Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 251–260

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Feature selection on database optimization for Wi-Fi fingerprint indoor positioning

Guilherme Henrique Apostolo[a*], Igor Garcia Ballhausen Sampaio[a], José Viterbo[a]

[a] *Universidade Federal Fluminense, Av. Gal. Milton Tavares de Souza, s/n, Niterói-RJ 24210-346, Brazil*

Abstract

Indoor location-based services have become very popular, principally, because of its wide and valuable applications. On that context, Wi-fi fingerprinting based on the received signal strength indicator (RSSI) has become very popular,due the fact that RSSI values are easily acquired. On the Wi-fi fingerprint method, machine learning algorithms are trained on the constructed fingerprint database and then used on a new entry to give the indoor location based on its estimations. Choosing the correct machine learning algorithm is one of the main problems in the literature. However the database sizes used during the training phase is also one of the main concerns. In this paper, a proposed feature selection method used on the original UJIIndoorLoc database created a smaller version of it, with the 30 highest RSSIs after the APIDs responsible for then in descending order, and created even smaller database subsets. Both databases, the original UJIIndoorLoc database and ours, were split into smaller subsets that were used on the classification problem according the DESIP method proposed in [1]. Six machine learning algorithms were deployed for training and testing the two database subsets with the classification attributes modified for symbolic localization. The J48 with the AdaBoost iterative algorithm gave the best results on both database subsets. The minimized database subsets showed smaller elapsed time results for all the classifications that were done. The accuracy results show similar results for both database subsets, on building and floor classification. Although, on the region attribute, the database subset with 520 attributes got better accuracy results than the reduced one.

Keywords: Wi-fi Indoor Localization; Database Optimization; Machine Learning Algorithms; Feature Selection .

1. Introduction

Knowing the exact location of something or someone is a very important and valuable information nowadays. Traditionally, outdoor localization using GPS systems has been widely used and can reach a very high accuracy.

* Guilherme Henrique. Tel.: +55-21-99491-5828.
E-mail address: guilherme_apostolo@id.uff.br

1877-0509
10.1016/j.procs.2019.09.180

Most of the mobile devices come with built in GPS sensors [11]. However, GPS signal cannot penetrate well in indoor environment [5], making indoor localization not accurately possible with this technology. A lot of research and development in indoor localization systems has been made on the last decade due the impossibility of using the GPS systems with acceptable precision [10].

Indoor location-based services have become very popular, principally, because of its wide and valuable applications. Indoor location-based services are estimated to worth US$ 10 billion by 2020 [5]. Location detection of a product in a warehouse, medical personnel or equipment in a hospital, firefighters in a building on fire and a persons location inside a shopping mall or convention center are some examples of real-world application of indoor location-based services [7]. Indoor location-based systems can be build on different signals for positioning interpretations like Wi-fi, Bluetooth, radio-frequency identification (RFID) and ultrasound for example [5]. Indoor localization using wireless local area network (WLAN) techniques are one of the most studied approaches, mostly, due to the rapid spread of both WLANs and mobile devices [9]. From WLAN approaches Wi-fi fingerprinting based on the received signal strength indicator (RSSI) has become a popular approach, since RSSI fingerprint are easily acquired [13].

Wi-fi fingerprint indoor positioning systems based on the RSSI values works associating the RSSI measured value to an indoor position. Wi-fi fingerprint is conduct in two phases, the offline phase followed by the online phase. On the offline phase reference samples, containing the RSSIs values of all detected access points (APs) and the reference coordinates of the known location, are collected and stored. The collection of reference samples forms the fingerprints database of the surveyed area. On the online phase, a user measures its RSSIs values and the corresponding APs that through algorithms, compares it to a location of the fingerprinted map and gives an estimated location [11] [14].

The UJIIndorLoc training database [13] provides 19938 reference samples containing RSSIs values from 520 APs collected by 20 users, using 25 different mobile devices. This database covers 933 different places from 3 different buildings, with 4 or 5 floors depending on the building. One relevant fact about the database is that none of the reference samples detected RSSIs values for all the 520 APs. As a matter of fact, the maximum number of RSSIs values for APs detected was 51. That way, most of the database is filled with blank, out of range, values.

On A Comparative Study on Machine Learning Algorithms for Indoor positioning [1] the database was divided into new subsets, the method proposed is named as Deductive Separation for Indoor Positioning (DESIP) [1]. Basically, the classification is divided in three steps where each of the symbolic location attributes, building, floor and a combination of Relative Position and Space ID attributes, called Region, is singly classified. These database subsets were trained and tested using different machine learning algorithms.

In this study, the DESIP method and six machine learning algorithms will be used on the database to create database subsets versions of the UJIIndorLoc training database [13]. We will perform some feature selection methods where we will filter the database into a new, and smaller, one that now contain the highest 30 RSSIs values and theirs corresponding 30 AP identification (APId) numbers instead of 520 RSSIs values. So, the database will no longer have 520 RSSIs values as features, each one representing its own corresponding APs, but only the highest 30 RSSIs values and 30 APId, each APID corresponding to the APs numbers that gave that measures. That way, the new filtered database will have 60 values, instead of 520 values of the original one. This new configuration of the database minimizes the stored data size making easier to upload and store the fingerprints registers, which is a problem related to Wi-Fi fingerprinting indoor location systems. Employing six different machine learning algorithms, using the WEKA library [4] for the experiments, it will be possible to determine the accuracy and computational time of the new filtered database subsets and then compare it to the results obtained by using the original database subsets versions of the UJIIndorLoc training database [13].

The rest of this paper is organized as follows: Section 2 will present related works found in the literature. Section 3 introduces the changes and modifications made in the original database and how is the database in this study made. Section 4 explains how is the experiment made. Section 5 shows the results obtained by the experiment.Finally, Section 6 gives the conclusion.

2. Related works

In WiGEM: A learning-based approach for indoor localization [3], a Wi-fi fingerprint localization algorithm is proposed, where a Gaussian Mixture Model and an Expectation Maximization algorithm are deployed to estimate de user position. On WiGEM approach the RSSIs values are collected, by the nearby APs, and stored in a vector. A

Gaussian Mixture Model is created and then used to estimate the most likely position that could have generated that RSSIs vector. The Expectation Maximization algorithm is used to learn the maximum likelihood of the parameters on the estimated model and used to recalculate these parameters on the propagation model, used for location estimation. The key point of WiGEM model is that the propagation model is recalculated at each new vector of RSSIs that arrives from the user being much more robust mobility, device and power variabilities. It also eliminates the offline training phase of the Wi-fi fingerprint method, saving computational time and space.

The UJIIndorLoc [13] created a wif-fi fingerprint database with 19938 reference samples containing RSSIs values from 520 APs collected by 20 users, using 25 different mobile devices. The database contains 529 features and it covers an area of 108703 m2 including 3 buildings. In [13], They have also used a KNN algorithm to provide a baseline for further comparisons. In [9], the UJIIndorLoc have been used for position estimation based on KNN and WKNN algorithms. The authors made the localization estimation into parts, where the building was estimated first, followed by the floor. The coordinate estimation is the centroid of the K nearest neighbors, calculated using the Euclidean distance.

The paper Smart probabilistic approach with RSSI fingerprint for indoor localization [11] propose a probabilistic approach using the UJIIndorLoc training database [13]. On that approach, first, wi-fi fingerprint probabilistic algorithms have been used for building and floor classification. After having the floor classification done, the database is divided into subsets, each one for each floor, containing only the APs that gave a valid RSSI for that floor. The Shannons Entropy is calculated for each one of the APs on the new subset and the APs that have a null entropy are removed. The last step is the position estimation based on the similarity of the test and trained fingerprints subsets. In [2], the constructed wi-fi fingerprinted radio map database built on the offline phase is reduced by APs and fingerprint filtering. Some rules based on the RSSI value are responsible for setting threshold values for a statistical method that will determine the number of useful APs to construct the database. The fingerprint filtering is made with the help of machine learning algorithms. Machine learning algorithms are used to classify the fingerprints into the rooms and those who were wrongly classified, are marked as outliers or bad samples and removed from the database. That approach reduced the size of the positioning radio map while maintaining the positioning precision result. In [6], the authors constructed a localization approach using Deep Belief Network algorithms where only a small fraction of labeled fingerprints is needed to construct an unsupervised deep feature learning model. The model use a small fraction of labeled fingerprints combined with unlabeled fingerprints to give a location estimation. The results of this approach had very similar results to tests using theirs same deep feature learning with a much bigger database.

In A Comparative Study on Machine Learning Algorithms for Indoor positioning [1] the UJIIndorLoc training database gets divided into smaller database subsets using the DESIP method. These subsets are then trained, by six different machine learning algorithms, and used to estimate the symbolic position of a measured point containing 520 RSSIs values. The estimation has been done in parts, where each element of the full symbolic location is distinctly classified, first the building, then the floor and lastly by Region, combination of both Relative Position and Space ID. The estimated locations, given by the machine learning algorithms, are compared, using accuracy and computational time measures, to determine the most satisfactory machine learning algorithm for a Wi-fi fingerprint-based indoor positioning system.

3. Proposed feature selection method

The UJIIndorLoc training database[13] is a real world RSSI collection database that has 19938 instances of 933 different locations, collected by 25 different mobile devices. The authors created two Android applications, one for training and another one for validation. These applications use a map based service with geographical information of the indoor areas. The training database construction app captures the RSSI values and other geographical information, inputted by the user according to the map provided, and send it to the server. The validation application captures the RSSI values of the detected WAPs and send, only, the detected WAPs and their corresponding RSSI value to a central server that, then, answers with the geographical indoor position to the user. That way, the indoor localization procedure through the fingerprints is entirely made on the server and transparent to the user.

There are several different implementations on how to deal with this database that has 529 features, 520 RSSIs values and 9 labeled attributes about the fingerprint, such as latitude and longitude for example, and estimate the indoor position in the server. Estimating the indoor localization with machine learning classifier algorithms is one

of them. Based on the Wi-Fi Fingerprint approach and its issues, we propose three heuristics rules on the RSSIs attributes part of the database to filter it, from 520 to 60 attributes. On the labeled attributes, regarding characteristics about the fingerprints, a filtering process have also been done, in order to adapt for the DESIP method and make the classification process as wide as possible. These feature selections proposals make the database smaller, which will enhance the overall time processing metric. After the new database formation, the new database subsets are formed, each one with its own particularity for the classification it will perform. The three heuristics feature selection filters, the labeled attribute filter and the database subsets labeling rule are described in details below.

3.1. RSSIs feature selection methods

The UJIIndorLoc training database [13] provides 19938 instances where there are 520 RSSIs attributes. Each of these attributes correspond to one of the RSSIs readings of the 520 APs used on the study. However, none of the captured fingerprints had detected RSSIs values for all the 520 APs. In fact, the maximum number of APs detected in a fingerprint was 51 and some fingerprints have detected no APs at all, no Wi-Fi coverage. Figure 1 shows the frequency distribution of APs detected on one fingerprint. It is possible to observe that most of the fingerprints on the database have around 18 APs detected. With that being said, if all the fingerprints on the database have 51 or less APs detected, it means that most of the 520 attributes are filled with null values. That makes that database with a extensive amount of null data.

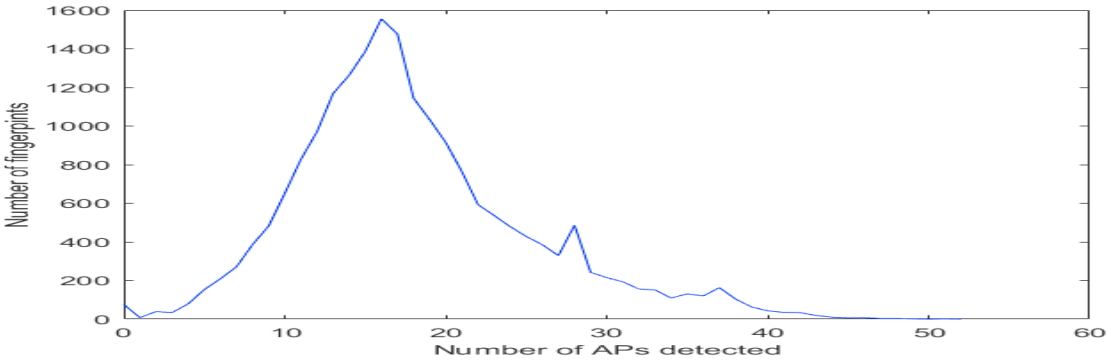

Fig. 1. Frequency distribution of the number of APs detected on a single fingerprint.

It is possible to calculate that 18584, 93,21%, of the fingerprints on the database have 30 or less APs detected. With that being said, most of the database, 93,21% of it, have more than 490 attributes set as null values, the rest of the undetected APs. Therefore, if we remove all the set as null attributes for each fingerprints, we are going to have a reduction on the 520 features, with no impact on the instances. Since most of the attributes must be removed and only a few APs for each fingerprint will remain, we have to rewrite the database. That is done by rewriting the RSSIs non null values inserting the APs that gave that specific values before it. Since each of the 520 APs were named, from WAP001 to WAP520, we can create a attribute called APID, AP identification. This APID is a numeric value that represent the number appearing on the attributes names, APID = 1 represent the WAP001 and so on.

The second rule applied on the database is the rearrangement of it according the descending order of the RSSIs values. Bigger RSSIs values means that the mobile device, responsible for the reading, is closer to it. So, APs that give the biggest RSSIs readings for one fingerprint, are the one that are closer to the device and therefore more useful to estimate its position. So the database is rearranged by the descending order of RSSIs values, where the APs responsible for the readings are put together to the corresponding RSSIs value. affecting

The third rule is that only the first 30 RSSIs values should be left on the database, representing a reduction on the database to only 60 attributes. Since 93,21%, of the fingerprints on the database have 30 or less APs detected, only 6,79% of the fingerprints readings are affected. That represents a small amounts of the total number of instances but a great reduction on the remaining number of null values. These rule can cause some undesirable accuracy performance loss, since it is removing information from some instance which can lead to wrong classifications, but can cause

significant improvements on the elapsed time metric that we want to reduce. It is important to highlight the fact that not all of the fingerprints have 30 valid RSSI readings. So, for fingerprints with less than 30 valid RSSI values, the remaining attributes, after the smallest RSSI reading,are filled with null values, APID = 0 and RSSI = -500. Table 1 gives a example of how a fingerprint reading is represented on the new dataset, as you can see AP 173, AP1, gave the highest RSSI reading, RSSI1, while AP29, AP30, RSSI29 and RSSI30 does not represent a valid reading.

Table 1. Example of one fingerprint on the new dataset with only 60 attributes

AP1	RSSI1	AP2	RSSI2	...	AP29	RSSI29	AP30	RSSI30	BuildingID	Floor	RegionID
173	-53	172	-54	...	0	-500	0	-500	B1	F2	S106R2

3.2. Labeled attributes filtering

The UJIIndorLoc training database has 9 labeled attributes that identifies different information about the fingerprints collected. Although all of them carry important information, some of them are not necessary for the scope of this paper and others must be rewritten. The database with all the 520 RSSIs values, hereafter called original database, and the filtered reduced one with only 60 attributes, hereafter called filtered database, are subjected to these changes and will contain the same labeled attributes.

Since we are interested on a classification problem with symbolic location, where a geographical area is translated into a label that express that area, longitude and latitude features are not necessary in the scope. Different phones, even from the same model, will display different RSSI readings for the same position and time [8] [3]. Indoor facilities have also a dynamic nature, where people and objects are always changing position, that causes changes on the radio waves propagation which leads to changes on the signal strength map over time [12]. With that being said, Phone ID, User ID and Time stamp features carries valuable data for indoor positioning studies that aim to mitigate their influences over the data. However, we will not use these features on that study, since we want create a more broad approach where just the RSSIs readings and symbolic location are given as features for the indoor positioning classification. That way Longitude, Latitude, User ID, Phone ID and Time stamp are, therefore, removed from the database.

The Space ID and Relative position attributes are numeric and identify the particular space, a room or an office for example, and position, inside or outside at the corridor, where the readings were taken. Since the DESIP method, in [1], concatenated both attributes to crate a new one, called Region, the same will be done here. Table 2 shows how the Region attribute was made in this paper.

Table 2. Relation between Region and Space ID, Relative Position

Space ID	Relative Position	Region
111	1	S111R1
101	2	S101R2

3.3. Database subsets descriptions

There are several database subsets created by the split of the databases, the original and the filtered. Each database subset is important for each step of the DESIP method. Since there are several database subsets, a name standardization was made in order to help on the understating of what each subset is classifying. Table 3 shows how the names were created and what each one of them is responsible for.

4. Experiment Description

Now that the filtered version of the database, with 60 attributes, is done and the original, with 520 attributes, is modified to have the same 3 labeled attributes, the experiment can start. The experiment consists of performing

Table 3. Database subset name description

Database subset name	Description
B	Building classification
Bx	Floor classification from the Building number x
Bx_Fy	Region classification from the building number x and floor number y

the classification of both, filtered and original, databases subsets, constructed according to DESIP method, using six different machine learning algorithms on WEKA[4]. Our aim was to evaluate the performance on the metrics, accuracy and elapsed time, of the filtered database subsets using the classification algorithms for indoor positioning. This evaluation was done by comparing the metric results of each filtered database subset with the corresponding metric results obtained with the original database subset.

Both databases will be split into database subsets containing the same labeled attributes according to the DESIP method. DESIP classification method is divided into steps where on the first step, only the RSSIs values and building information are used for building classification. After classifying the building, the database is split, based on the building information, creating 3 new database subsets, one for each building, containing the RSSIs values, Building ID and Floor ID that are used for floor classification. The last stage is splitting the database based on the floor information, one for each floor from each building, creating new database subsets containing the RSSIs values, Building ID, Floor ID and Region that are used for region classification.

Six different machine learning algorithms, J48, BayesNet, KNN, SMO, Adaboost with J48 and Bagging with J48 were comparatively used. These algorithms were used for training and testing the original and filtered databases subsets, created for this paper. All of the mentioned machine learning algorithms were used with the standard, automatic, selections parameters of the Weka library [4]. These machine learning algorithms were choose for presenting the best results, based on the results obtained on the paper "A comparative study on machine learning algorithms for indoor positioning" [1]. The training database subsets, used during the training phase, correspond to 95% of the full database, while the reminder, 5%, compose the test database subset, used for testing.

Accuracy and elapsed time are used as results metrics, in this work, for performance comparison reason. Accuracy gives the percentage of the number of instances correctly classified over the total number of instances after the test phase. It is a important and simple metric for machine learning algorithms comparison and for indoor positioning, since we want to know the algorithm with the highest correctness, and that is why it is used as the main comparison tool in this work. Elapsed time gives us the time taken, from start to end, for each one of the machine learning algorithms to train and test the subset. When talking about indoor positioning, time is a really precious since location based systems are meant to work in real time, that being said the algorithms can not take to long to give an appropriate answer and that is why elapsed time is the second result metric used in this paper. All the results showed on the tables are equivalent to the average accuracy and elapsed time of 10 different iterations of the algorithms.

5. Experiment Results

Table 4 gives the accuracy results of the six machine learning algorithms for all the original database subsets, all DESIP steps sequentially. It is possible to notice that building classification, B subset, has reached almost 100% accuracy in almost all algorithms, that is probably due to the huge amount of training instances and the small number of classification classes, only 3. Table 4 also tells us that floor classification, Bx subsets, is done with a high accuracy in most cases, more than 90%, reaching 98% on the best ones. Region classification, Bx_Fy subsets, showed the weakest accuracy results when compared to floor and building classification steps, that is probably due to the smaller number of available instances and bigger number of classes. Adaboost with J48 gave almost always the best result for Region classification, being surpassed only by the BayesNet algorithm on some subsets.

Table 4. Machine Learning algorithms accuracy results for the original database subsets

	J48	BayesNet	KNN	SMO	Adaboost J48	Bagging J48
B	99,80	99,90	99,70	99,90	99,90	99,50
B0	95,80	95,80	94,28	91,99	98,47	98,09
B0_F0	62,26	58,49	66,04	66,04	69,81	77,36
B0_F1	70,59	75,00	64,71	64,71	77,94	76,47
B0_F2	51,39	70,83	55,56	61,11	65,28	72,22
B0_F3	72,86	71,43	65,71	64,29	74,29	72,86
B1	98,08	95,38	98,08	96,15	99,61	99,23
B1_F0	73,53	67,65	69,12	54,41	86,77	79,41
B1_F1	67,57	72,97	62,16	64,87	66,22	70,27
B1_F2	71,43	84,29	85,71	77,14	90	84,29
B1_F3	70,21	80,85	72,34	59,58	70,21	72,34
B2	95,58	93,26	98,74	96	97,68	97,68
B2_F0	77,32	74,23	77,32	75,26	78,35	78,35
B2_F1	69,45	76,85	73,15	76,85	77,78	75,96
B2_F2	72,15	84,81	72,15	70,89	83,54	79,75
B2_F3	59,26	77,78	72,59	65,19	77,78	72,59
B2_F4	50,91	65,45	58,18	56,36	50,91	56,36

The accuracy results of the six machine learning algorithms for all DESIP steps, all the filtered database subsets sequentially, are shown in Table 5. The building classification on the filtered, 60 attributes, B subset gave worse results on all the six algorithms, however the Adaboost with J48 gave a result of 99,60% which is very close to the accuracy results achieved with the original database subset. Floor classification with the filtered database, Bx subsets, also gave worse results than the original database ones. On floor classification we start to notice a more clear distance between the accuracy results of the original subsets and the filtered ones. However the greatest distance on the best accuracy for floors classification are no more than 3% which still made them relatively close.The region classification also showed worse results, probably due to a combination of smaller number of attributes, smaller number of available instances and a bigger number of classes. Adaboost with J48 gave almost always the best result for all filtered database subsets classification, being surpassed only by the BayesNet algorithm on some subsets and once by bagging with J48.

Table 5. Machine Learning algorithms accuracy results for the filtered database subsets

	J48	BayesNet	KNN	SMO	Adaboost J48	Bagging J48
B	98,50	99,30	83,45	74,32	99,60	99,00
B0	92,75	89,31	56,49	51,15	95,80	94,66
B0_F0	64,15	69,81	37,74	28,30	83,02	83,02
B0_F1	60,29	60,29	38,24	44,12	76,47	72,06
B0_F2	56,95	47,23	25,00	31,95	61,12	58,34
B0_F3	54,29	32,86	37,14	34,29	65,71	57,14
B1	95,00	91,92	79,62	67,69	98,85	96,92
B1_F0	66,67	59,42	36,23	33,34	73,91	65,22
B1_F1	59,72	69,45	48,61	47,22	61,11	66,67
B1_F2	65,71	78,57	42,86	34,29	77,14	70
B1_F3	70,21	72,34	36,17	27,66	70,21	78,72
B2	93,90	82,53	52,00	48,42	97,06	96,21
B2_F0	75,23	53,61	38,15	37,11	74,23	77,32
B2_F1	51,85	34,26	32,41	43,52	64,82	60,19
B2_F2	67,09	60,76	29,11	39,24	78,48	77,22
B2_F3	51,11	50,37	26,67	41,48	65,19	67,41
B2_F4	36,36	49,09	30,91	32,73	36,36	43,64

The Figure 2 shows a graphic comparison between the best accuracy result, out of the six machine learning algorithms used,on each of the classification steps for both original and filtered database subsets.Through Figure 2 it is possible to compare the best accuracy scenario for both databases and it is possible to see that for building and floor classification the results are very alike. Such behavior it is not notice on the region classification, where the relation between the original and filtered database shows an more unbalanced scenario. On the region classification the original database have, mostly, a better accuracy, but the filtered one have some very close accuracy results on most of the subsets and one of the filtered subset, for region classification, has even a superior accuracy percentage.

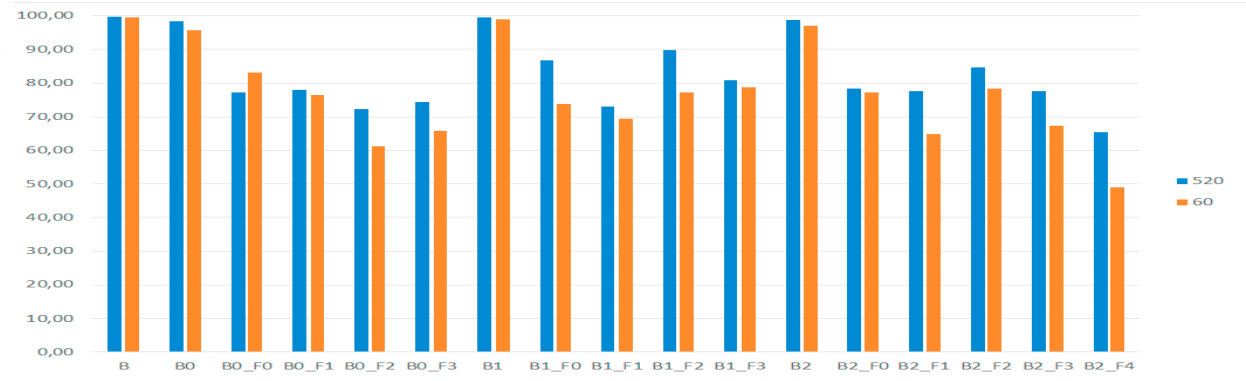

Fig. 2. Comparison between the best accuracy results achieved for the original, labeled as 520, and filtered, labeled as 60, database on each subset

Tables 6 and Table 7 show the elapsed time for the six machine learning algorithms on all database subsets, Table 6 for the original database and Table 7 for filtered one. In general, it is possible to notice that the BayesNet and J48 algorithm have a smaller elapsed time than the others, followed by the KNN and Adaboost with J48 for both databases. It is also very eminent the fact the the elapsed time on the filtered subsets, with few exceptions, are much smaller than the ones on the original. That is probably due the smaller size of the database which make the size and processing time of the algorithms smaller both for training and testing phases.

Table 6. Machine Learning algorithms elapsed time results for the original database subsets

	J48	BayesNet	KNN	SMO	Adaboost J48	Bagging J48
B	43.04	8.54	17.17	3.07	170.82	235.86
B0	3.95	0.87	0.69	7.91	44.39	31.41
B0_F0	0.33	0.22	0.02	4.32	3.63	2.75
B0_F1	0.45	0.34	0.05	5.78	4.97	3.60
B0_F2	0.56	0.42	0.06	4.46	6.08	4.63
B0_F3	0.53	0.19	0.06	4.08	5.73	4.27
B1	3.6	1.28	1.40	6.24	54.87	28.09
B1_F0	0.05	0.42	0.08	6.96	5.3	3.88
B1_F1	0.63	0.33	0.09	2.85	3.38	3.98
B1_F2	0.42	0.36	0.09	4.32	5.01	4.00
B1_F3	0.31	0.12	0.05	1.50	1.39	2.33
B2	8.2	1.71	3.86	10.22	66.21	51.39
B2_F0	0.06	0.45	0.16	4.47	7.3	5.23
B2_F1	0.91	0.68	0.22	7.69	11.24	7.72
B2_F2	0.53	0.41	0.14	4.96	6.25	4.11
B2_F3	1.29	1.04	0.33	11.77	16.28	10.60
B2_F4	0.3	0.26	0.06	3.61	2.06	2.8

Table 7. Machine Learning algorithms elapsed time results for the filtered database subsets

	J48	BayesNet	KNN	SMO	Adaboost J48	Bragging J48
B	2.08	0.69	4.33	135.89	15.22	17.14
B0	0.08	0.56	0.46	7.49	4.21	2.47
B0_F0	0.08	0.06	0.01	3.54	0.93	0.58
B0_F1	0.11	0.03	0.08	3.24	1.65	0.88
B0_F2	0.16	0.04	0.03	3.72	2.43	1.06
B0_F3	0.11	0.06	0.03	3.49	1.78	1.25
B1	0.27	0.08	0.31	7.84	4.24	2.59
B1_F0	0.13	0.05	0.03	3.87	1.94	0.8
B1_F1	0.11	0.04	0.05	1.93	1.21	0.85
B1_F2	0.11	0.05	0.02	3.54	1.72	1.11
B1_F3	0.03	0.02	0	1.16	0.34	0.38
B2	0.77	0.21	0.99	55.3	8.74	2.29
B2_F0	0.16	0.06	0.03	3.86	1.44	1.23
B2_F1	0.25	0.05	0.06	6.48	3.65	1.86
B2_F2	0.13	0.05	0.03	4.35	2.06	1.10
B2_F3	0.28	0.10	0.09	10.27	5.60	2.76
B2_F4	0.06	0.04	0.02	2.87	0.19	0.61

6. Conclusion

In this paper, the proposed filtered, smaller, and the original database were split into subsets, each one for one attribute classification, then trained and tested with six different machine learning algorithms. Accuracy and elapsed time were compared between the subsets of both databases, each with its corresponding pair. The filtered database did well on building, with a difference not bigger than 0,30%, and floor, with a difference not bigger than 3,00%, classification using the J48 with the AdaBoost iterative algorithm. The filtered database even showed one accuracy

result that were better than the one of the original database subsets in one region classification. However, it was the original database subsets, with 520 attributes, that have better accuracy results in almost all algorithms, with significant better accuracy results in most of the regions classified. That result on the region attribute is probably due to a combination of the small amount of instances for that attribute classification and the loss of information on some of the instances. On the other hand, the filtered database subsets showed much smaller elapsed time, with fewer exceptions, in almost all algorithms, specially on the J48 with the AdaBoost iterative algorithm that gave the best accuracy results for this database. With that being said, we can assume that is possible to reduce the database from its original size with great improvements to the elapsed time, that the algorithm requires to train and give an indoor position location, with a low loss on its accuracy. Although, it is clear that minimizing the database, making it waste RSSIs values, can cause great losses on the accuracy results if just a few instances are available for training and testing.

References

[1] Bozkurt, S., Elibol, G., Gunal, S., Yayan, U., 2015. A comparative study on machine learning algorithms for indoor positioning, in: Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on, IEEE. pp. 1–8.

[2] Eisa, S., Peixoto, J., Meneses, F., Moreira, A., 2013. Removing useless aps and fingerprints from wifi indoor positioning radio maps, in: Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on, IEEE. pp. 1–7.

[3] Goswami, A., Ortiz, L.E., Das, S.R., 2011. Wigem: A learning-based approach for indoor localization, in: Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies, ACM. p. 3.

[4] 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.

[5] He, S., Chan, S.H.G., 2016. Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons. IEEE Communications Surveys & Tutorials 18, 466–490.

[6] Le, D.V., Meratnia, N., Havinga, P.J., 2018. Unsupervised deep feature learning to reduce the collection of fingerprints for indoor localization using deep belief networks, in: 2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN), IEEE. pp. 1–7.

[7] Liu, H., Darabi, H., Banerjee, P., Liu, J., 2007. Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 37, 1067–1080.

[8] Lui, G., Gallagher, T., Li, B., Dempster, A.G., Rizos, C., 2011. Differences in rssi readings made by different wi-fi chipsets: A limitation of wlan localization, in: Localization and GNSS (ICL-GNSS), 2011 International Conference on, IEEE. pp. 53–57.

[9] Moreira, A., Nicolau, M.J., Meneses, F., Costa, A., 2015. Wi-fi fingerprinting in the real world-rtls@ um at the evaal competition, in: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), IEEE. pp. 1–10.

[10] Njima, W., Ahriz, I., Zayani, R., Terre, M., Bouallegue, R., 2017a. Comparison of similarity approaches for indoor localization, in: Wireless and Mobile Computing, Networking and Communications (WiMob),, IEEE. pp. 349–354.

[11] Njima, W., Ahriz, I., Zayani, R., Terre, M., Bouallegue, R., 2017b. Smart probabilistic approach with rssi fingerprinting for indoor localization, in: Software, Telecommunications and Computer Networks (SoftCOM), 2017 25th International Conference on, IEEE. pp. 1–6.

[12] Pan, S.J., Zheng, V.W., Yang, Q., Hu, D.H., 2008. Transfer learning for wifi-based indoor localization, in: Association for the advancement of artificial intelligence (AAAI) workshop, The Association for the Advancement of Artificial Intelligence Palo Alto.

[13] Torres-Sospedra, J., Montoliu, R., Martínez-Usó, A., Avariento, J.P., Arnau, T.J., Benedito-Bordonau, M., Huerta, J., 2014. Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems, in: Indoor Positioning and Indoor Navigation (IPIN), 2014 International Conference on, IEEE. pp. 261–270.

[14] Wang, X., Gao, L., Mao, S., Pandey, S., 2017. Csi-based fingerprinting for indoor localization: A deep learning approach. IEEE Transactions on Vehicular Technology 66, 763–776.