Comparison of Machine Learning Techniques in a Smart Home-Oriented System

Alejandro H. García R., Salvador Ibarra M., José A. Castán R., Jesús D. Terán V., Julio Laria M., Mayra G. Treviño B., Julissa Pérez C. and Emilio Castán R.

***Abstract*—The implementation of automatisms and intelligent systems in our environment ensures the achievement of many tasks. However, experts must address several challenges to increase the integration of technology more efficiently. A system that facilitates the performance of activities to users can be reactive. It means, reacted at the moment based on changes in the environment or proactive by predicting user preferences to provide appropriate solutions based on prior learning. Predicting preferences is one of the most considerable interests to researchers because it allows developing intelligent systems to focus more effectively on users. Our approach consists of an efficient inference system for decision-making based on users' preferences. In this sense, this paper presents the application of five of the machine learning techniques most commonly used by researchers in the field of classification and decision-making. We prove the techniques' efficiency with a dataset from a smart home-oriented system based on user preferences. The experimentation consists of acquiring each technique's efficiency for its later comparison employing a Friedman Test. The results show that C4.5 and ANN techniques are suitable for the development of the inference system.**

***Index Terms*— Human-Centered Computing; Intelligent Home; User Preferences; Internet of Things; Machine Learning**

I. Introduction

According to automation in a home, a home can be classified into three types: automated, digital, and intelligent. The automated home consists of a building integrated with electronic devices such as sensors, actuators, and software for specific tasks [1]. For its part, the digital home incorporates the Internet's use into the home as the primary communication channel with the devices. This feature allows controlling the digital home from the outside. However, both automated and digital homes require significant interaction with users to perform diverse tasks. Therefore, these tend to be limited in their applications.

To solve that situation, researchers have carried out in the homes the integration of intelligent systems to serve as support for decision-making [1], [2]. In this way, it is possible to reduce the number of user interactions with the building automatisms. With this, the concept of a smart home emerges. Smart homes are an opportunity to face the complicated lifestyle in which we are involved. An example of this is to leave a lamp on, a situation that, in addition to affecting energy consumption (i.e., our economy) and comfort, could lead to a short circuit that causes material damage and even threatens our lives.

Machine learning in household systems allows the development of adequate inference systems to grant intelligent households enough autonomy for proactive and efficient decision-making without direct user supervision [3]. In this way, smart homes can achieve a high level of context-awareness. In this sense, this paper presents the results obtained from applying some machine learning techniques on a set of data obtained from a smart home-oriented system based on user preferences.

The paper's remainder is as follows: Section II addresses some fundamental aspects of the machine learning techniques used. Section III shows the methodology used for the application of the techniques. Section IV presents the results of the comparison carried out. Section V describes the discussion, and Section VI mentions some conclusions and future research work.

II. Machine Learning Techniques

This section describes the machine learning techniques selected from state-of-the-art for this paper's purposes. These techniques were selected based on their large application field. The selected techniques are listed below:

- K-nearest neighbors (**KNN**)
- Induction Decision Tree (**ID3**)
- **C4.5**
- Naive Bayes (**NB**)
- Artificial Neuronal Network (**ANN**)

2.1 K-nearest neighbors

The **KNN** technique has been used widely in classification problems. Based on a distance metric, the **KNN** measures the difference or similarity between two cases. Given a case *x* of an unknown class, we calculate the distance between *x* and all the training database cases. Finally, the class determined by the ***K*** cases closest to *x* is assigned [4]. **KNN** is a typical example of slow learning that stores training data at the time of training and delays its learning until classification. Despite this, **KNN** has been widely used as a classifier for decades [5].

Manuscript received March 06, 2020; revised December 09, 2020. This work was supported partially by the National Council of Science and Technology (CONACYT) of Mexico.

A. H. Garcia is a PhD Student at the Autonomous University of Tamaulipas, Mexico (a2113330227@alumnos.uat.edu.mx).

S. Ibarra, J. A. Castán, J. Terán, J. Laria, M. G. Treviño and J. Perez are members of the Computer Technology Group at the Engineering School at the Autonomous University of Tamaulipas, Mexico.

E. Castán is professor at the Electric and Electronic Department at the TecNM/Madero Institute of Technology.

The **KNN** algorithm consists of two phases: training and classification. In the training phase, the training examples are vectors (each with a class label) in the space of multidimensional characteristics. In this phase, are stored the characteristic vectors and class labels of the training samples. In the classification phase, the user defines a value for the constant ***K*** and then is classified an unlabeled vector assigning a label based on the most recurrent class among the ***K*** training samples closest to the vector. This way of classifying the input vector based on its distance to training samples is a simple but effective way to classify new points.

2.2 Induction Decision Tree

An Inductive Decision Tree (**ID3**) is an algorithm encompassed within the so-called inductive and supervised learning proposed by Quinlan in 1986. This algorithm aims to model the data through a decision tree. In this tree, the intermediate nodes are attributes of the cases presented, the branches represent values of those attributes, and the final nodes are the class's values. Its main application is decision problems. Its use focuses on the so-called classification problems: diagnosis of diseases given the symptoms, granting of loans, among others.

As an advantage, it has good results in a wide range of applications, and the accuracy of the result is usually high. As a disadvantage, the attributes and classes must be discrete, and sometimes the trees are too leafy, which makes interpretation difficult [6].

2.3 C4.5

Due to the need to process a large amount of data on many occasions, it is necessary to use fast learning algorithms. Many inductive learning algorithms build decision trees. **C4.5** is one of the most used and efficient decision tree learning algorithms [7].

It allows us to extract classification rules from the observations of a system. The entry consists of a set of records. Each record contains some attributes and a decision attribute. An expert in the domain must decide which variable depends on others, so the decision attribute should be considered. Although **C4.5** has been used traditionally as a classifier, it can find temporal relationships [8].

C4.5 creates a decision tree by calculating each attribute's information content and pruning the attributes to create more straightforward classification rules than the original input records. The output is decision rules that can later classify future records.

2.4 Naïve Bayes

A Bayesian classifier is considered a particular case of a Bayesian network, where one variable fulfills the class role, and the other variables are considered attributes. The classification process consists of identifying in which class c_i of a set of classes $C=\{ c_1, c_2, c_3, \ldots, c_m \}$ a new object is found $o=\{ a_1, a_2, a_3, \ldots, a_n \}$ characterized by of individual observations of a set of characteristics (or attributes) $A=\{ A_1, A_2, A_3, \ldots, A_n \}$ [9].

The most straightforward probabilistic approach is the Naive Bayes (**NB**) classification proven successful in many applications [10]. The computational complexity of it is considered very low compared to other methods such as decision trees. Since the classifier combines simple functions of univariate densities, this procedure's complexity is *O(nm)*. This classifier has several advantages, such as those listed below [11]:

- Simple and easy to understand.
- Easy to adapt for incremental learning environment models.
- Resistant to irrelevant attributes.

2.5 Artificial Neuronal Network

An Artificial Neural Network (**ANN**) consists of a mathematical model inspired by the biological systems adapted and simulated in conventional computers [12]. Biological neural networks are active neurons specialized in tasks like as: mathematical calculations, positioning, and memory [13]. Although **ANN**s have a lower degree of complexity than a biological neural network, they are suitable for performing complex calculations and processing information [14].

2.5.1 Multilayer Perceptron

This network model emerged in the 80s of the twentieth century as a solution to overcome the problem detected in the simple perceptron, that is, the inability to learn classes of nonlinearly separable functions [14]. This model is the most used neural network model for resolving classification and regression problems, having demonstrated its status as a universal approximator of functions, which justifies its individual and detailed study. It is a neuronal model with "forward programming", which is characterized by its organization in layers of disjoint cells, so that no neuronal output constitutes an input for neurons of the same or previous layers, thus avoiding connections "towards back" or "auto recurring".

III. Methodology

We use a smart home-oriented dataset based on user preferences for the application of machine learning techniques. This dataset consists of the decision desired by diverse users to solve different contexts originated within the home. The dataset was divided into subsets through the K-Means clustering algorithm with a value of ***K=4*** generated. Based on these groups, machine learning techniques were applied.

For calculating efficiency, we consider splitting each subset into training data and test data. This segmentation consisted of 80%, 20%, respectively. Subsequently, the Friedman nonparametric statistical test was applied to compare the techniques with the efficiency calculations obtained for each subset of data.

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Regarding the **KNN** and **ANN** techniques, we considered some configuration parameters to obtain the best efficiency results. For **KNN**, the similarity metric selected was the Euclidean distance. This metric measures the similarity between the new case to be classified and the rest of the training data set cases. Besides, using a value of ***K=40,*** we observed that this technique provided better efficiency values for test data subsets. For the application of **ANN**, is used the multilayer perceptron model. This model consisted of an input layer, two hidden layers, and a single output neuron. Both the input layer and the hidden layers consisted of a total of **5** neurons each.

The efficiency results calculated by learning technique for each of the subsets are shown in Table I.

TABLE I
CALCULATED EFFICIENCY RESULTS FOR THE SUBSETS

	KNN	ID3	C4.5	NB	ANN
Subset 1	80	93	93	85	93
Subset 2	87	91	97	81	95
Subset 3	83	89	95	86	91
Subset 4	85	85	93	83	91

IV. RESULTS

We use the Friedman test to determine if the efficiency values of machine learning techniques evaluated in this work show differences among the different subsets generated [15]. Equation 1 shows the applied statistical test hypotheses, where ***Ho*** is the null hypothesis, and ***Ha*** is the alternative hypothesis.

$$Ho: \textit{There is no difference among the efficiency calculated for the diverse learning techiniques}$$
$$Ha: \textit{There is a difference among the efficiency calculated for the diverse learning techiniques} \quad (1)$$

Table II shows the parameters of the Friedman test applied.

TABLE II
APPLIED FRIEDMAN TEST PARAMETERS

Learning Techniques (k):	5
Observations (n):	4
Significance Level (α):	0.05
Friedman's Critical Value:	9.48
Confidence level:	95%

We ranked the machine learning techniques according to each one's average efficiency with the different subsets. The best technique was that with the highest efficiency average, thus obtaining first place and so on. In this test, the **C4.5** technique was ranked the best by the Friedman test (Table III).

TABLE III
RANKED LEARNING TECHNIQUES ACCORDING TO THEIR EFFICIENCY

Learning Techniques	Classification
C4.5	1.25
RNA	2.0
ID3	2.875
KNN	4.375
NB	4.5

According to Table II's values, we calculate the Friedman test. In this test, if the statistic is greater than Friedman's critical value or if the p-value is less than α, then the null hypothesis is rejected and accepted the alternative hypothesis [15]. Table IV shows the results of the Friedman test calculated.

TABLE IV. FRIEDMAN TEST RESULTS

Friedman statistic:	13.5
p-value:	0.010566

The value of the Friedman statistic calculated is greater than the critical value. Moreover, the calculated p-value is less than ***α***; this indicates a difference in the efficiency of the machine learning techniques. However, this result does not provide enough information to determine which technique is best for our work. In this sense, we considering the application of a post hoc procedure based on the adjustment of p-values (APV) [15] and multiple comparisons (Friedman test (1xN)). The **C4.5** technique ranked best (Table III) was selected as a control technique and compared against the remaining techniques. We calculated the adjusted p-values with the Holm procedure. Table V shows the results obtained.

TABLE V
UNADJUSTED P-VALUES AND HOLM P-VALUES

Learning Techniques	Unadjusted p-values	Holm p-values
NB	**0.003650**	**0.0125**
KNN	**0.005188**	**0.0166**
ID3	**0.146100**	**0.025**
ANN	0.502334	0.05

The highlighted values indicate those measures in which the p-value obtained by Holm is less than the value of ***α***, that is, those where statistically there is a significant difference between these measures and the **C4.5** learning technique.

V. DISCUSSION

Friedman test results (Table 4) show that the Friedman statistic calculated value is greater than the critical value. Moreover, the p-value obtained is less than ***α***, so there is enough statistical evidence to confirm a difference in the set of machine learning techniques' efficiency.

Additionally, in the post hoc applied by Holm procedure, the highlighted adjusted p-values (Table V) reject the null hypothesis that there are no differences with the **C4.5** technique. In this sense, there is not enough statistical

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

evidence to ensure a difference among the efficiency calculated for **C4.5** and **ANN** techniques.

VI. Conclusions

The creation of intelligent systems focused on users is a growing trend in researchers. With the learning of user preferences, the intelligent system can react proactively. This way, the system provides users' desired responses to the different contexts in their daily work. With this, the users' quality of life benefits from satisfaction, security, and energy savings. The above allows users to consider a smart home more as an extension of themselves and not just as a tool. Such argumentation is one of the most pursuit objectives of modern computing.

Due to the above, this paper addressed this area of opportunity through the application of some machine learning techniques on a set of data obtained from a smart home-oriented system based on user preferences. According to the efficiency results obtained, we concluded that integrating **C4.5** and **ANN** learning techniques in the inference of a smart home-oriented system is adequate to support decision-making. In this sense, the intelligent system can provide the users with the desired responses to the home's different contexts, considering the user preferences. The future work consists of developing and validating a new intelligent system using some of the techniques selected in this research work.

Acknowledgment

A. H. Garcia thanks the National Council of Science and Technology (CONACYT) of Mexico.

References

[1] N. Attoue, I. Shahrour, R. Younes, N. Attoue, I. Shahrour, and R. Younes, "Smart Building: Use of the Artificial Neural Network Approach for Indoor Temperature Forecasting," *Energies*, vol. 11, no. 2, p. 395, Feb. 2018.

[2] O. Hernandez Uribe, J. P. San Martin, M. C. Garcia-Alegre, M. Santos, and D. Guinea, "Smart Building: Decision Making Architecture for Thermal Energy Management.," *SENSORS*, vol. 15, no. 11, pp. 27543–27568.

[3] N. Nesa and I. Banerjee, "IoT-Based Sensor Data Fusion for Occupancy Sensing Using Dempster–Shafer Evidence Theory for Smart Buildings," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1563–1570, Oct. 2017.

[4] H. Rajaguru and S. K. Prabhakar, *KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis*. Anchor Academic Publishing, 2017.

[5] L. Wang, L. Jiao, G. Shi, X. Lu, and J. Liu, *Fuzzy Systems and Knowledge Discovery: Third International Conference, FSKD 2006, Xi'an, China, September 24-28, 2006, Proceedings*. Springer Berlin Heidelberg, 2006.

[6] J. A. Piedra Fernández, *Aplicación de los sistemas neurodifusos a la interpretación automática de imágenes de satélite*. Editorial Universidad de Almería, 2008.

[7] X. Wu *et al.*, *Research and Development in Knowledge Discovery and Data Mining: Second Pacific-Asia Conference, PAKDD'98, Melbourne, Australia, April 15-17, 1998, Proceedings*. Springer, 1998.

[8] K. S. Leung, L. Chan, A. Learning, H. Meng, and I. C. I. D. Engineering, *Intelligent Data Engineering and Automated Learning - IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents: Second International Conference Shatin, N.T., Hong Kong, China, December 13-15, 2000. Proceedings*. Springer, 2000.

[9] R. A. Sánchez, C. R. Osorio, and H. P. Molina, *VIII Congreso Ibérico de Agroingenier{\'\i}a: "Retos de la nueva agricultura mediterránea."* Universidad Miguel Hernández, 2016.

[10] A. Gelbukh and C. A. Reyes-Garcia, *MICAI 2006: Advances in Artificial Intelligence: 5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, November 13-17, 2006, Proceedings*. Springer Berlin Heidelberg, 2006.

[11] O. Z. Maimon and L. Rokach, *Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications*. World Scientific, 2005.

[12] J. R. H. González and V. J. M. Hernando, *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. RA-MA, 1994.

[13] S. V Verdú and C. S. Blanes, *Aplicación de un modelo de red neuronal no supervisado a la clasificación de consumidores eléctricos*. Editorial Club Universitario, 2013.

[14] R. F. López and J. M. F. Fernández, *Las Redes Neuronales Artificiales*. Netbiblo, 2008.

[15] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 3–18, Mar. 2011.