zhu et al., 2015; Trawinsky et al., 2015; Mehdiyev et al., 2015). Notice that, classifier ensembles are well-recognized machine learning tools capable of obtaining better performance than a single component classifier. They are able to deal with complex and high dimensional classification problems (Kuncheva, 2001), obtaining high accuracy performance. In (Dietterich, 2001), Dietterich provides a reasoning from statistical, computational and representational point of view indicating why ensembles can improve results. Accordingly, fuzzy classifier ensembles are due to increase the performance of WiFi localization systems in real-world environments.

This work focuses on localization of mobile devices indoors, taking as the only source of information the RSS from the already existing APs in the environment. We present a novel framework which emerges of combining our previous approaches for fingerprint WiFi-based semantic indoor localization. Namely, we have refined the hierarchical localization framework described in (Hernández et al., 2016) by embedding the multiclassifier approach first introduced in (Trawinsky et al., 2013) and then enhanced with fuzzy classifier ensembles in (Trawinsky et al., 2015). Experiments carried out in a real scenario at the University of Edinburgh prove how the new framework successfully provides accurate and reliable localization indoors.

The rest of the manuscript is structured as follows. Section 2 revisited our preliminary works which are required to understand the proposal made in this paper. Section 3 presents the core of this work. It thoroughly describes how to combine our previous hierarchical and fuzzy classifier ensembles approaches in a common framework for WiFi-based semantic indoor localization. Then, Section 4 goes in depth with the experimental evaluation of the new framework. Finally, we summarize the main contributions in this work along with some future research lines in Section 5.

2. Preliminary Approaches to Address WiFi Indoor Localization in Large Environments

A well known problem when developing WiFi localization systems at real-world large environments (with a high number of APs and positions) is that performance drops dramatically down in comparison with trials made at simulated environments or labs where conditions are fully under control. To tackle this problem, the authors have previously developed two independent and alternative approaches: (1) a hierarchical approach (Hernández et al., 2016) to reduce the localization error in real-world large environments and (2) a multiclassifier approach (Trawinsky et al., 2015) to increase accuracy of WiFi-based localization systems. Both approaches follow a two-stage WiFi-based fingerprint localization. Firstly, the reference map is built in an off-line training stage. Secondly, it is exploited in the on-line localization stage when the location of the device is estimated. The rest of this section is devoted to introduce the specific bases behind each approach.

2.1. Hierarchical Localization Approach

Following the well-known divide and conquer principle, we proposed to split the test environment into smaller sub-zones with a reduced number of APs and reference positions to be identified (Hernández et al., 2016). This way the localization task can be carried out in an intuitive, hierarchical manner.

Once the radio map is built with fingerprints taken from all visible APs, the process of automatically learning the localization hierarchy starts. To do so, a hierarchical partition of the map is

created using a clustering algorithm. Namely, we applied the KMeans clustering algorithm (MacQueen, 1967) and the Caliński-Harabasz Index (Caliński & Harabasz, 1974) to choose the most suitable value of k. Then, a classifier was trained for each zone. This way, the system was able to locate the device through the different levels of the hierarchy using the previously trained classifiers. We considered three different classifiers: KNN (Kibler & Aha, 1987), FURIA (Hühn & Hüllermeier, 2009) and SVM (Cortes & Vapnik, 1995).

The goodness of this localization approach was successfully validated with experiments in a real-world multifloor environment at the University of Alcalá. It yielded an improvement of accuracy around 11% (getting an error reduction around 22.5%) versus the non-hierarchical counterpart. The interested reader is kindly referred to (Hernández et al., 2016) for further details.

2.2. Multiclassifier Localization Approach

Classifier ensembles (CE), or multiclassifiers in short, combine the outputs provided by groups of machine learning systems, in the literature called weak learners, in order to yield better accuracy (e.g. lower error rates) (Kuncheva, 2001). Their performance strongly relies on diversity of the weak learners (Brown et al., 2005), in the way that ideally they make their errors on different parts of the problem space. In other words, an individual classifier must provide different patterns of generalization in order to obtain a diverse set of classifiers composing a highly accurate ensemble.

The most common group of CE approaches is based on manipulating training sets (Dietterich, 2000). In particular, it considers data resampling with the aim of generating different training sets to derive each individual classifier. Two algorithms can be distinguished from this family, bagging (Breiman, 1996) and boosting (Freund & Schapire, 1996; Pedrycz & Kwak, 2006). On the one hand, bagging, also called bootstrap aggregation, improves the stability by reducing the variance. The training sets (*bags*) for each base classifier are randomly selected with replacement from the original training data set. Then, the base classifiers separately (in parallel) learn from the generated *bags*. On the other hand, boosting sequentially generates the weak classifiers by selecting the training set for each of them based on the performance of the previous classifier(s) in the series. Opposed to bagging, this resampling process gives a higher weight to the incorrectly predicted examples by the previous classifiers. On the contrary, examples already classified correctly are assigned lower weights for training new classifiers. In consequence, the next weak classifiers focus more on the examples misclassified by the previous weak learners.

Furthermore, we can emphasize a second group, which includes a more diverse set of methods, inducing the individual classifier diversity through some alternative ways (different from resampling (Zhou, 2005)). For instance, feature selection is a key component in many of them, using different subsets of the original features to derive each base classifier (Tsymbal et al., 2005). Random subspace (Ho, 1998a), generating each feature subset at random, is the most representative method of this group.

It is worthy to note that we have already proved the goodness of considering CE for WiFi indoor localization (Trawinsky et al., 2015). Namely, we combined Random Linear Oracles (RLO) (Kuncheva & Rodríguez, 2007) with fuzzy CE (Trawiński et al., 2011) in a framework for fingerprint WiFi-based localization (Trawinsky et al., 2013). In the off-line training stage, we considered FURIA for generating the base fuzzy classifiers along with RLO, which is a fast and generic method able to induce more diversity thus improving the final performance. In the on-line localization stage, we aggregated the information provided by the set of component classifiers with the aim of computing the final output, assuming that all the classifiers were trained over the entire feature space.

3. Fuzzy Classifier Ensembles for Hierarchical Localization

In this section, we provide a description of the proposed approach for the WiFi indoor localization. It combines the hierarchical approach (Hernández et al., 2016) with the CE approach (Trawinsky et al., 2013) in order to improve the accuracy performance in large environments.

Both frameworks have been improved and updated by the incorporation of hierarchical clustering into the hierarchical approach to obtain a better division of the environment and two new CE configurations, using Random Linear Oracles (RLO) and Random Spherical Oracles (RSO), to identify the location of the device through the hierarchy reducing the mean localization error in comparison with other methods. The versatility of both frameworks have been proved (as will be shown in Section 4) being able to deal with different clustering and classification methods, reducing the localization error in a very challenging environment.

A block diagram of the entire system is shown in Figure 1. It is divided in two stages: the training stage and the localization stage. During the first one, the localization system is created by using the available training data. Then, during the localization stage, the trained system is used to estimate the position of the device. Each stage will be explained in the following subsections.

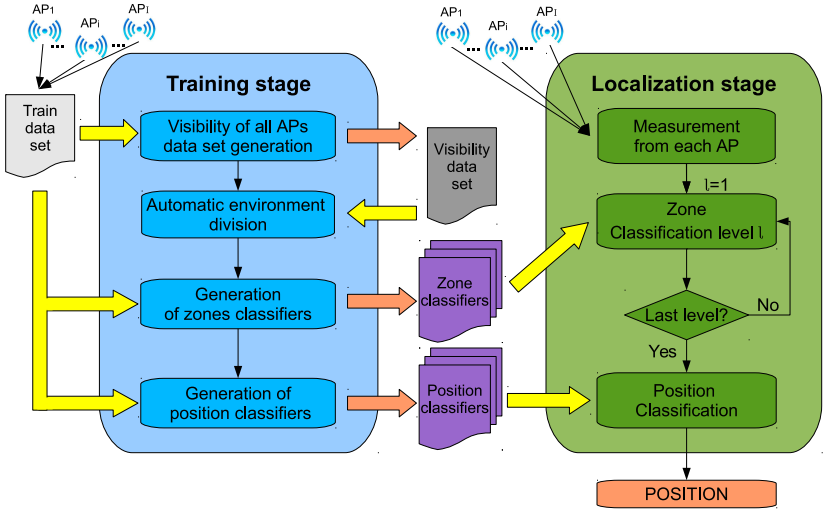


Figure 1: Hierarchical localization approach.

3.1. Training Stage

The objective of this off-line stage is to create the WiFi-based localization system. To do so, the first step is to collect the training data to be used to train the localization system. This process consists on measuring the RSS at the positions on the environment to be identified by the localization system.

The collected data is now converted to the so-called "visibility dataset" and used to divide the environment. The visibility of an AP at a certain position is computed as the percentage of samples measured out of the total samples intended to be collected. As an example, if we were collecting 10 samples from an AP at a certain position and only received 8, the visibility for the

AP will be 80%. The visibility dataset have been used to divide the environment as it has been proved the one providing the lowest localization error.

The enviroment division is created by using a clustering algorithm to find groups of positions with similar APs visibility. This division in sub-zones is repeated until each sub-zone contains a maximum number of positions, creating a hierarchical tree of positions (Figure 2).

Once the environment division tree is created, the classifiers for each new sub-zone are trained. Two different kinds of classifiers are created: Firstly, the so-called "zone classifiers" (represented by squares in Figure 2) which are in charge of deciding the most likely sub-zone where the device is in the immediately lower level. Secondly, the so-called "position classifiers" (represented by circles in Figure 2) which are in charge of decide the position of the environment where the device is.

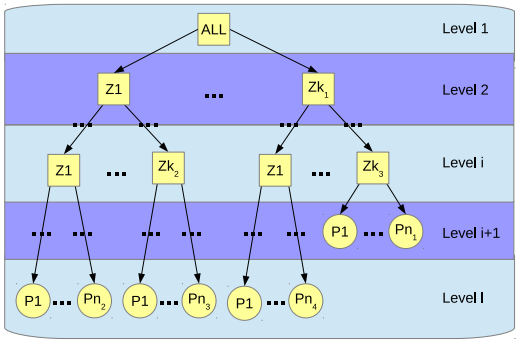


Figure 2: Environment division tree.

3.2. Localization Stage

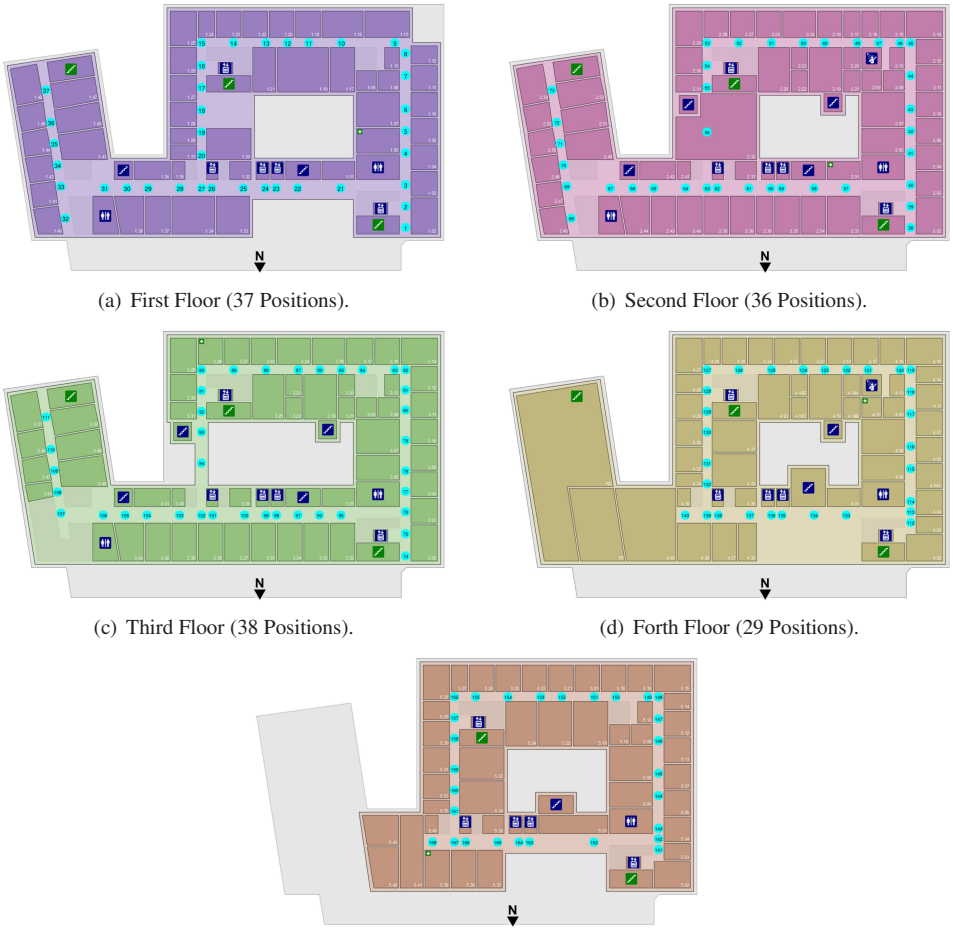
The objective of this on-line stage is to estimate the position of a WiFi device using the system created during the training stage.

A RSS sample collected at an unknown position of the environment is classified through the hierarchical tree to obtain the most likely position of the device. To do so, the sample is first classified using the "Level 1" zone classifier, obtaining the sub-zone where the sample belongs to. Then, the sample is classified again using the "zone classifier" corresponding to the sub-zone previously identified. The classification continues until a "position classifier" is reached. This means that the lowest level of the hierarchy tree branch has been reached. The output of the "position classifier" is the output of the system (the estimated position where the sample has been collected).

4. Experimental Analysis

The proposed approach has been tested in a complex real-world environment. The experiments have been performed on the Informatics FORUM at the University of Edinburgh (Figure 3). Many interferences affect the signal measure since the building is located on the city centre. In this building, mainly made of glass, the attenuation of the signal is lower than usual making harder to distinguish between positions. Moreover, there is an open area in the center of each

floor, making even harder to differentiate between floors and positions located at apposite sides of the building (this effect will be seen when analysing the environment division shown in Figure 9). The environment is made up of five floors with a surface of 2500m^2 each. In our experiments, we detected 164 APs that were deployed over the environment with the aim of providing Internet access to the students but disregarding localization purposes. It is important to highlight that we do not have any information about these APs (their location, configuration, etc is unknown). In consequence, the RSS collected from all the available APs is used to train the system and the localization task is performed using the RSS from the APs with no prior knowledge about their physical location. We have considered 168 significant semantic positions (distributed over the five floors) represented by circled numbers in Figure 3. In this environment, the minimum distance between neighbour positions is 1.64 m, the maximum distance 13.22 m and the mean distance 5.10 m.



(a) First Floor (37 Positions).

(b) Second Floor (36 Positions).

(c) Third Floor (38 Positions).

(d) Forth Floor (29 Positions).

(e) Fifth Floor (28 Positions).

Figure 3: FORUM test-bed environment.

4.1. Learning Algorithms

This section provides a description of the algorithms proposed to solve the localization problem and a brief revision of the algorithms used to be compared with.

- ***Environment Division (Clustering Algorithms):*** Hierarchical (Johnson, 1967) and KMeans clustering algorithm (MacQueen, 1967) are used to obtain the hierarchical partition of the environment. The objective is to create a partition of the environment maximizing intra-cluster similarity. For this process, deciding the right number of clusters is a key task. To choose the number of clusters (sub-zones) in which each zone will be divided, the Caliński-Harabasz Index (Caliński & Harabasz, 1974), also known as Variance Ratio Criterion (VRC), has been used. It performs a quantitative evaluation of clusters looking for compact and well-separated clusters within the feature space.

  Caliński-Harabasz Index has been chosen since it is one of the criteria providing the highest hit rates while having the lowest computational complexity as stated in the study carried out in (Vendramin et al., 2010).

- ***Localization (Zone and Position Classifiers):*** Besides the proposed CE, three classifiers have been tested for comparison purposes: K-Nearest Neighbours (KNN) (Kibler & Aha, 1987), Support Vector Machines (SVM) (Cortes & Vapnik, 1995), Fuzzy Unordered Rule Induction Algorithm (FURIA) (Hühn & Hüllermeier, 2009), and five CE: Random Forest (RF) (Breiman, 2001), AdaBoost (Freund & Schapire, 1997), Bagged Trees and Subspace KNN and Subspace Discriminant (Ho, 1998b).

  - **KNN** is one of the simplest machine learning algorithms. It classifies an object with a voting mechanism selecting the most voted class among the k nearest neighbours to the object. KNN was used in RADAR (Bahl & Padmanabhan, 2000) which is world-wide recognized as one of the pioneers in the research field of WiFi indoor localization and usually used as baseline to compare with new indoor WiFi localization systems (Youssef & Agrawala, 2008; Wu et al., 2007).
  - **SVM** is a kernel-based classifier. It has been proved as one of the most accurate algorithms for hierarchical WiFi indoor localization in our previous research (Hernández et al., 2016).
  - **FURIA** is a fuzzy rule learning algorithm which extends the well-known RIPPER algorithm (Cohen, 1995), getting better accuracy while preserving its advantages, such as simple rule sets. It is admitted as one of the algorithms able to yield the most accurate single fuzzy classifiers in most benchmark datasets.
  - **CE**, as explained in the previous section, are well-recognized machine learning tools capable of obtaining better performance than a single component classifier. They are able to deal with complex and high dimensional classification problems (Kuncheva, 2001), obtaining high accuracy performance.
    - **RF** is an ensemble learning method that creates a combination of decision trees, each one depending on the values of a random vector sampled independently. Random forest makes use of multiple decision trees, trained on different parts of the same training set, with the goal of reducing the overfitting. It is known for providing high accuracy in complex indoor localization problems (Mo et al., 2014)(Jedari et al., 2015)(Calderoni et al., 2015).

- **AdaBoost**: Is a boosting CE that sequentially generates the weak classifiers by selecting the training set for each of them giving a higher weight to the incorrectly predicted examples of the previous classifiers.
- **Bagged Trees**: It is a bagging CE that trains each model using training sets randomly selected with replacement from the original training data set.
- **Subspace ensembles**: Improve the accuracy of discriminant analysis or KNN classifiers by selecting random features instead of the complete feature set.
- **Random Oracles (RO)**: Our proposal is based on RO-based bagging fuzzy rule-based classification systems as described in (Cordón & Trawiński, 2013): Namely, we have considered two versions using random linear oracle (RLO) (Kuncheva & Rodríguez, 2007) (Rodríguez & Kuncheva, 2007) (for now on RLO-CE) and using random spherical oracle (RSO) (Rodríguez & Kuncheva, 2007) (for now on RSO-CE). The first one uses a randomly generated hyperplane to divide the feature space, while the second one uses a hypersphere.

Table 1 summarizes the results achieved training the proposed classifiers to cover all the environment (without using the hierarchical approach). In consequence, in this experiment, the localization system is composed of only one classifier to locate the device over the whole environment. This experiment has been designed to get the baseline to compare with the complete system, when the hierarchical approach is applied. As can be seen, the lowest localization error is obtained when using the proposed CE classifiers, achieving an accuracy higher than a 75% with a mean distance error around 2.2 m and a distance error under 7 m for the 90% of the classified samples.

Table 1: Classifiers comparison using a single classifier.

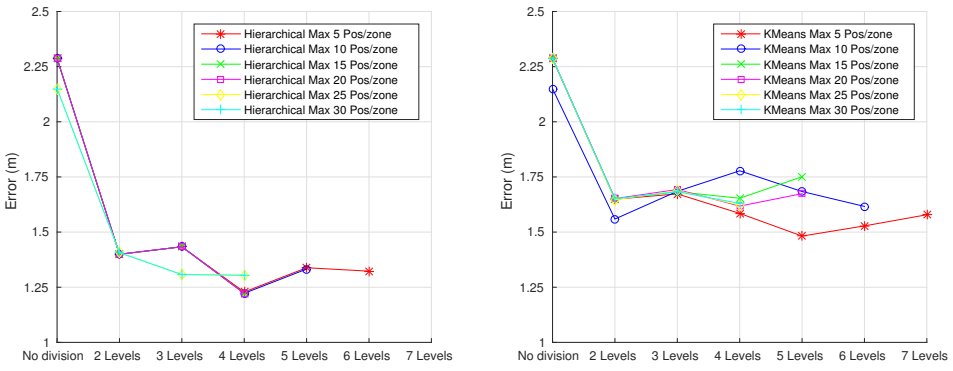
	Accuracy (%)	Mean Error (m)	Error for the 90th Percentile (m)
FURIA	51.04%	5.688 m	17.300 m
SVM	51.49%	4.477 m	12.900 m
KNN	60.12%	3.257 m	9.600 m
RF	65.33%	2.930 m	8.600 m
AdaBoost	43.00%	10.330 m	35.700 m
Subspace Discriminant	48.96%	5.541 m	15.000 m
Subspace KNN	63.09%	3.456 m	9.600 m
Bagged Tree	71.06%	2.824 m	7.900 m
RSO-CE	**75.60%**	2.287 m	6.700 m
RLO-CE	75.22%	**2.150 m**	**6.200 m**

We have used the implementations of KNN, SVM, FURIA and RF provided by the data mining tool Weka (Hall et al., 2009; Witten et al., 2011) and the AdaBoost, Bagged Tree, Subspace KNN and Subspace Discriminant implementations provided by MATLAB (MATLAB, 2016).

It is worthy to note that in the following experimentation, the mean error using FURIA and SVM is always higher than using the rest of classifiers and the mean error using AdaBoost, Subspace KNN and Subspace Discriminant is always higher than using the rest of CE (similar to the results shown in Table 1). In consequence, only the results using KNN, RF, Bagged Tree and the proposed RSO-CE and RLO-CE will be discussed in depth in order to keep the rest of the section easy to read. However, the complete set of results is presented as supplementary material.

4.2. Environment Division Analysis

In this section, the results obtained using the hierarchical clustering algorithm to create the hierarchical partition of the environment are compared with the results using the KMeans clustering algorithm. Figure 4(a) shows the results using the hierarchical clustering algorithm and Figure 4(b) shows the results using the KMeans clustering algorithm to divide the environment. In both subfigures, six graphs are shown corresponding to the results stopping the division of each zone when the number of positions goes under a given threshold (5, 10, 15, 20, 25 and 30 positions). The horizontal axis represents the levels of the environment division tree. The results labelled as "No division" correspond to the results obtained when the hierarchical approach is not applied (this means the localization is performed using one classifier containing all the positions in the environment and corresponds to the results shown in Table 1). The results labelled with the maximum number of levels are obtained using the complete hierarchical partition (it is the mean error using the proposed method). Finally, the results labelled from "Level 2" to the maximum number of levels minus one are intermediate results just shown to observe the trend of the results using the hierarchical approach when the localization process is stopped once the level is reached. Finaly, the mean error shown for each configuration corresponds with the combination of zone and position classifiers providing the lowest mean error. However, the trends of the rest of combinations are similar to the ones in the figure.



(a) Division using Hierarchical clustering.

(b) Division using KMeans clustering.

Figure 4: Mean error using (a) Hierarchical clustering vs. (b) KMeans clustering.

As can be seen, the mean error using the hierarchical clustering is always lower than using the KMeans clustering algorithm, independently of the maximum number of positions in each zone that was set as stopping criterion. In addition, the selection of the maximum number of positions per zone is not critical since the highest reduction of the mean error is obtained after the first

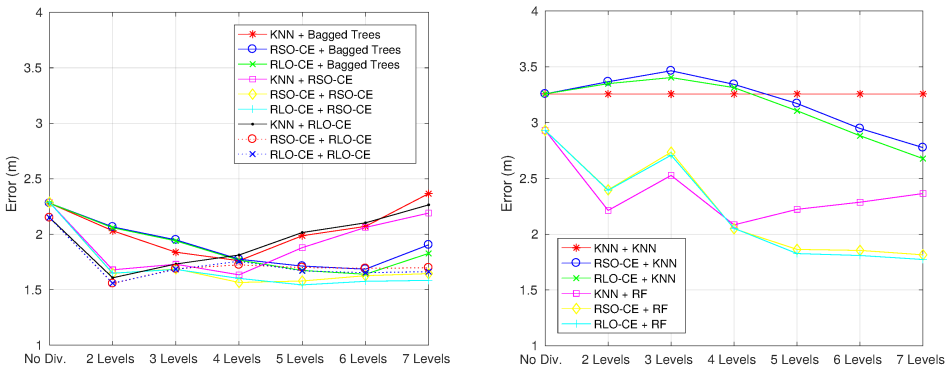
division (around 39% using the hierarchical clustering and 32% using the KMeans clustering algorithm).

The maximum number of positions per zone could be selected to obtain the lowest mean error or the lowest complexity of the system, achieving in both cases improvements in the mean error higher than 30%. To obtain the best results regarding the mean distance error, the division of the environment should be composed of zones with maximum 15 positions using the hierarchical clustering and maximum 5 positions using the KMeans clustering algorithm.

4.3. Localization Error Evaluation

Figures 5 and 6 show the mean error for different classifier combinations using the best environment division using the hierarchical and KMeans clustering algorithms as explained in the previous section.

In both figures, the subfigures on the left (Figure 5(a) and 6(a)) show the mean error obtained using the Bagged Tree, RLO-CE and RSO-CE as "position classifiers", while the subfigures on the right (Figure 5(b) and 6(b)) show the mean error obtained using KNN and RF as "position classifiers". All the classifiers have been used as "zone classifiers" creating different combinations as can be seen in the legend of the figures. As an example, "KNN+RLO-CE" means KNN has been used for the "zone classifiers" while RLO-CE has been used for the "position classifiers".

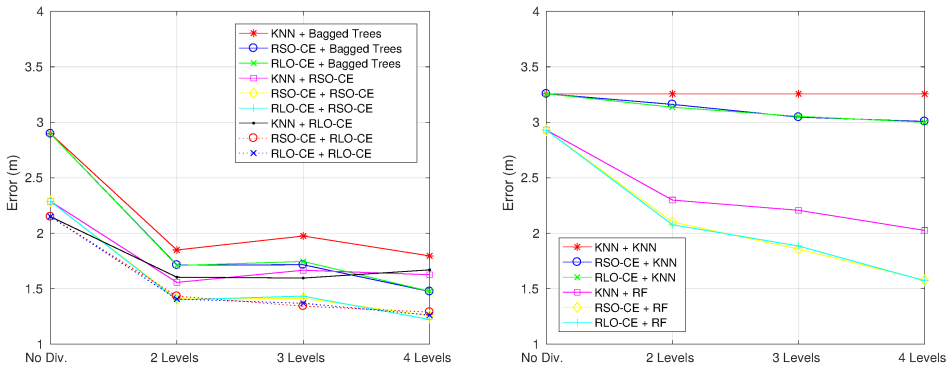


(a) (b)

Figure 5: Mean error using different classifiers combinations for the best division using KMeans clustering algorithm (maximum 5 positions per zone).

The results are similar using both clustering algorithms, although lower mean errors are obtained when the environment is divided using the hierarchical clustering algorithm. The following conclusions can be drawn from the information contained in Figures 5 and 6:

- It is more critical to choose an accurate classifier for the "position classifiers" than for the "zone classifiers", especially when the environment division is created using the hierarchical clustering algorithm.
- KNN is the worst classifier for both the "zone" and "position" classifiers among the ones represented in the graph. As explained before, the error obtained using FURIA, SVM,



(a) (b)

Figure 6: Mean error using different classifiers combinations for the best division using Hierarchical clustering algorithm (maximum 15 positions per zone).

AdaBoost, Subspace KNN and Subspace Discriminant are higher than the error using KNN.

- The CE provide the best results. Among them, the CE proposed in this article (RSO-CE and RLO-CE) provide the lowest mean errors. The best combination of classifiers are the ones using the RLO-CE and RSO-CE for both the "zone" and "position classifiers".

- The mean error difference when using the four combinations with the RSO-CE and RLO-CE (for both zone and position classifiers) is almost negligible (with a difference of 6 cm from the best to the worst combination). This suggests that the decision on using either of them is not critical looking at the localization accuracy.

- The lowest mean error is obtained by using RLO-CE for the "zone classifiers" and RSO-CE for the "position classifiers" achieving a mean error of 1.22 m (43.16% error reduction comparing the hierarchical versus the non-hierarchical approach, 38.96% compared with the hierarchical approach using RF classifiers and 26.43% compared with the Bagged Tree using the hierarchical approach).

The results in this section have been obtained using the division of the environment providing the lowest mean errors (maximum 15 positions per zone using the hierarchical clustering and 5 positions using KMeans clustering) as exposed in the previous section. However, the conclusions previously exposed are applicable to the rest of configurations with the only difference of higher mean errors.

The following conclusions can be obtained after analysing the complete set of results:

- For all the classifiers there is at least one environment division achieving mean localization error reduction. In fact, for most of them, almost any environment division improves localization, especially using the hierarchical clustering algorithm (this is true except for AdaBoost which does not perform well in this problem and for Subspace KNN which is neither benefited nor worsened by the hierarchical approach).

- The tested Boosting CE algorithm (AdaBoost) does not perform well in our experimentation achieving errors around 10.5 meters using a single classifier. However, using the hierarchical approach the error is reduced from 10.5 meters to 7.6 meters using the Hierarchical clustering with maximum 30 positions per zone. The results using AdaBoost suggest that this algorithm performs better with bigger zones, being not benefited by the hierarchical approach using zones smaller than 20 positions.

  In addition, when using AdaBoost as a "zone classifier" the error is highly increased (getting errors up to 36 metres). This effect is caused because "zone classifiers" have to classify among a small number of sub-zones (usually under 4-5 sub-zones) which is the case when AdaBoost does not perform well as explained before.

- Regarding the "zone classifiers", the best results are achieved using the RLO-CE or RSO-CE independently of the division of the environment and the algorithm used as "position classifier". There are some exceptions when the environment division is created using KMeans and the classification is performed using KNN-based algorithms (this is an expected behaviour as previously explained and proved in (Hernández et al., 2016) since KMeans and KNN are distance-based algorithms). When the rest of algorithms are used as "zone classifiers" the mean localization error is increased in most of the cases when compared with the single classifier approach.

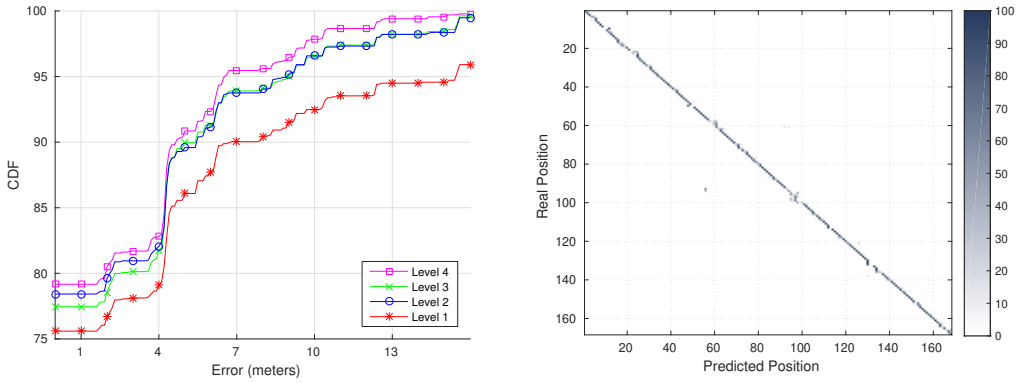
- For any environment division, the minimum mean distance error is always provided by a RLO-CE and RSO-CE combination for both "zone" and "position classifiers". Using these classifiers the method and maximum number of positions per zone used to divide the environment is not critical: using the hierarchical clustering the mean error varies from 1.22 to 1.30 metres when changing the maximum number of positions per zone and, using the KMeans algorithm, from 1.55 to 1.60 metres.

- Finally, Subspace KNN provides low mean distance errors when using a single classifier (3.45 meters), but is one of the less benefited by the use of the hierarchical approach (obtaining similar error with and without using the hierarchical approach).

4.4. Final System Evaluation

Figure 7 shows the Cumulative Distribution Function (CDF) along with the confusion matrix for the configuration providing the highest accuracy (hierarchical clustering with maximum 15 positions per zone, RLO-CE for the "zone classifiers" and RSO-CE for the "position classifiers"). The CDF (Figure 7(a)) shows an analysis of the distance to the real positions in the different levels of the hierarchical system. As can be seen, the error decreases as the number of levels increases, obtaining 90% of the classified samples with an error under 4.5 metres and 80% of the samples with an error under 2 metres. The confusion matrix (Figure 7(b)) details the predicted positions by the system related to the positions where the device really was. Looking at the figure, it can be seen that most of the classification errors occur within the nearest positions. Notice that, since we perform indoor localization using discrete semantic locations, the minimum error in distance depends on the minimum distance between the semantic positions (1.64 metres in this environment).

Figure 8 shows a tree representing the division of the environment for the configuration providing the best results. In this figure, the horizontal dotted lines show the division between the different levels of the hierarchy, the squares represent the "zone classifiers" and the circles denote



(a) CDF. (b) Confusion Matrix.

Figure 7: CDF and Confusion Matrix using the proposed method: Environment division by hierarchical clustering (Figure 8) plus hierarchical classification by RLO-CE for the "zone classifiers" and RSO-CE for the "position classifiers".

the "position classifiers". The number under the circles correspond to the number of positions of the corresponding zone. Finally, the lines joining the nodes represent the hierarchy between the different zones, showing the number of subzones in which a zone is divided. As can be seen, the environment has been divided in 4 different levels, obtaining 37 final zones containing 1 to 12 positions each.

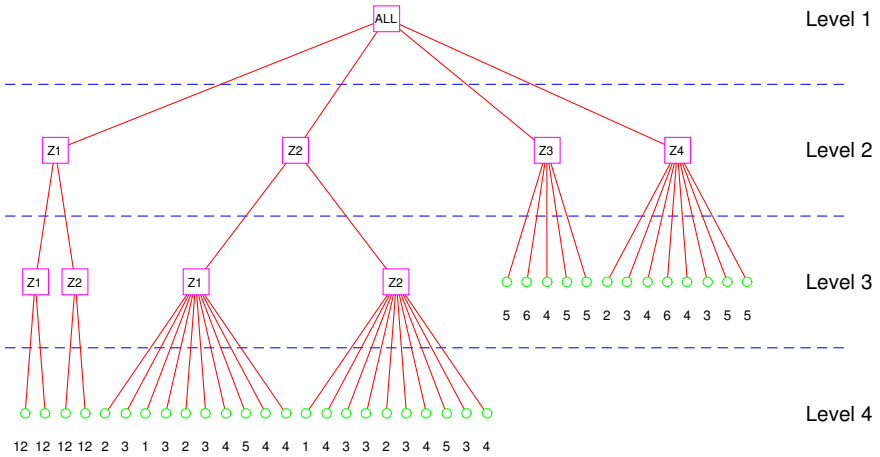
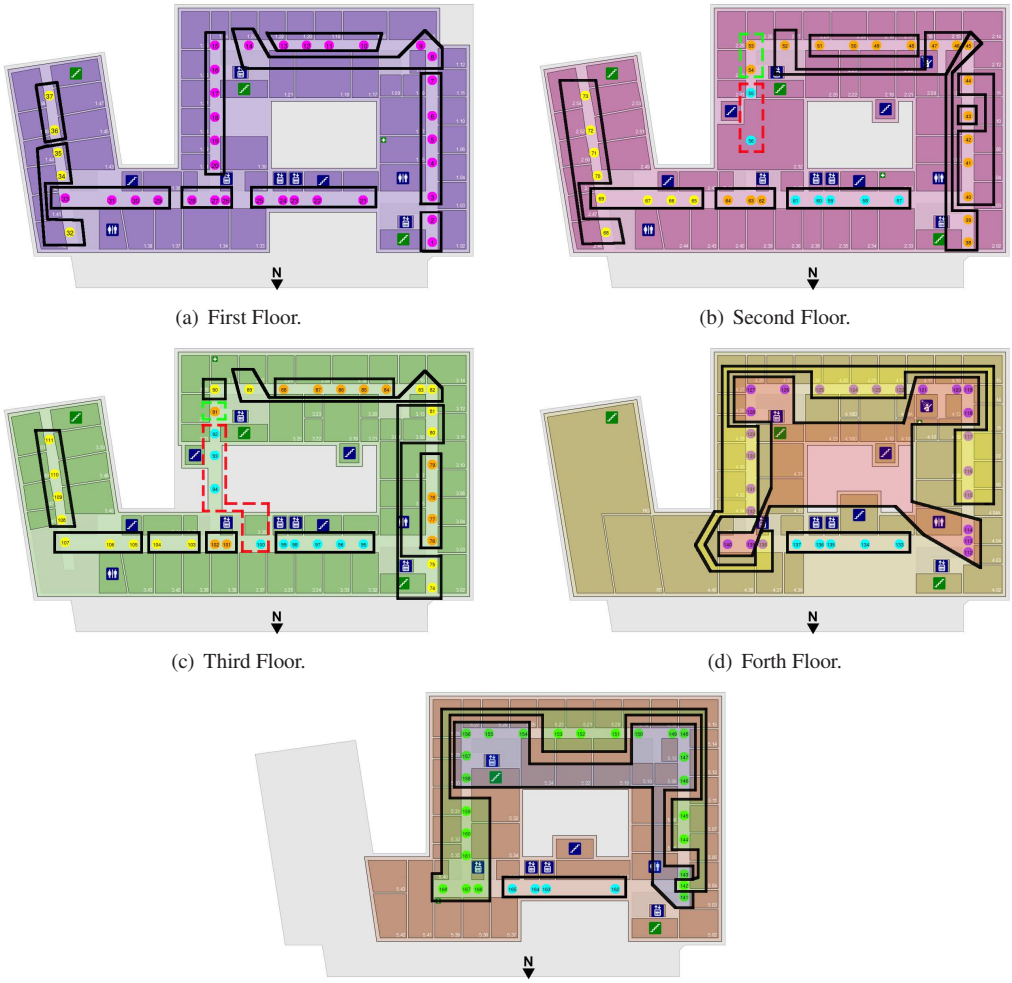


Figure 8: The best environment division with the proposed approach.

Finally, Figure 9 shows the distribution of the positions in the environment division where the black squared areas represent the positions that belong to the same zone in the last level of each branch of the division tree. As can be seen most of the positions belonging to the same zone are neighbouring positions. Except for the positions at the corners of each floor that are

similar to each other (this can be better seen in the maps of the fourth and fifth floors Figures 9(d) and 9(e)). Finally, all grouped positions are in the same floor, suggesting that the RSS is more similar in neighbouring positions than in positions on top or bottom of the others. There are two exceptions (represented using dashed red and green lines so they are easier to identify) where the groups are distributed over two floors (namely the second and third floors). However, as can be seen in Figure 10, the positions corresponding to these zones are located on an open area (second floor) with a walkway (third floor), so it makes sense these positions are grouped together in the same zone.



(a) First Floor.

(b) Second Floor.

(c) Third Floor.

(d) Forth Floor.

(e) Fifth Floor.

Figure 9: FORUM environment division.

(a) View of the open area on the second floor from the open area in the center of the building.

(b) View from the walkway on the third floor to the open area on the second floor.

Figure 10: FORUM open area in the second floor with a walkway on the third floor.

5. Conclusions

In this paper we have presented a novel framework for WiFi-based semantic indoor localization. It takes as starting point our previous fingerprint approach based on classifier ensembles that has been proved as a reliable and accurate proposal for indoor WiFi-based localization. Then, it has been improved by applying a hierarchical localization approach, previously designed by the authors, that has reported accurate localization reducing the mean error of fingerprint-based localization systems.

The proposed system has been tested in a real-world environment, under real conditions (people moving around, doors opening and closing, etc). The APs already existing in the environment have been used during the experimentation, with no prior information about their location, their configuration or their working times. Two datasets have been collected on different days and under different conditions, the first one to be used as a training dataset, the second one to be used as test dataset. The proposed hierarchical CE approach has achieved a mean error of 1.22 metres, achieving an error reduction around 43% versus the non-hierarchical CE approach and around 58% compared with the RF-based localization system (which reported the best results among the usual fingerprint-based localization systems). When compared with Bagged Tree using the hierarchical approach, the proposed CE achieves a 26.43% error reduction.

The contribution of the paper lies in two important points: On the one hand, the hierarchical framework has been updated and improved by including hierarchical clustering and a new kind of classifiers (classification ensembles) that have not been previously tested. The idea behind classifier ensembles is to combine the outputs of multiple classifiers in one unique classifier, while the idea behind the hierarchical approach is to combine the outputs of multiple classifiers through the use of all of them independently in a hierarchical tree structure. Both methods seek for similar goals using different approaches and, as a consequence, it is not evident that its combination will improve the localization accuracy. The experimentation performed in this manuscript proves that the hierarchical framework is able to reduce the localization error even for classifier ensembles reducing the mean localization error when compared with the non-hierarchical approach.

On the other hand, the experimentation has been performed in a very challenging five floor

environment made up of glass walls, located in the Edinburgh city centre causing the attenuation of the signal to be very low. As a consequence, classical fingerprinting algorithms like SVM or KNN, that usually provide acceptable results, obtain around 50% - 60% accuracy and localization errors higher than 3-4 metres. Using the proposed approach the accuracy of the system has been increased reaching an accuracy around 80% and reducing the mean localization error to 1.22 metres.

In the future, we are planning on designing a procedure to select some of the available APs to perform the localization, maintaining the most reliable ones. This way, the localization accuracy is expected to be improved and the complexity of the system reduced, especially in large environments. In addition, we are planing on including the information provided by other sensors (such as compasses, accelerometers, etc) to also improve the localization while locating a device in motion. Finally, we would like to focus on the energy efficiency of the method as proposed by (Bisio et al., 2016) and the reduction of the training effort as proposed by (Bisio et al., 2014) and (Caso & Nardis, 2015).

Acknowledgment

This work has been funded by the Spanish Ministry of Economy and Competitiveness under TIN2014-56633-C3-3-R (ABS4SOW project), the "Consellería de Cultura, Educación e Ordenación Universitaria" (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF). We would also like to thank the WiMo Research Group of the University of Edinburgh (UK) where Noelia Hernández was a visiting researcher and collected the data presented in the experimental section.

References

Alonso, J. M., Ocaña, M., Hernández, N., Herranz, F., Llamazares, A., Sotelo, M. A., Bergasa, L. M., & Magdalena, L. (2011). Enhanced WiFi localization system based on soft computing techniques to deal with small-scale variations in wireless sensors. *Applied Soft Computing*, *11*, 4677–4691.

Alonso, J. M., Ocaña, M., Sotelo, M. A., Bergasa, L. M., & Magdalena, L. (2009). WiFi localization system using fuzzy rule-based classification. In *Computer Aided Systems Theory - EUROCAST 2009* (pp. 383–390). volume 5717 of *Lecture Notes in Computer Science*.

Alvarez-Alvarez, A., Alonso, J. M., & Trivino, G. (2013). Human activity recognition in indoor environments by means of fusing information extracted from intensity of WiFi signal and accelerations. *Information Sciences*, *233*, 162–182.

Astrain, J. J., Villadangos, J., Garitagoitia, J. R., de Mendívil, J. R. G., & Cholvi, V. (2006). Fuzzy location and tracking on wireless networks. In *Proceedings of the ACM International Workshop on Mobility Management and Wireless Access* (pp. 84–91).

Bahl, P., & Padmanabhan, V. N. (2000). RADAR: An in-building RF-based user location and tracking system. In *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies* (pp. 775–784).

Benavente-Peces, C., Puente, M., Domínguez-García, A., Lugilde-Rodríguez, M., de la Serna, E., Miguel, D., & García, A. (2009). Global system for localization and guidance of dependant people: Indoor and outdoor technologies integration. In *Ambient Assistive Health and Wellness Management in the Heart of the City* (pp. 82–89). volume 5597 of *Lecture Notes in Computer Science*.

Bisio, I., Cerruti, M., Lavagetto, F., Marchese, M., Pastorino, M., Randazzo, A., & Sciarrone, A. (2014). A trainingless wifi fingerprint positioning approach over mobile devices. *IEEE Antennas and Wireless Propagation Letters*, *13*, 832–835.

Bisio, I., Lavagetto, F., Marchese, M., & Sciarrone, A. (2016). Smart probabilistic fingerprinting for wifi-based indoor positioning with mobile devices. *Pervasive and Mobile Computing*, *31*, 107–123.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorization. *Information Fusion*, *6*, 5–20.

Calderoni, L., Ferrara, M., Franco, A., & Maio, D. (2015). Indoor localization in a hospital environment using random forest classifiers. *Expert Systems with Applications*, *42*, 125–134.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, *3*, 1–27.

Camposa, R. S., Lovisolob, L., & de Campos, M. L. R. (2014). Wi-fi multi-floor indoor positioning considering architectural aspects and controlled computational complexity. *Expert Systems with Applications*, *41*, 6211–6223.

Campuzano, F., Sánchez, A., & Botía, J. A. (2015). Hybrid indoor location: Simultaneous zone and coordinates based location for AAL environments with 802.11 fingerprinting technology. *Journal of Ambient Intelligence and Smart Environments*, *7*, 315–327.

Cardieri, P., & Rappaport, T. (2001). Statistical analysis of co-channel interference in wireless communications systems. *Wireless Communications and Mobile Computing*, *1*, 111–121.

Caso, G., & Nardis, L. D. (2015). On the applicability of multi-wall multi-floor propagation models to wifi fingerprinting indoor positioning. In *Future Access Enablers for Ubiquitous and Intelligent Infrastructures* (pp. 166–172). volume 159 of *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*.

Chen, T. (2016). A fuzzy integer-nonlinear programming approach for creating a flexible just-in-time location-aware service in a mobile environment. *Applied Soft Computing*, *38*, 805–816.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the International Conference on Machine Learning* (pp. 115–123).

Cordón, O., & Trawiński, K. (2013). A novel framework to design fuzzy rule-based ensembles using diversity induction and evolutionary algorithms-based classifier selection and fusion. In *Advances in Computational Intelligence* (pp. 36–58). volume 7902 of *Lecture Notes in Computer Science*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Dharne, A., Lee, J., & Jayasuriya, S. (2006). Using fuzzy logic for localization in mobile sensor networks: simulations and experiments. In *Proceedings of the American Control Conference* (pp. 2066–2071).

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 139–157.

Dietterich, T. (2001). Ensemble methods in machine learning. In J. Kittler, & F. Roli (Eds.), *Multiple Classifier Systems* (pp. 1–15). Springer volume 1857 of *Lecture Notes in Computer Science*.

Elnahrawy, E., Li, X., & Martin, R. P. (2004). The limits of localization using signal strength: A comparative study. In *Proceedings of the Annual IEEE Communications Society Conference on Sensor Ad Hoc Communications and Networks* (pp. 406–414).

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning* (pp. 148–156). Bari.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

García-Valverde, T., García-Sola, A., Hagras, H., Dooley, J., Callaghan, V., & Botía, J. A. (2013). A fuzzy logic-based system for indoor localization using WiFi in ambient intelligent environments. *IEEE Transactions on Fuzzy Systems*, *21*, 702–718.

Garcia-Villalonga, S., & Perez-Navarro, A. (2015). Influence of human absorption of Wi-Fi signal in indoor positioning with Wi-Fi fingerprinting. In *International Conference on Indoor Positioning and Indoor Navigation* (pp. 1–10).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, *11*, 10–18.

Hammadi, O. A., Hebsi, A. A., Zemerly, M. J., & Ng, J. W. P. (2012). Indoor localization and guidance using portable smartphones. In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 337–341). volume 3.

Hernández, N., Alonso, J. M., & Ocaña, M. (2016). Hierarchical approach to enhancing topology-based WiFi indoor localization in large environments. *Journal of Multiple-Valued Logic and Soft Computing*, *26*, 1–24.

Ho, T. (1998a). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 832–844.

Ho, T. (1998b). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 832–844.

Hühn, C., & Hüllermeier, E. (2009). FURIA: An algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, *19*, 293–319.

Jedari, E., Wu, Z., Rashidzadeh, R., & Saif, M. (2015). Wi-Fi based indoor location positioning employing random forest classifier. In *Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation* (pp. 1–5).

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254.

Kelley, K. J. (2014). Wi-Fi location determination for semantic locations. *The Hilltop Review*, *7*, 57–69.

Kibler, D., & Aha, D. (1987). Learning representative exemplars of concepts: An initial case study. In *Proceedings of*

the International Workshop on Machine Learning (pp. 24–30).

Kuncheva, L. (2001). Combining classifiers: Soft computing solutions. In S. Pal, & A. Pal (Eds.), *Pattern Recognition. From Classical to Modern Approaches* chapter 15. (pp. 427–452). Singapore: World Scientific.

Kuncheva, L. I., & Rodríguez, J. J. (2007). Classifier ensembles with a random linear oracle. *IEEE Trans. Knowl. Data Eng.*, *19*, 500–508.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).

Malagon-Soldara, S. M., Toledano-Ayala, M., Soto-Zarazua, G., Carrillo-Serrano, R. V., & Rivas-Araiza, E. A. (2015). Mobile robot localization: A review of probabilistic map-based techniques. *IAES International Journal of Robotics and Automation*, *4*, 73–81.

MATLAB (2016). Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

Mehdiyev, N., Krumeich, J., Werth, D., & Loos, P. (2015). Sensor event mining with hybrid ensemble learning and evolutionary feature subset selection model. In *IEEE International Conference on Big Data* (pp. 2159–2168).

Mo, Y., Zhang, Z., Lu, Y., Meng, W., & Agha, G. (2014). Random forest based coarse locating and KPCA feature extraction for indoor positioning system. *Mathematical Problems in Engineering*, *2014*, 1–8.

Neves, A. M., Carvalho, A., & Ralha, C. (2014). Agent-based architecture for context-aware and personalized event recommendation. *Expert Systems with Applications*, *41*, 563–573.

Pedrycz, W., & Kwak, K. (2006). Boosting of granular models. *Fuzzy Sets and Systems*, *157*, 2934–2953.

Rodríguez, J. J., & Kuncheva, L. I. (2007). Naïve bayes ensembles with a random oracle. In *Multiple Classifier Systems* (pp. 450–458). volume 4472 of *Lecture Notes in Computer Science*.

Schiller, J., & Voisard, A. (2004). *Location Based Services*. Morgan Kaufmann.

Stella, M., Russo, M., & Begušić, D. (2014). Fingerprinting based localization in heterogeneous wireless networks. *Expert Systems with Applications*, *41*, 6738–6747.

Tang, F., Li, J., You, I., & Guo, M. (2015). Long-term location privacy protection for location-based services in mobile cloud computing. *Soft Computing*, (pp. 1–13).

Tesoriero, R., Tebar, R., Gallud, J., Lozano, M., & Penichet, V. (2010). Improving location awareness in indoor spaces using rfid technology. *Expert Systems with Applications*, *37*, 894–898.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic Robotics*. Intelligent robotics and autonomous agents series (4th ed.). MIT Press.

TNS, Kantar Group (2012a). How mobile raises the bar for brand communications. Accessed on October 2016.

TNS, Kantar Group (2012b). Mobile life study. Accessed on October 2016.

Torres-Sospedra, J., Montoliu, R., Mendoza-Silva, G. M., Belmonte, O., Rambla, D., & Huerta, J. (2016). Providing databases for different indoor positioning technologies: Pros and cons of magnetic field and wi-fi based positioning. *Mobile Information Systems*, *216*, 1–22.

Trawiński, K., Cordón, O., & Quirin, A. (2011). On designing fuzzy rule-based multiclassification systems by combining FURIA with bagging and feature selection. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *19*, 589–633.

Trawinsky, K., Alonso, J. M., & Cordón, O. (2015). Applying random linear oracles with fuzzy classifier ensembles on WiFi indoor localization problem. In *Enric Trillas: Passion for Fuzzy Sets. A Collection of Recent Works on Fuzzy Logic* (pp. 277–287). volume 322 of *Studies in Fuzziness and Soft Computing*.

Trawinsky, K., Alonso, J. M., & Hernández, N. (2013). A multiclassifier approach for topology-based WiFi indoor localization. *Soft Computing*, *17*, 1817–1831.

Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion*, *6*, 83–98.

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, *3*, 209–235.

Werner, M. (2014). *Indoor Location Based Services. Prerequisites and Foundations*. Springer.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Data Management Systems Series (3rd ed.). Morgan Kaufmann.

Wu, B., Jen, C., & Chang, K. (2007). Neural fuzzy based indoor localization by Kalman filtering with propagation channel modeling. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 812–817).

Yim, J. (2008). Introducing a decision tree-based indoor positioning technique. *Expert Systems with Applications*, *34*, 1296–1302.

Youssef, M., & Agrawala, A. (2003). Small-scale compensation for WLAN location determination systems. In *Proceedings of the IEEE Wireless Communications and Networking* (pp. 1974–1978). volume 3.

Youssef, M., & Agrawala, A. (2008). The Horus location determination system. *Wireless Networks*, *14*, 357–374.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zhou, Z. (2005). Ensembling local learners through multimodal perturbation. *IEEE Transactions of Systems, Man, and*

Cybernetics, Part B: Cybernetics, *35*, 725–735.

Zhu, S., Sun, K., & Du, Y. (2015). A multi-classifier-based multi-agent model for Wi-Fi positioning system. In *Proceedings of the 4th International Conference on Computer Engineering and Networks* (pp. 1299–1305). Springer.