Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 191 (2021) 361–366

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2th International Workshop on Artificial Intelligence & Internet of Things (A2IOT)
August 9-12, 2020, Leuven, Belgium

Predictive model for the identification of activities of daily living (ADL) in indoor environments using classification techniques basedon Machine Learning

García-Restrepo, Johanna[a] , Ariza-Colpas, Paola Patricia[a*,b],Oñate-Bowen, Alvaro Agustín[d], Suarez-Brieva ,Eydy del Carmen[d], Urina-Triana, Miguel[c*], De-la-Hoz-Franco, Emiro[a], Díaz-Martínez, Jorge Luis[a] and Butt, Shariq Aziz[e] and Molina_Estren, Diego[a]

[a] *Universidad de la Costa, CUC,Street. 58 # 55 - 66 BarranquillaColombia*
[b] *Universidad Pontificia Bolivariana, Medellin Colombia. Street. 1 #70-01, Medellín, Colombia*
[c] *Universidad Simón Bolívar. Street. 58 # 55-132.Barranquilla – Colombia*
[d] *Universidad Popular del Cesar. Street 15 # 12-54. Valledupar – Colombia*
[d] *University of Lahore, 1-Km Defence Road. Lahore- Pakistan*

Abstract

AI-based techniques have included countless applications within the engineering field. These range from the automation of important procedures in Industry and companies, to the field of Process Control. Smart Home (SH) technology is designed to help house residents improve their daily activities and therefore enrich the quality of life while preserving their privacy. An SH system is usually equipped with a collection of software interrelated with hardware components to monitor the living space by capturing the behavior of the resident and their occupations. By doing so, the system can report risks, situations, and act on behalf of the resident to their satisfaction. This research article shows the experimentation carried out with the human activity recognition dataset, CASAS Kyoto, through preprocessing and cleaning processes of the data, showing the Vía Regression classifier as an excellent option to process this type of data with an accuracy 99.7% effective

Keywords: HAR, Human Activity Recognition, Machine Learning, ADL, Activity Daily Living

* Corresponding author. Tel.: +57 035 3225498; fax: +355 22 66 999
E-mail address: pariza1@cuc.edu.co

10.1016/j.procs.2021.07.069

1. Introduction

Older adults usually have problems in relation to their health and the management of their quality of life. There are many studies that have made it possible to strengthen the social well-being of the elderly using technology. One of the technological aspects that contribute to improving the quality of life of adults, is precisely the enrichment of physical spaces with sensors, video surveillance equipment and actuators, which favor the performance of their activities of daily living (ADL) in indoor environments, which allows discovering patterns of human actions generated from the movement and interaction of individuals with the environment, in such a way that they facilitate data monitoring and understanding of the activity of older adults in surveillance environments, technology-based, with the purpose of automatically detecting abnormal patterns, which affect your health or could put your life at risk. Generating in this way, conditions that allow the elderly to have a comfortable, comfortable, and independent life. These types of technological tools allow monitoring of daily actions that range from basic hygiene management, personal care, cleaning and even meal preparation. All these basic activities give older adults the possibility of interacting in community [1] with the tranquility of a personalized and functional medical attention through the implementation of technology. Although the list of activities that a person can perform is extensive, this study focused on those that take place in indoor environments. Where the domain of cognitive ability prevails to carry out these actions. To identify this type of activity, various studies have been developed from different lines of work: from the stochastic perspective, making use of the hidden Márkov chains [2], from the treatment of uncertainty through of fuzzy logic [3], and from web ontology [4] from the concepts on which web developments are structured. Through this article, the results of the research and the predictive model and its characteristics are shown. First, the brief review of the literature section 2 is shown, second, the functional characteristics of the model are shown. Third, the results of the experiments are detailed, finally the conclusions are shown.

2. Literature Review

As a result of the analysis of the literature review, a series of categories related to: the implementation of preprocessing techniques in HAR [6-11]; Techniques based on supervised and unsupervised learning, [12-15] and the different areas of HAR development that allow the processing of the data extracted from the sensors to solve problems related to human activities in different contexts [16-18]. With the rapid development of current technology and the popularization of smartphones, ubiquitous sensing has become a research field and its universal purpose is to extract knowledge from data obtained by sensors. One of the keys to a successful HAR is selecting the appropriate features, representing the sensor data, and implementing classifiers. Ronao [6] performs a study based on the selection of characteristics. He proposes a convolutional neural network (Convolutional Neural Networks - Convnet) that allows the extraction of characteristics and their classification using smartphone sensors. With a multilayer convnet with alternating convolution and grouping membranes, the entities extracted from the raw time series sensor data, the lower layers that extract the most basic features, and the upper layers that can obtain more complex characteristics. Showing how different Convnet architectures affect overall performance and how this system does not require advanced preprocessing or tedious manual build functions and can outperform other next generation algorithms in the HAR field. The quality of the data according to Gudivada [8] plays a fundamental role in preprocessing, and data intensive applications. Acquisition and verification are the biggest challenges in data- intensive applications. The high quality of the information enables faster decisions to be made. The degree of adequacy of the data for a specific purpose must be considered, that it is complete, consistent, without duplication, that it is accurate and timely. The application of relevant practices and controls that improve the quality of reports is known as data quality. Furthermore, quality assessment is domain specific, less objective, and requires significant human participation.

The analysis presented by Galván [10] proposes the collection of data from different sources using a device and the use of a modified version of the Adaptive Boosting (AdaBoost) algorithm for the selection of characteristics. A similar work was proposed by Sean Eddy [11] in the use of hidden Markov models (HMM) to classify certain activities. A relevant example of this technique is the one proposed by Sean Eddy, who built a model called the Discriminative Conditional Restricted Boltzmann Machine (DCRBM). This model combines a discriminatory approach with the capabilities of the Conditional Restricted Boltzmann Machine (CRBM). The model enables the discovery of Essential Predicates of Social Interaction (ESIP) actionable components to train the DCRBM model and use it to generate low-level ESIP data with a high degree of precision.

3. Predictive Functional Model

The construction process of the predictive model for the recognition of ADL from the CASAS Kyoto dataset, implied the approach of an experimental process, divided into a series of phases (see figure No. 1), that is: 1) integration and debugging,

2) grouping of instances, 3) application of techniques of representation of characteristics by subset of data, 4) training and testing of models for classification, and (5) evaluation of the quality metrics of the models to identify with which hybridization of technique, the best results were generated in terms of hit rate, see figure 1, Each of the aforementioned phases are detailed below.

To generate the ADL Activities dataset, the CASAS laboratory researchers recruited 20 participating volunteers to carry out five (5) activities which are: making phone calls, washing hands, cooking, eating, and cleaning. The information was collected through the following sensors: Motion sensors, Sensors for the use of kitchen elements, Sensor in the medicine container, sensor in kitchen utensils, telephone directory, cabinet sensor, water sensor, sensor ignition of the kitchen and use of the telephone. For the integration of the data, a dataset was first generated consisting of 120 files that correspond to 86-day measurements of the interactions of individuals vs sensors in the environment, these instances were ordered consecutively by the TimeStamp column, this integration was carried out due to because each of the files have the same structure. The structure of each of the files is made up of date, TimeStamp, sensor, status, start and end of the activity. As a result, a first-row dataset was obtained with a total of 6,425 data instances. For the construction of the preprocessed dataset, the row dataset was taken in which, considering the TimeStamp, a representation in columns of the states of each of the sensors of the respective timelinewas generated, indicating in the same way, the activity that was running on that line.

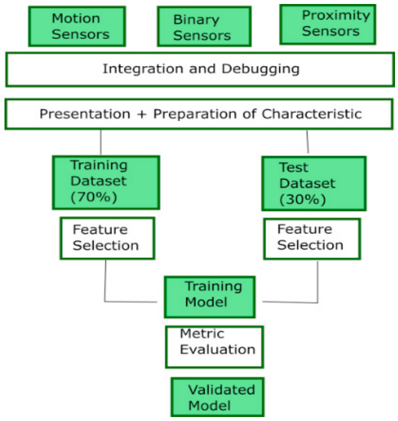

Fig. 1. Predictive Functional Model. Source: Created by author

The preprocessed dataset is made up of five (5) types of characteristics or data columns (which are represented in table No. 1) and collects a total of 26 characteristics.

Table 1. Description of the characteristics of the Preprocessed dataset.

#	Characteristics	Amount	Description
1	TimeStamp	1	Contains the date and time of the sequence of activities.
2	Motion sensors	11	Identified like this: M01, from M07 to M09, from M13 to M18 and M23. Each one takes ON and OFF values.
3	Binary sensors	10	Identified like this: from I01 to I08 and D01. Each one takes ABSENT and PRESENT values
4	Proximity sensors	3	Identified like this: AD1-A, AD2- B and AD1-C. Each one takes numerical values.
5	Telephone use	1	It takes values of START and END.
	Total	26	

After the process described above, in which the preprocessed dataset is obtained, the grouping of the data instances continues; For this, the dataset was segmented with sliding windows of three (3) seconds, while, for the process of v generating new characteristics, the values of the attributes corresponding to the proximity sensors (AD1-A, AD2- B and AD1-C), applying aggregation functions such as: standard deviation, kurtosis, mean, maximum, minimum, bias and range.

As a result of this, 21 new characteristics were obtained, generating a segmented dataset with 45 characteristics including activity and 5,736 instances. See table No 2.

Table 2 . Description of the characteristics of the segmented dataset

#	Characteristics	Amount	Description
1	TimeStamp	1	Contains the date and time of the sequence of activities.
2	Motion sensors	11	Identified like this: M01, from M07 to M09, from M13 to M18 and M23. Each one takes ON and OFF values.
3	Binary sensors	10	Identified like this: from I01 to I08 and D01. Each one takes ABSENT and PRESENT values
4	Proximity sensors	21	Identified like this: AD1-A, AD2- B and AD1-C. Each one takes numerical values.
5	Telephone use	1	It takes values of START and END.
	Total	**26**	

The data set after the segmentation process is unbalanced, that is, the number of instances of each of the activities are in different proportions. One solution to this is class balancing. To illustrate how this technique works, minority classes must be identified and then oversampled, that is, taking a sample from the data set and considering its closest neighbors, to create a more synthetic data point. For the balancing of the classes, the SMOTE technique (Synthetic Minority Oversampling Technique) [19] was used, obtaining as a result a balanced dataset as shown in Figure 2.

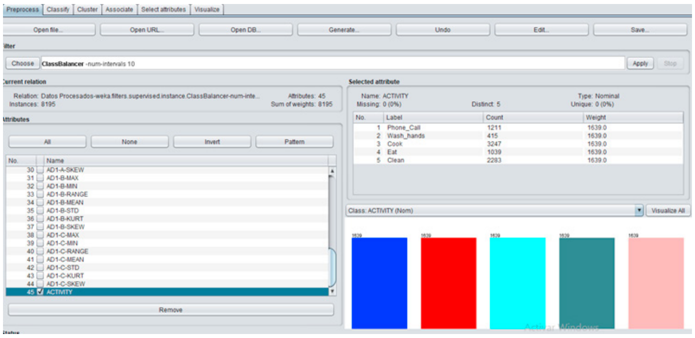

Fig 2. Balanced Dataset Classes. Source: Created by author.

When balancing the classes as evidenced in Figure No. 2, the symmetry in each of the instances can be observed. Next, two subsets of data were generated, the first for training the model and the second for testing, with a 70-30 ratio, given that the literature [20] indicates 70% of the data for training and 30 % for test, with different records in each dataset.For the process of applying characteristics selection techniques, the level of incidence that the attributes of the dataset have with respect to the process of identifying the class criteria is identified. This process was carried out with the application of different characteristics selection techniques such as: Gain Ratio, Chi square, Info Gain, OneR, ReliefF and Symmetrical Uncert. This process seeks a reduction in the number of characteristics to reduce the computational time required in the construction of the final model.

Subsequently, the training of the model is carried out using the attributes with the best results according to the selection of characteristics. With the purpose that it can predict the activity carried out by a person according to the entered data set, in this stage several models are obtained, which later will be applied different quality metrics to find the one with the best result throw. Finally, the evaluation of the quality metrics of the obtained model is carried out to identify which hybridization of the classification technique with the characteristic selection technique, yields the best results in terms of the hit rate and then the selection of the model is carried out with the best results. The quality techniques that were evaluated are the following: FPR, Precision, Recall and ROC Area. Understanding the methodology used in this study, we proceeded to recreate a series of experimentation scenarios, using different types of segmentation techniques, selection of characteristics and classification.

4. Experimentations

The processed dataset (consisting of 45 characteristics, including the class criterion) was evaluated using the following nineteen (19) classification techniques: Classification Via Regression [21], Random SubSpace [22], Bagging [23], Random Forest [24], Attribute Selected, J48 [25], OneR [26], LMT [27], REP Tree [28], Randomizable Filtered [29], Random Tree [30], JRip [31], Iterative Classifier Optimizer, LogiBoost [32], Multi Class [33], Multilayer Perceptron [34], Simple Logistic [35] and IBK [36]. See table 3.

Table 3 . Classificator´s Results

#	Classificator	FPR	Accuracy	Recall	ROC Area
1	Classification Via Regression	1,20%	97,60%	97,60%	99,70%
2	Random SubSpace	4,10%	89,80%	89,90%	98,60%
3	Bagging	5,20%	85,10%	86,40%	97,40%
4	RandomForest	4,10%	86,90%	87,40%	97,20%
5	Attribute Selected	2,70%	91,60%	91,80%	96,90%

5. Conclusions

To determine which of these classification techniques generated better results, the different quality metrics were evaluated. With the Classification Via Regression classifier, the ROC area with the highest rate was 99.70%, the IBK classifier with the lowest ROC area rate was 71.60%, with the Random SubSpace the accuracy rate was 89.80%, FPR was 4.10% and recall with 89.90%. After carrying out this experimentation, it can be identified that the classification methods constitute a good tool to be able to carry out the adequate identification of activities of daily life and that they allow to validate predictive methods and support the making of these types of decisions.

References

[1][Nagi, S. Z., Burk, R. D., & Potter, H. R. (1965). Back disorders and rehabilitation achievement. Journal of Chronic Diseases, 18(2), 181–197. https://doi.org/10.1016/0021-9681(65)90101-3

[2] Lladó, M. R., Código, H., Lennin, A., Quiroz, P., & Lima -Perú, V. (2020). ENTORNO DOMÓTICO ADAPTADO A PERSONAS CON DISCAPACIDAD FÍSICA UTILIZANDO MODELOS OCULTOS DE MARKOV Tesis para optar el Título Profesional de Ingeniero de Sistemas. In Repositorio Institucional - Ulima. Universidad de Lima. http://repositorio.ulima.edu.pe/handle/20.500.12724/11664

[3] Carlos, A., D'negri, E., De Vito, E. L., & Zadeh, L. A. (2006). Introducción al razonamiento aproximado: lógica difusa. In Revista Argentina de Medicina Respiratoria Año (Vol. 6).

[4] Marcondes, C. H., & Almeida Campos, M. L. de. (2008). ONTOLOGIA E WEB SEMÂNTICA: O ESPAÇO DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO. PontodeAcesso, 2(1), 107. https://doi.org/10.9771/1981-6766rpa.v2i1.2669

[5] DANE. Archivo Nacional de Datos ANDA. 2014. [Citado Marzo 20,2016]. Available in: http://formularios.dane.gov.co/Anda_4_1/index.php/home

[6] Ronao, C. A., & Cho, S. B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. Expert Systems with Applications, 59, 235–244. https://doi.org/10.1016/j.eswa.2016.04.032

[7] Capela, N. A., Lemaire, E. D., & Baddour, N. (2015). Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. PLoS ONE, 10(4), e0124414. https://doi.org/10.1371/journal.pone.0124414

[8] Gudivada, V. N., Ding, J., & Apon, A. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Flow Cytometry of 3-D structure View project Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transf. October, 1–20. https://www.researchgate.net/publication/318432363

[9] Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. Proceedings of the IEEE International Conference on Computer Vision, 1, 10– 17. https://doi.org/10.1109/iccv.2003.1238308

[10] Galván-Tejada, C. E., Galván-Tejada, J. I., Celaya-Padilla, J. M., DelgadoContreras, J. R., Magallanes-Quintanar, R., Martinez-Fierro, M. L., GarzaVeloz, I., López-Hernández, Y., & Gamboa-Rosales, H. (2016). An Analysis of Audio Features to Develop a Human Activity Recognition Model Using Genetic Algorithms, Random Forests, and Neural Networks. Mobile Information Systems, 2016, 1–10. https://doi.org/10.1155/2016/1784101

[11] Eddy, S. R. (1998). Profile hidden Markov models. Academic.Oup.Com, 144(9), 755–763. https://academic.oup.com/bioinformatics/articleabstract/14/9/755/259550 Envejecimiento y salud. (2018, February 5).

[12] Shah,C. (2020). Supervised Learning. In A Hands-On Introduction to Data Science (pp.235–289).

[13] Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review, 33(4), 275–306.https://doi.org/10.1007/s10462-010-9156-z

[14] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. ACM International Conference Proceeding Series, 148, 161–168. https://doi.org/10.1145/1143844.1143865

[15] Mejia-Ricart, L. F., Helling, P., & Olmsted, A. (2018). Evaluate action primitives for human activity recognition using unsupervised learning approach. 2017 12th International Conference for Internet Technology and Secured Transactions, ICITST 2017, 186–188. https://doi.org/10.23919/ICITST.2017.8356374

[16] Crandall, A. S. (2011). BEHAVIOMETRICS FOR MULTIPLE RESIDENTS IN A SMART ENVIRONMENT. https://scihub.si/http://research.wsulibs.wsu.edu/xmlui/handle/2376/2855

[17] Hoey, J., Pltz, T., Jackson, D., Monk, A., Pham, C., & Olivier, P. (2011). Rapid specification and automated generation of prompting systems to assist people with dementia. Pervasive and Mobile Computing, 7(3), 299–318. https://doi.org/10.1016/j.pmcj.2010.11.007

[18] Fahad, L. G., Tahir, S. F., & Rajarajan, M. (2015). Feature selection and data balancing for activity recognition in smart homes. IEEE International Conference on Communications, 2015-Septe, 512–517. https://doi.org/10.1109/ICC.2015.724837310

[19] Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In Journal of Artificial Intelligence Research (Vol. 16). https://scihub.si/http://www.jair.org/index.php/jair/article/view/10302

[20] López Saca, F., Ferreyra Ramírez, A., Avilés Cruz, C., Villegas Cortez, J., Zúñiga López, A., & Rodrigez Martinez, E. (2018). Preprocesamiento de bases de datos de imágenes para mejorar el rendimiento de redes neuronales convolucionales. Research in Computing Science, 147(7), 35–45. https://doi.org/10.13053/rcs-147-7-3

[21] Ruan, Y. X., Lin, H. T., & Tsai, M. F. (2014). Improving ranking performance with cost-sensitive ordinal classification via regression.

Information Retrieval, 17(1), 1–20. https://doi.org/10.1007/s10791-013-9219-2

[22] Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. https://doi.org/10.1109/34.709601

[23] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/bf00058655

[24] Nagalla, R., Pothuganti, P., & Pawar, D. S. (2017). Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests. Procedia Computer Science, 109, 474–481. https://doi.org/10.1016/j.procs.2017.05.312

[25] Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, 16(3), 235–240. https://doi.org/10.1007/bf00993309

[26] Kumar, K., Kumar, G., & Kumar, Y. (2013). Feature Selection Approach for Intrusion Detection System. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 2(5), 47–53.

[27] Ariza Colpas, P., Vicario, E., De-La-Hoz-Franco, E., Pineres-Melo, M., Oviedo-Carrascal, A., & Patara, F. (2020). Unsupervised human activity recognition using the clustering approach: A review. Sensors, 20(9), 2702.

[28] Chandra, S., & Maheshkar, S. (2017). Verification of static signature pattern based on random subspace, REP tree and bagging. Multimedia Tools and Applications, 76(18), 19139–19171. https://doi.org/10.1007/s11042-017-4531-2

[29] Asaju, L. B., Shola, P. B., Franklin, N., & Abiola, H. M. (2017). Intrusion Detection System on a Computer Network Using an Ensemble of Randomizable Filtered Classifier, K-Nearest …. Ftstjournal.Com, 2(1), 550– 553. www.ftstjournal.com

[30] Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News.

International Journal of Innovative Science, Engineering & Technology, 2(2), 438–446. www.ijiset.com

[31] Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S. P., & Raghuvanshi, M. (2011). J48 and JRIP rules for e-governance data. International Journal of Computer Science and Security, 5(2), 201–207.11

[32] Cai, Y. D., Feng, K. Y., Lu, W. C., & Chou, K. C. (2006). Using LogitBoost classifier to predict protein structural classes. Journal of Theoretical Biology, 238(1), 172–176. https://doi.org/10.1016/j.jtbi.2005.05.034

[33] Qian, H., Mao, Y., Xiang, W., & Wang, Z. (2010). Recognition of human activities using SVM multi-class classifier. Pattern Recognition Letters, 31(2), 100–111. https://doi.org/10.1016/j.patrec.2009.09.019

[34] Suykens, J. A. K., & Vandewalle, J. (1999). Training multilayer perceptron classifiers based on a modified support vector method. IEEE Transactions on Neural Networks, 10(4), 907–911. https://doi.org/10.1109/72.774254

[35] Khalajzadeh, H., Mansouri, M., & Teshnehlab, M. (2014). Face recognition using convolutional neural network and simple logistic classifier.

Advances in Intelligent Systems and Computing, 223, 197–207. https://doi.org/10.1007/978-3-319-00930-8_18

[36] Choudhury, S., & Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings, 89–95. https://doi.org/10.1109/ICSTM.2015.7225395