Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 32 (2014) 453 – 460

Procedia Computer Science

5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014)

Evaluation of Four Classifiers as Cost Function for Indoor Location Systems

Carlos E. Galván-Tejada[a,*], Juan P. García-Vázquez[a], Enrique García-Ceja[a], José C. Carrasco-Jiménez[a], Ramón F. Brena[a]

[a] *Instituto Tecnologico de Monterrey, Monterrey, Nuevo León, México, Autonomous Agents in Ambient Intelligence.*

Abstract

In our previous research work, we proposed a methodology that uses magnetic-field and multivariate methods to estimate user location in an indoor environment. In this paper, we propose the use of this methodology to evaluate the performance of four different classification algorithms: Random Forest, Nearest Centroid, K Nearest Neighbors and Artificial Neural Networks; each classifier will be considered as a cost function of a genetic algorithm (GA) used in the feature selection process task of the methodology. The motivation to evaluate the algorithms of classification was that several ILSs use a classification algorithm in order to estimate the location of the user, but the classifiers performance vary from application to application. In order to evaluate the performance of each classification algorithm, the following issues were considered: (1) the time of the training phase to obtain the final classification algorithm; (2) the number of features needed for getting the model; (3) the type of the features from the final model; and (4) the sensitivity and specificity of the model. Our results indicate that Nearest centroid is the classfier algorithm that is best suited to be implemented in an end-user application given the obtained results on the evaluated criteria for the indoor location system (ILS).

Keywords: Indoor Location, Classifier Algorithms, Nearest Centroid, K Nearest Neighbors, Artificial Neural Networks, Multivariate Methods

1. Introduction

Indoor Location Systems (ILS) conform the basis of many ubiquitous services and as a consequence improvements in location systems are the focus of many reasearch works [1,2,3,4,5,6]. Locating users with high accuracy can offer a deal of services that range from turning on the lights in a user's room to targeting users with advertisments and special offers while in a shopping center. There are several approaches to locate a user in an indoor environment. For instance, indoor location systems rely on a variety of techniques and instruments to sense the environment, which can be combined to approximate the location of a user with higher precision.

Indoor location systems can be divided into four different categories based on the type of technology that mediates between the environment and the user. The categories are divided in i) those that exploit technology that was not originally intended for location systems per se, such as Bluetooth and RFID, which can be extended and adapted to

* Corresponding author. Tel.: +52-81-8358-2000
E-mail address: ericgalvan@uaz.edu.mx

doi:10.1016/j.procs.2014.05.447

complex indoor location systems[1,2]; ii) systems that reuse sensory devices included in conventional smart phones such as accelerometer sensors; iii) ILSs that rely on specialized infrastructure that consists of arrays of sensors fixed at specific locations, among which indoor light systems and WiFi access points play a crucial role[3,4]; and iv) systems that use natural environmental signals captured from natural phenomena such as magnetic-field and environmental audio[5,6]. Regardless of the information source (e.g. light, WiFi, magnetic-field), several ILSs use classification algorithms in order to estimate the location of the user, but it is important to point out that different classifiers behave differently when used in different scenarios of ILS development.

In our previous work, we explored the development of ILSs in which both artificially generated data (e.g. WiFi and Bluetooth), and natural data (e.g. magnetic-field and indoor light intensity), were studied in more depth.[6,7,8]. The implementation of the indoor location systems allowed us to propose a general methodology based on multivariate models that approximate a user's location in an indoor environment such as a building or a shopping center[8]. In this paper, we propose the use of this methodology to evaluate the performance of four different classification algorithms. Each classifier will be considered as part of our ILS methodology. The aim of this work is to evalute a number of classifiers with different characteristics, focusing on the use of magnetic-field data to approximate the location of an individual. Magnetic-field, as opposed to infrastructure that relies on Bluetooth and Wifi, does not require a predefined arrangement of devices or the installation of devices to generate data.

In order to evaluate the classification algorithms, a Genetic Algorithm was used to select the classifier that optimizes the cost function. For this, we take into account four factors that impact the performance of the algorithms: i)the time of the training phase to obtain the final classification algorithm; ii)the number of features needed for building the model; iii) the type of the features from the final model; and iv) the sensitivity and specificity of the model.

The paper is organized as follows: section 2 introduces relevant information about the selected classification algorithms and the methodology used to estimate the user location in an indoor environment. The experiments are explained in section 3. In section 4, the results are analized and discussed, and the conclusions and future work are presented in section 5.

2. Background

In this section, the classification methods evaluated in this work are explored, as well as an explanation of the methodology used to evaluate indoor location systems.

2.1. Classification methods

The set of methods evaluated consists of four different classification algorithms: i) Random Forest (RF), ii) Nearest Centroid (NC), iii) K Nearest-Neighbors (K-NN) and iv) Artificial Neural Networks (ANN). These approaches were considered for this work in order to evaluate the performance of methods that belong intrinsically to different categories. For example, Random Forest follows a decision based approach[9], while K-NN is an instance based method[10]. On the other hand, Artificial Neural Networks are connectionist and the Nearest Centroid method can be considered a hybrid approach that encompasses an instance based approach as well as a statistical one.

2.1.1. Random Forest

This classifier was proposed by Breiman et al.[11]; it provides tree ensambles that depends on the values of a random feature vector and provides the same distribution to all the trees included in the forest. The decision of this classifier depends on the decision of several trees. A random forest differs each time that is performed given the randomness introduced in the tree building process. This randomness can be modified to acquire specific purpose forests to certain problems[12]. This classifier is commonly used because the interpretation involving logical relation between variables, values, and classes is very simple[13]. An example of an ILS based on feature extraction that generates a classification model to estimate the location based in a Random Forest as classifier was presented in our previous research work[14].

2.1.2. Nearest Centroid

For a given set of samples of the same class C, the centroid of C is defined as the mean or median value[13] of the corresponding samples in C. The nearest centroid for an unknown sample is the centroid whose euclidean distance

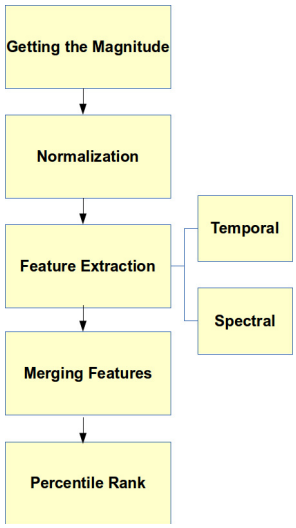

Figure 1. Methodology to obtain a ILS

is minimum. For instance, Chen et al.[15] proposed the use of an optimized nearest centroid classifier to increase the estimation accuracy while reducing the power consumption by intelligently selecting the number of WiFi access points (APs) used for location estimation.

2.1.3. K Nearest Neighbors (K-NN)

For a given sample, its nearest neighbor is the sample that has the minimum distance to it. Hence, a measure of distance is needed to determine how close two samples are. Because of its generality, euclidean distance is often preferred as a distance measure. Although K-NN is widely used in a variety of applications, it has also been used in indoor location systems. LANDMARC[16] is an example of an indoor location system that uses Radio Frequency Identification (RFID) as information source; this ILS was implemented using K-NN approach to find an unknown tracking RFID tag in a pre-deployed network of radio frequency (RF) readers.

2.1.4. Artificial Neural Networks

Artificial Neural Networks (ANN) are based in the biological neural networks. A neuron is a biological cell that processes information in the brain. These cells have two types of branches to connect with other neurons called axon and dendrites. The basic function of neural cells is done when a neuron receives signlas from other neurons througth its dendrites and transmit the signal to other neurons along the axon. Eventually, when a signal is passing throuth the same neurons, it can be learned by the brain[17]. Inspired on this knowledge, artificial neurons have been defined as a mathematical function that produces an output based on a weighted sum of inputs if the sum is greater than a threshold[13].

Fang et al.[18] presents a paper proposing a localization algorithm, that uses an adaptive neural network, which takes the received signal strength (RSS) from the access points (APs) as inputs to infer the client position in the wireless local area network (LAN) environment.

2.2. Indoor Location System Explanation

In our previous work[8,14], we developed an efficient indoor location system based on the fingerprint approach to estimate the location of the user using the magnetic-field as an information source and a conventional smartphone as a sensor. The methodology to develop an ILS is depicted in figure 1, and a brief description of each phase is given.

2.2.1. Phase 1: Data collection

As in the fingerprint approach, a data collection phase is needed. This phase consists of collecting samples of earth magnetic-field readings, througth the sensor of a smartphone.

To get data entries, the user must walk 10 seconds around the indoor environment with the Smartphone in the palm of his hand with the screen up and keep it up to waist level.

The data collected is used to generate signatures, a basic set of data entries that represent the spectral and temporal data of a room. In order to estimate the number of signatures needed to create a model, we used Eq.(1) proposed by *Eberhardt*[19] to determine the minimal number of experiments in a multivariate process with the aim to have static validation. In Eq. (1), x is the minimum number of experiments, and N is the number of variables.

$$x = log_2(N) + 1 \tag{1}$$

2.2.2. Phase 2: Data analysis

Data analysis consists of five steps described as follows:

- *Collecting Magnetic-Field data*. Given the nature of the magnetic-field phenomena, the magnetic-field sensor obtains a vector of three components Bx, By, and Bz[4]; and in order to have a single dimension, we compute the magnitude of the field as described in Eq. (2), where Mx, My, and Mz are the three physical axes along x, y, and z respectively.

$$|M| = \sqrt{Mx^2 + My^2 + Mz^2} \tag{2}$$

- *Signature Normalization*. To acquire a robust ILS and avoid outliers, a data normalization is applied to all collected data, eliminating spatial scaling and shifting by normalizing the data using Eq. (3), where $z_{i,d}$ is the normalized reading, $r_{i,d}$ refers to the i^{th} observation in the collected data in dimension d; μ_d is the mean value of data for dimension d and σ_d is the standard deviation of the data for dimension d.

$$\forall i \in m : z_{i,d} = \frac{r_{i,d} - \mu_d}{\sigma_d} \tag{3}$$

  Eq. (3) is applied for all dimensions in R^d
- *Feature Extraction*. A process that consists on data signal reduction that works by extracting features that characterize the behavior of the signal. The features that were chosen in the analysis are shown in Table 1. These features are taken from temporal and spectral domains.
- *Feature Vector*. Once all the features are extreacted, all of them are merged into one vector to characterize each second of the collected data.
- *Percentile Rank*. In order to keep all the features in a range of 0 to 1 and make them have the same impact in the model development, a percentile rank is applaied using Eq.(4).

$$PR = \frac{trunc(rank(x))}{length(x)} \tag{4}$$

2.2.3. Phase 3: ILS model development

In order to build the model that allows us to estimate the user's location, a series of tasks are required:

- *Task 1: Feature Selection*. To reduce the number of features needed and improve the accuracy of the ILS, a feature selection is done. To accomplish this task a Genetic Algorithm(GA) is used as an optimizer and feature selector with a specific classifier as the cost function. The final model built is then used with this classifier. This classifier can be changed to obtain different models with different behaviors that are going to be evaluated in this work.
- *Task 2: Forward selection* (FS). During the GA process the features are ranked by importance, and to develop a model with better accuracy a FS strategy is applied. This method generates nested models using the rank of

Table 1. Features Extracted

Features	Temporal Domain	Spectral Domain
Kurtosis	*	*
Mean	*	*
Median	*	*
Standard Deviation	*	*
Variance	*	*
Coefficient of Variation (CV)	*	*
Inverse CV	*	*
1,5,25,50,75,95,99 100-Quantile	*	*
Trimmed Mean	*	*
Shannon Entropy		*
Slope		*
Spectral Flatness		*
Spectral Centroid		*
Skewness		*
1-10 Spectrum Components		*

features, adding the next best ranked feature, one at a time in an iterative process, so that it selects the features that most increase fitness.

- *Task 3: Backward elimination* (BE). The model acquired using the FS could have redundant information, and to eliminate this, a backward elimination (BE) strategy is then applied. This strategy generates models by deleting features one by one from the FS model to improve the fitness of the model; this process is repeated until no improvement is detected. This final model is the one used in the ILS using a specific classifier, which is proposed as a cost function in the GA.

3. Experimental Setup

The data from the magnetic-field that was used for these experiments was collected in a residential home that consists of 4 rooms: living room, dining room, kitchen and bathroom, as it is shown in figure 2. The data was collected with the magnetic sensor of a smartphone device (Samsung S3 i9300 with the official samsung android compilation 4.1), with a mobile application developed with Java using the Google API Level 7. Data used in this evaluation is available in the AAAMI research group website [1] which includes the magnetic-field data.

In the second phase, the R Project [2], a free multiplataform software (GNU project) environment for statistical computing, was used to implement the methodology proposed in this research work. The third phase of the methodology, in which GAs, FS and BE strategies are required, the Galgo R package [13] was used to solve the optimization problems.

The Galgo R package does the work in four basic steps:

1. A preprocess of the data to comply with the requirements of Galgo.
2. A search of multivariate models through a GAs procedure.
3. When the search process is completed, an analysis of the best chromosomes is done.
4. The best model is chosen from the different models generated by Galgo.

[1] http://aaami.mty.itesm.mx/
[2] http://www-r-project.org/

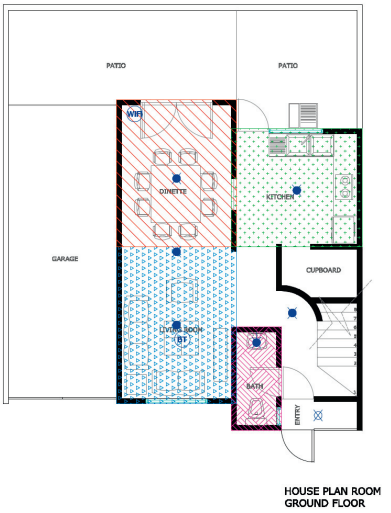

Figure 2. First floor house plans with furniture

4. Results

In this section we present the evaluation of the classifiers during the development of the ILS using each algorithm as the cost function in the GA. Table 2 shows several performance characteristics that were considered in this study, among which sensitivity and specificity are two well-known indicators of the performance of the classifiers. *Sensitivity* is the ability of a test to correctly classify an individual as *diseased* and *specificity* is the ability of a test to correctly classify an individual as *disease-free*[20]. hese terms are important to evaluate the performance of the classifier in this context because we can interpret them as a user belonging to a certain location and not belonging to a different location. We evaluate sensitivity and specificity instead of accuracy because the importante of the false positives in an ILS. The time to develop the final model is also a critical characteristic to be evaluated if we consider a system that recalibrates the model periodically. We know that most of the ILS must be recalibrated when the indoor distribution changes, for example a model for different intervals of the day, etc.

Figure 3 shows that the Nearest Centroid method outperformed dramatically all other classifiers in terms of the required time to learn the final model, which took about 5 minutes as opposed to 82 minutes obtained using Artificial Neuron Networks, its closest competitor. We also consider the number of features needed to develop the final model because it is related with the amount of information needed to develop the ILS.

In order to learn if some processes such as the Fast Fourier Transform can be avoided using a specific classifier, we split the features into two types, spectral and time features. The idea that drove us to analize if some processes could be avoided in order to locate an individual with high accuracy, was mainly to reduce the time to build the feature vectors, and as a consequence to reduce the time to build the models.

On the other hand, sensitivity and specificity vary around one percent (1%) regardless of the classifier used, therefore the performance in terms of these factors was about the same for all the classifiers.

The minimum number of features needed to implement an ILS was 2. All the classifiers required only spectral features with the exception of Random Forest, which required one spectral feature and one temporal feature.

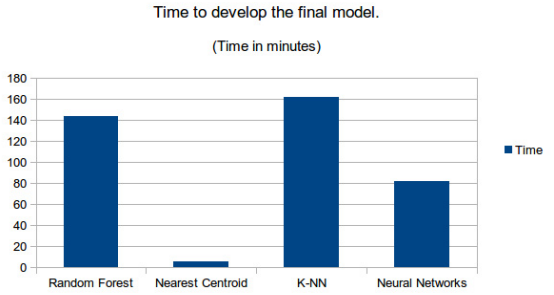

Figure 3. Time in minutes needed to develop the final model.

Table 2. Classifiers Performance

Method	Sensitivity	Specificity	Time (In Minutes)	Spectral Features	Temporal Features
Random Forest	0.9883401	0.9961134	144	1	1
Nearest Centroid	0.9890737	0.9963579	5	2	0
K-NN	0.9889345	0.9963115	162	2	0
Neural Networks	0.9782658	0.9927553	82	2	0

5. Conclusions and Future Work

The main contribution of this paper is an empirical evaluation of four different classifiers that can be used to develop an ILS based on magnetic-field signal. This evaluation allows us to know the behavior of the classifiers in several important aspects, such as, the time required to build the final model, model features, sensitivity and specificity.

The following important aspects regarding the classifiers were identified :

- *Sensitivity and Specificity:* As it is shown in table 2, Sensitivity and Specificity vary around one percent (1%) regardless of the classifier used. Therefore, the algorithms have the same performance in these terms, so we conclude that these features are not determinant to take a decision about what classifier should be used in the development of an ILS.
- *Most of the classifiers work with spectral evolution features:* In our results, we identify that the classifiers require a minimum of two (2) features to estimate the user location. These features are from the spectral evolution of signatures. This fact allowed us to learn that the important information is embedded in the spectra. However, Random Forest was the exception, requiring one (1) feature from the temporal evolution.
- *Time is the most important issue to take into account.* Along the evaluation performed in this work, we also show that time is the variable with higher variability. Time became the most important characteristic to take into account when developing an ILS that requiers recalibration.

Regarding the classifiers, we conclude that Nearest Centroid classifiers have better performance in all the aspects, even in sensitivity and specificity although the difference is small for this experimentation data. Time was the most important characteristic when developing an ILS because the system needs to be recalibrated frequently.

As part of the future work, we propose to increase the number of classifiers given the fact that there are other approaches to develop ILS. Moreover there are other variables that must be taken in consideration such as the information source, for example, Bluetooth and WiFi signals. And finally an overfitting evaluation is requiered to prove independence from the data used. Our future work has a twofold purpose: increase the number of features as well as to extend the sources of information and the number of classifiers.

References

1. S. S. Saad, Z. S. Nakad, A standalone RFID indoor positioning system using passive tags, Industrial Electronics, IEEE Transactions on 58 (5) (2011) 1961–1970.
2. L. Chen, L. Pei, H. Kuusniemi, Y. Chen, T. Kröger, R. Chen, Bayesian fusion for indoor positioning using bluetooth fingerprints, Wireless personal communications 70 (4) (2013) 1735–1745.
3. M. Paciga, H. Lutfiyya, Herecast: an open infrastructure for locationbased services using WiFi., in: WiMob (4), Citeseer, 2005, pp. 21–28.
4. W. Storms, J. Shockley, J. Raquet, Magnetic field navigation in an indoor environment, in: Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2010, IEEE, 2010, pp. 1–10.
5. X. Liu, H. Makino, Y. Maeda, Basic study on indoor location estimation using visible light communication platform, in: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, IEEE, 2008, pp. 2377–2380.
6. C. E. Galván-Tejada, J. C. Carrasco-Jimenez, R. Brena, Location Identification Using a Magnetic-field-based FFT Signature, Procedia Computer Science 19 (2013) 533–539.
7. C. E. Galván-Tejada, J. C. Carrasco-Jiménez, R. F. Brena, Bluetooth-WiFi based Combined Positioning Algorithm, Implementation and Experimental Evaluation, Procedia Technology 7 (2013) 37–45.
8. C. E. Galvan-Tejada, J. P. Garcia-Vazquez, R. Brena, Magnetic-Field Feature Reduction for Indoor Location Estimation Applying Multivariate Models, in: Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on, IEEE, 2013, pp. 128–132.
9. L. Rokach, Data mining with decision trees: theory and applications, Vol. 69, World scientific, 2008.
10. D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Machine learning 6 (1) (1991) 37–66.
11. L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
12. G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, The Journal of Machine Learning Research 9 (2008) 2015–2033.
13. V. Trevino, F. Falciani, GALGO: an R package for multivariate variable selection using genetic algorithms, Bioinformatics 22 (9) (2006) 1154–1156.
14. C. E. Galván-Tejada, J. P. García-Vázquez, R. Brena, Magnetic-Field Feature Extraction for Indoor Location Estimation, in: Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction, Springer, 2013, pp. 9–16.
15. Y. Chen, Q. Yang, J. Yin, X. Chai, Power-efficient access-point selection for indoor location estimation, Knowledge and Data Engineering, IEEE Transactions on 18 (7) (2006) 877–888. doi:10.1109/TKDE.2006.112.
16. L. M. Ni, Y. Liu, Y. C. Lau, A. P. Patil, Landmarc: indoor location sensing using active rfid, Wireless networks 10 (6) (2004) 701–710.
17. A. K. Jain, J. Mao, K. M. Mohiuddin, Artificial neural networks: A tutorial, IEEE computer 29 (3) (1996) 31–44.
18. S.-H. Fang, T.-N. Lin, Indoor Location System Based on Discriminant-Adaptive Neural Network in IEEE 802.11 Environments, Neural Networks, IEEE Transactions on 19 (11) (2008) 1973–1978. doi:10.1109/TNN.2008.2005494.
19. F. Eberhardt, C. Glymour, R. Scheines, N-1 experiments suffice to determine the causal relations among n variables, in: Innovations in machine learning, Springer, 2006, pp. 97–112.
20. R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, R. Thomas, Understanding and using sensitivity, specificity and predictive values, Indian journal of ophthalmology 56 (1) (2008) 45.