UNIVERSIDAD DE SEVILLA

DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMÁTICOS

Desarrollo y Versatilidad del Algortimo de Discretización Ameva

(Development and Versatility of the Discretization Algorithm Ameva)

International Doctoral Dissertation presented by
D. Miguel Ángel Álvarez de la Concepción
advised by
Dr. Juan Antonio Álvarez García
and **Dr. Luis Miguel Soria Morillo**.

April 2018.

To my wife and children.

I would never have become who I am without you all.

Acknowledgments

I would like to thank to my thesis directors, Dr. Juan Antonio Álvarez García and Dr. Luis Miguel Soria Morillo, all the effort, support and time dedicated to the elaboration of this work and to many others that we have shared in the past. They have encouraged me to conclude what I saw so distant and actually was so close. Also, I would like to thank to Dr. Juan Antonio Ortega Ramírez, Dr. Luis González Abril and Dr. Francisco Velasco Morente for the good advises, both professional and personal, that they have given me and for believing in me since the beginning of my stage as researcher. I will be eternally grateful.

To my parents, Rafael and María de los Ángeles, for having instilled in me the values of hard work and sacrifice, which have allowed me to achieve great goals in my life, and my sister, Sara Inmaculada, for her unconditional love.

To all my colleagues in the Computer Languages and Systems department, of the research group and especially of the Scientific Computer Center of Andalusia where I have spent most of the time. We have worked and enjoyed some free time together, we have laughed, we have cried and those moments are what have made us great and unforgettable friends. In particular I want to thank to José Ignacio Sánchez Venzalá and Antonio Manuel Bellido Moreno for their friendship reflected in the good times we have spent inside and outside the workplace. I will never forget you.

Finally, and in a more special and unique way, thanks to my wife Cristina and my children Adriana and Hugo. You are special and I tell you every day. You are the engine of my life and there is not even once time I have made a hard decision without thinking of you all before. Thanks for the love that you give me, especially in the most difficult moments where your devotion and blind trust is evident. From now on we will have all the time in the world to play with dolls.

To all, many thanks from heart.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

This thesis, presented as a set of research papers, studies the problem of activity recognition and fall detection in mobile systems where the battery draining and the accuracy are the main areas of researching. These problems are tackled through the establishment of a new selection, discretization and classification algorithm based on the core of the algorithm Ameva. Thanks to the discretization process, it allows to get an efficient system in terms of energy and accuracy.

The new activity recognition algorithm has been designed to be run in mobile systems, smartphones, where the energy consumption is the most important feature to take into account. Also, the algorithm had to be efficient in terms of accuracy giving an output in real time. These features were tested both in a wide range of mobile devices by applying usage data from recognized databases and in some real scenarios like the EvAAL competition where non-related people carried a smartphone with the developed system. In general, it had therefore been possible to achieve a trade-off between accuracy and energy consumption.

The developed algorithm was presented in the Activity Recognition track of the competition EvAAL (Evaluation of Ambient Assisted Living Systems through Competitive Benchmarking), which has as main objective the measurement of hardware and software performance. The system was capable of detecting some activities through the established set of benchmarks and evaluation metrics. It has been developed for multi-class datasets and obtains a good accuracy when there is approximately the same number of examples for each class during the training phase. The solution achieved the first award in 2012 competition and the third award in 2013 edition.

PART I

Preface

CHAPTER 1

INTRODUCTION

It is strange that only extraordinary men make the discoveries, which later appear so easy and simple - Georg Christoph Lichtenberg

1.1 Research motivation

This doctoral thesis will present the work of the doctorate in discretization and classification systems based on discrete techniques and its application in the area of activity recognition. The energy saving and the low computational complexity, as well as the increasing of the performance, are the main objectives that have been defined to optimize the execution of algorithms in mobile devices with the lower possible impact.

As part of the research team, we decided to continue with one of the main topics that my research group did. It is the Ameva discretization algorithm and there were some open discussions around the possibilities that it had, starting with the relation between discretization and classification.

The problem of classification is one of the main issues in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of features. The induction of classifiers from data sets of classified instances is a central problem in machine learning. For that purpose, a large number of methodologies based on SVM [46], Naive Bayesian [116], C5.0 [49], etc. have been developed.

One of the most important preprocess in classification is the discretization. This process establishes a relationship between continuous variables and their discrete transformation through functions. Therefore, it is possible to model qualitatively a series of continuous values if a label is assigned to them. Some studies [34] have shown that execute a prior process to discretize continuous features is more efficient than work directly with the continuous values. This process reduces the computation time and memory usage in classification algorithms and it is used to manage the set of values of a feature more effectively. One of the contributions in this vein of the present doctoral thesis is a discretization method based on the Ameva discretization algorithm.

Ameva algorithm reduces execution time and the number of intervals as it was demonstrated in [47]. This behavior is outstanding when the data set has a large number of classes, although, under certain circumstances, it has a slight reduction in the capacity of identification.

The Ameva discretization algorithm performs this process effectively and quickly, so the set of values of a feature is greatly reduced, but do not reduce the number of features. Because Ameva uses the statistic χ^2 to determine the relationship between features and classes, it is possible to use this algorithm to determine the relationship between features. In fact, this is one of the open questions derived from this research.

In the other hand, activity recognition and fall detection are two areas of great interest as can be seen in Figure 1.1 and detailed in Table 1.1, where the result of some Web of Science reports have been exported. The number of research papers in these areas has been grown in the last decades due to the facilities offered by the technologies companies with new and really good devices. They do not only allow to get relevant information from their embedded sensors, but share the information across health entities or other users in general-purpose applications.

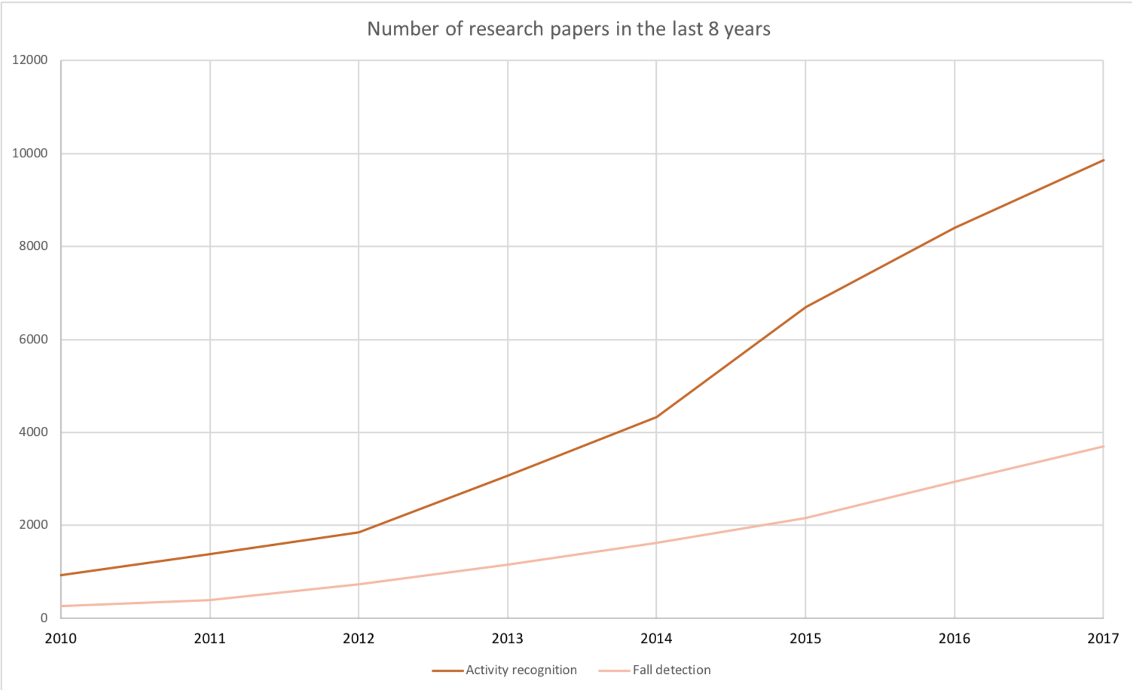


Figure 1.1: Growth in activity recognition and fall detection research papers since 2010.

In terms of activity recognition, the appearance of affordable gyroscopes and accelerometers allowed researches to use them in their investigations. From 2000 to now, wearable sensors and smartphone devices increased the number of possibilities of getting information and data in some ways. This means that the generation of products and services around the world is also more efficient.

In the other hand, the number of research papers in fall detection is lower because it didn't have an own well defined area (the majority of researchers included it in activity recognition), but the increasing of the published papers in the last years is proportionally bigger. Since 2010, fall detection has gained importance for elderly people, for this reason, medical companies are investing several resources.

Year	# research papers	
	Activity recognition	Fall detection
2010	930	267
2011	1373	390
2012	1848	726
2013	3062	1147
2014	4329	1620
2015	6697	2167
2016	8402	2931
2017	9862	3699

Table 1.1: Detail of number of research papers reported by Web of Science since 2010.

Within this context, this thesis proposal continues the thesis of one of the advisors, Luis Miguel Soria Morillo [86], where one of the topics was mobile physical activity recognition systems. He ended up that one of the needs was the decreasing of the computational time and therefore of battery draining and, here, it is where a new algorithm based on the Ameva discretization algorithm was developed as part of this thesis.

All this research was done as part of two research projects: HERMES (Healthy and Efficient Routes in Massive Open-Data Based Smart Cities-Citizen) and Simon (Saving Energy by Intelligent Monitoring)(1) where the objectives were to achieve energy efficiency in the day-to-day life of citizens and monitoring the citizens of a smart city.

(1)This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness HERMES (TIN2013-46801-C4-1-r) and Simon (P11-TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

1.2 Research methodology

This research follows the standard scientific research technique [66] which includes the following phases:

1. **Define research problem.** Most of the activity recognition systems applied to mobile devices have a high computational cost and in this dissertation, a new one is designed that has a trade-off between battery draining and performance.

2. **Review the literature.** During the period of this doctoral thesis, a deep searching and reading of the research based on activity recognition systems has been done and the references of each paper show it.

3. **Formulate hypotheses.** As part of the previous works made in our research group, we discussed how to apply discretization techniques in order to reduce the computational cost in the analysis of the data acquired by the mobile devices without decreasing the performance.

4. **Design research.** The whole system was analyzed in order to find the key points that are susceptible to changes and we found some of them where the discretization algorithm could be applied. The main ones were the pre-process and the post-process steps.

5. **Collect data.** Hundreds of megabytes of real data were collected using some devices carried out by dozens of people of diverse age. Also, a few well-known databases from other researchers were used as well.

6. **Execution of the project.** The designed algorithm was executed using the previous data as training data and after that, it was executed in real environments.

7. **Analyze data.** The information obtained by the system was analyzed using standard metrics and was compared with previous results.

8. **Interpret and report.** Once the information was analyzed and interpreted, some papers were published as a result of our hypothesis.

1.3 Research question

The research question that leads this thesis dissertation is: *are we able to reduce the computational cost in the activity recognition and fall detection system developed by us without decreasing the performance and therefore increasing the battery life using Ameva?*

Within the context of activity recognition, we focused on obtaining and discretizing data, generating some statistics and clustering them in activities using Ameva. To this end, activity recognition systems from the literature and new proposals were tested and compared. Also, a software was specially developed for this purpose and tested in real scenarios.

Within the context of specific activities, we focused on rating as the maximum priority the detection of fall detection. The question became how to adapt the system in order to isolate the recognition of that activity in the set of activities configured in the system using the Ameva discretization algorithm. As part of the system, the same software developed in the previous case was used.

1.4 Success criteria

The success will be achieved if the research question is resolved. This means that we have to check that both the model and the supporting algorithms actually recognize and identify the activities and demonstrate their validity through real scenarios. The developed software demonstrates that the output matches the prediction that was formulated in the hypothesis.

Using the sensors in the mobile devices, hundreds of Megabytes of data were collected and computed from a set of 10 users. The learning process of each activity recognized by the system consisted of its performance for a time of 6 minutes and, for the recognition process, users were followed over a period of 72 hours. Later, new experiments were carried out on real scenarios by a set of a wide age people [27]. Apart of that, other well-know databases [101, 131, 70, 106] were used as well.

1.5 Properties analyzed and discussed

Activity recognition and fall detection is approached from two sides:

- **Performance.** In relation to the number of activities recognized by the system, we improved it developing a new methodology using the Ameva algorithm as an extension of the Luis Miguel Soria Morillo's thesis.
- **Battery draining.** In relation to the computational cost, the methodology uses the Ameva algorithm as well. It allowed to reduce the amount of data that has to be analyzed.

The work presented in this document provides a solution for the improvement of activity recognition in order to satisfy user preferences, by and minimizing energy consumption with high performance, using the Ameva algorithm.

1.6 Thesis outline

The document is structured as follows: In Part. I, the current introduction is presented in Chapter 1. Chapter 2 describes the Ameva discretization algorithm based on the χ^2 statistic and Chapter 3 introduces the problem of recognizing the actions of one or more subjects and presents the difficulty of identifying a specific activity among others, where is a major health risk that diminishes the quality of life for elderly people. In Part. II, three selected journal papers are provided. These journals are included in the Thomson Reuters JCR ranking integrated with the Institute for Scientific Information (ISI) web of knowledge. All of these papers are related to the problem of activity recognition in mobile systems:

- **Discrete techniques applied to low-energy mobile human activity recognition. A new approach**. M.Á. Álvarez de la Concepción, L.M. Soria Morillo, L. González-Abril, J.A. Ortega Ramirez. Published in Journal of Expert Systems with Applications, Elsevier, ISSN: 0957-4174, Date of Publication: October 2014, Volume: 41, Issue: 14, On Pages: 6138-6146, DOI: `https://doi.org/10.1016/j.eswa.2014.04.018`, [JCR-2014 2.240, JCR-5 2.571] [Q1 in three categories: Operations Research & Management Science (12/81), Engineering, Electrical & Electronic (48/249) and Computer Science, Artificial Intelligence (29/123)].

- **Low Energy Physical Activity Recognition System on Smartphones**. L.M. Soria Morillo, L. González-Abril, J.A. Ortega Ramirez and M.Á. Álvarez de la Concepción. Published in Journal of Sensors, MDPI AG, ISSN: 1424-8220, Date of Publication: March 2015, Volume: 15, Issue: 3, On Pages: 5163-5196, DOI: `https://doi.org/10.3390/s150305163`, [JCR-2015 2.033, JCR-5 2.437] [Q1 in Instruments & Instrumentation (12/56)].

- **Mobile activity recognition and fall detection system for elderly people using Ameva algorithm**. M.Á. Álvarez de la Concepción, L.M. Soria Morillo, J.A. Álvarez García, L González-Abril. Published in Journal of Pervasive and Mobile Computing, Elsevier, ISSN: 1574-1192, Date of Publication: January 2017, Volume: 34, On Pages: 3-13, DOI: `https://doi.org/10.1016/j.pmcj.2016.05.002`, [JCR-2016 2.349, JCR-5 2.874] [Q2 in two categories in 2016: Telecommunications (34/89) and Computer Science, Information Systems (53/146)].

A summary of these journal papers and rankings can be found in Table 1.2.

Title	Journal	I.F.	Ranking
Discrete techniques applied to low-energy mobile human activity recognition. A new approach	Expert Systems with Applications 2014	2.240	Q1
Low Energy Physical Activity Recognition System on Smartphones	Sensors 2015	2.033	Q1
Mobile activity recognition and fall detection system for elderly people using Ameva algorithm	Pervasive and Mobile Computing 2017	2.349	Q2 (2016)

Table 1.2: Summary of papers published in JCR indexed journals

In Part. III, other related relevant research work, published in conference proceedings, are included in this thesis. These include:

- **A qualitative methodology to reduce features in classification problems**. M.Á. Álvarez de la Concepción, et al. Published in the 25th International Workshop on Qualitative Reasoning. Date of Publication: July 2011, On Pages: 1-5.

- **The CICA GRID - A Cloud Computing Infrastructure on Demand with Open Source Technologies.** M.Á. Álvarez de la Concepción, et al. Published in the 14th International Conference on Enterprise Information Systems. ISBN: 978-989-8565-11-2, Date of Publication: 2012, Volume: 2, On Pages: 301-304, DOI: `https://doi.org/10.5220/0003992603010304`, CORE ranking C.

- **Activity Recognition System Using AMEVA Method**. L.M. Soria Morillo, et al. Published in the Journal of Evaluating AAL Systems Through

Competitive Benchmarking. ISBN: 978-3-642-37418-0, Date of Publication: 2013, Volume: 362, On Pages: 137-147, DOI: https://doi.org/10.1007/978-3-642-37419-7_11.

- **An adaptive methodology to discretize and select features**. M.Á. Álvarez de la Concepción, et al. Published in the Journal of Artificial Intelligence Research. ISSN: 1927-6974, Date of Publication: February 2013, Volume: 2, Issue: 2, On Pages: 77-86, DOI: https://doi.org/10.5430/air.v2n2p77.

- **Activity Recognition System Using Non-intrusive Devices through a Complementary Technique Based on Discrete Methods**. M.Á. Álvarez de la Concepción, et al. Published in the Journal of Evaluating AAL Systems Through Competitive Benchmarking. ISBN: 978-3-642-41042-0, Date of Publication: 2013, Volume: 386, On Pages: 36-47, DOI: https://doi.org/10.1007/978-3-642-41043-7_4.

- **Evaluating Wearable Activity Recognition and Fall Detection Systems**. J.A. Álvarez García, et al. Published in the 6th European Conference of the International Federation for Medical and Biological Engineering. ISBN: 978-3-319-11127-8, Date of Publication: 2015, Volume: 45, On Pages: 653-656, DOI: https://doi.org/10.1007/978-3-319-11128-5_163.

Finally, in Part IV, final remarks are made, conclusions are drawn and future work is discussed.

CHAPTER 2

AMEVA

Science is the acceptance of what works and the rejection of what does not. That needs more courage than we might think - Jacob Bronowski

2.1 Introduction

Currently, the classification algorithms have become an important part of any process of execution of tasks and usually determine the decisions to be taken when performing the next step. The rules that determine whether an example belongs to one class or another are given by a comprehensive study of its features, but sometimes it is possible that at first glance there are no such rules and it has to be deduced through the comparison between different samples and the similarity between them (feature analysis).

The class where an example belongs is the value of one feature that is deduced and the others are data collected by an comprehensive study. Also, each class should be defined by specifying the rules that differentiate it from the rest. In the other hand, establish exact values that decides whether an example belongs to one class or another is quite difficult. Thus, these rules are specified in ranges.

The values of the features are mostly continuous and rarely discrete. The discretization process enables the relationship between continuous variables and a discrete transformation through functions. This process is used to manage more effectively the set of values of a property.

In mathematics, the discretization is defined as the process of transforming continuous models and equations into the equivalent discrete models. Usually, it is done as a first step in making appropriate numerical evaluation and implementation on computers.

From this, it is proposed to discretize the values of the sets of examples, so inferring the classification rules is a much simpler process and it is not necessary to deal with infinite values, but to a set of them.

The Ameva discretization algorithm allows this process effectively and quickly, so the set of values of a characteristic is greatly reduced. This reduction of set of possible data of a characteristic favors the classification, making it faster and less expensive.

2.2 The Ameva discretization algorithm

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a data set of a continuous attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of ℓ classes of the variable denoted by

$$\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2$$

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \to \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of the property that is being discretized.

Let us consider a discretization $\mathcal{D}$ which discretizes the continuous domain of $\mathcal{X}$ into k discrete intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{[d_0, d_1], (d_1, d_2], \cdots, (d_{k-1}, d_k]\}$$

In this discretization, d_0 is the minimum value and d_k is the maximum value of the attribute $\mathcal{X}$, and the d_i values are in ascendant order.

If L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2, 3, \ldots, k$, then

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \cdots, L_k\}$$

Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, k$. The discretization variable $\mathcal{L}(k)$ of attribute $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view. Having two discrete attributes, a two-dimensional frequency table (called contingency table) as shown in the Table 2.1 can be built.

$C_i \mid L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell k}$	$n_{\ell\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

Table 2.1: Contingency table

In Table 2.1, n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j. $n_{i\cdot}$ is the total number of instances belonging to the class C_i, and n_j is the total number of instances that belong to the interval

L_j, for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$. So that:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, \quad N = \sum_{i=1}^{\ell}\sum_{j=1}^{k} n_{ij}$$

Given discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$, these are statistically independents if for all $C_i \in \mathcal{C}$ and $L_j \in \mathcal{L}(k)$, then

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$. In other case, a dependence between these attributes exists.

Therefore, a way to measure the dependence between $\mathcal{C}$ and $\mathcal{L}(k)$ is the study of value

$$\sum_{i=1}^{\ell}\sum_{j=1}^{k} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}\right)^2$$

Considering a relative measure, $\chi^2(k) \stackrel{def}{=} \chi^2(\mathcal{L}(k), \mathcal{C}|X)$ is defined by

$$\chi^2(k) = \sum_{i=1}^{\ell}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{N}} = N\left(-1 + \sum_{i=1}^{\ell}\sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}}\right) \tag{2.1}$$

It is straightforward to prove that

$$\max_{X, \mathcal{L}(k), \mathcal{C}} \chi^2(k) = N(\min\{\ell, k\} - 1) \tag{2.2}$$

Hence, the Ameva coefficient, $Ameva(k) \stackrel{def}{=} Ameva(\mathcal{L}(k), \mathcal{C}|X)$, is defined as follows:

$$Ameva(k) = \frac{\chi^2(k)}{k(\ell - 1)}$$

for $k, \ell \geq 2$. The Ameva criterion has the following properties:

- The minimum value of $Ameva(k)$ is 0 and when this value is achieved then both discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are statistically independent and viceversa.
- The maximum value of $Ameva(k)$ indicates the best correlation between class labels and discrete intervals. If $k \geq \ell$ then, for all $x \in C_i$ a unique j_0 exists such that $x \in L_{j0}$ (remaining intervals $(k - \ell)$ have no elements); and if $k < \ell$

then, for all $x \in L_j$, a unique i_0 exists such that $x \in C_{i0}$ (remaining classes have no elements) i.e. the highest value of the Ameva coefficient is achieved when all values within a particular interval belong to the same associated class for each interval.

- The aggregated value is divided by the number of intervals k, hence the criterion favors discretization schemes with the lowest number of intervals.
- From Equation 2.2, it is followed that $Ameva_{max}(k) \stackrel{def}{=} \max_{X,\mathcal{L}(k),\mathcal{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ if $k < \ell$ and $\frac{N}{k}$ otherwise. Hence, $Ameva_{max}(k)$ is an increasing function of k if $k \leq \ell$, and a decreasing function of k if $k > \ell$. Therefore, $\max_{k\geq 2} Ameva_{max}(k) = Ameva_{max}(\ell)$ i.e. the maximum value of the Ameva coefficient is achieved in the optimal situation, it is to say, when all values of C_i are in a unique interval L_j and viceversa.

Therefore, the aim of the Ameva method is to maximize the dependence relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k.

2.3 Optimizations of the algorithm

A deep study was done in order to understanding its strengths and weaknesses and there are two main characteristics based on the Ameva coefficient that should be presented.

- The problem of finding a discretization scheme $L(k)$ with a global optimal value of Ameva criterion is highly combinatorial. The first approach reduces the task of discretization to a reasonable computational cost, so it can be applied to continuous attributes with a large number of different values and computes a local maximum value of the coefficient Ameva. The algorithm works with a top-down scheme by starting as a single interval and dividing it into two new intervals using as criterion the results obtained in the search for the optimal value of $Ameva(k)$ after the division. The algorithm has two parts and is detailed in the Algorithm 1. This algorithm finds the best and smallest number of intervals based on the value of the coefficient Ameva.

- The number of class labels is small and, as the Ameva criterion provides a reduced number of intervals, it is possible to define a genetic algorithm that finds the maximum overall value of the Ameva coefficient with a not excessive computational cost.

Algorithm 1: The Ameva algorithm

Data: $X = \{x_1, x_2, \ldots, x_N\}, x_i < x_{i+1}$ and $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2$

Result: D: discretization scheme

$D = \{[x_0, x_N]\}$;

$A = \{x_0, x_N\}$;

$global = 0$;

while *true* **do**

  for $j \leftarrow 1$ **to** $N-1, x_j \nexists A$ **do**

    $B = sort(A \cup x_j)$;

    $local_j = Ameva(\mathcal{L}(B), \mathcal{C}|X)$;

  end

  if $\max(local_1, local_2, \ldots, local_{N-1}) > global$ **then**

    $A = sort(A \cup x_j)$;

    $D = \{[x_0, x_1], [x_1, x_2], \ldots, [x_{|A|-1}, x_{|A|}]\}$;

    $global = local$;

  else

    break;

  end

end

Based on the previous findings, two optimization of Ameva discretization algorithm that provides improvements in two ways were performed:

- In the first optimization, the accuracy was increased and the computational cost was decreased over other discretization algorithms.

- In the second optimization, the accuracy was increased over its original version and the computational cost was reduced compared to other discretization algorithms.

2.3.1 First optimization

In this optimization, the algorithm stops when the number of intervals is equal to the number of existing classes in the class variable. As a result, the accuracy is reduced and the computational time is increased, but in comparison with other discretization algorithms, both are better.

The algorithm is detailed in the Algorithm 2.

Algorithm 2: The first optimization of the Ameva discretization algorithm

Data: $X = \{x_1, x_2, \ldots, x_N\}, x_i < x_{i+1}$ and $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2$

Result: D: discretization scheme

$D = \{[x_0, x_N]\}$;

$A = \{x_0, x_N\}$;

while $|D| < \ell$ **do**

  for $j \leftarrow 1$ **to** $N-1, x_j \nexists A$ **do**

    $B = sort(A \cup x_j)$;

    $local_j = Ameva(\mathcal{L}(B), \mathcal{C}|X)$;

  end

  $A = sort(A \cup x_j / \max(local_1, local_2, \ldots, local_{N-1}) = local_j)$;

  $D = \{[x_0, x_1], [x_1, x_2], \ldots, [x_{|A|-1}, x_{|A|}]\}$;

end

2.3.2 Second optimization

After studying the Ameva discretization algorithm and checking the partial results at each step of the value of the Ameva coefficient, it was discovered that there are more than one max point. Thus, the implemented function got in some cases more than one local maximum.

The original Ameva discretization algorithm ends when the next value of the Ameva coefficient is less than the previous one as can be seen in the Algorithm 1, so if there is another local maximum value of the Ameva coefficient, then it will not be considered.

Because it has to be considered, in this optimization, the algorithm stops also when the number of intervals is the number of classes, but it gets the best discretization scheme.

The algorithm is detailed in the Algorithm 3.

2.4 Feature reduction

If $\ell = 1$ or $k = 1$ then it is not possible to use the Ameva discretization algorithm[(1)]. Let us see the two cases in Table 2.2 and Table 2.3.

Equation 2.1 can not be calculated using Table 2.2 because it is not possible to divide by 0. Nevertheless, all the instances belong to the same class, therefore can be concluded that the dependence is maximum. In this case, let us indicate that $A^*(1) = 1$.

Regarding to Table 2.3, the Ameva discretization algorithm can not be used because $X^2(k) = 0$ and the Ameva coefficient does not give any information about the dependence. However, the dependence is not minimum and a new coefficient is

[(1)]Let us indicate these pathological cases do not happen in a standard discretization, but it will be necessary taking into account in the presented methodology in the current section

Algorithm 3: The second optimization of the Ameva discretization algorithm

Data: $X = \{x_1, x_2, \ldots, x_N\}, x_i < x_{i+1}$ and $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2$

Result: D: discretization scheme

$D = \{[x_0, x_N]\}$;

$A = \{x_0, x_N\}$;

$A' = \{x_0, x_N\}$;

$global = 0$;

while $|D| < \ell$ **do**

  for $j \leftarrow 1$ **to** $N-1, x_j \nexists A'$ **do**

    $B = sort(A' \cup x_j)$;

    $local_j = Ameva(\mathcal{L}(B), \mathcal{C}|X)$;

  end

  $A' = sort(A' \cup x_j / \max(local_1, local_2, \ldots, local_{N-1}) = local_j)$;

  if $\max(local_1, local_2, \ldots, local_{N-1}) > global$ **then**

    $A = A'$;

    $D = \{[x_0, x_1], [x_1, x_2], \ldots, [x_{|A|-1}, x_{|A|}]\}$;

    $global = local$;

  end

end

necessary. By taking into account that if all instances are distributed equally in all classes, the dependence is minimum, and if exists i such that $n_{i1} = N$, the dependence is maximum. Hence the following coefficient, called Entropy, is considered:

$$A(1) = 1 + \frac{1}{N \ln \ell} \sum_{i=1}^{\ell} n_{i1} \ln(\frac{n_{i1}}{N})$$

It holds that $0 \leq A(1) \leq 1$ and the next:

- If $A(1) = 0$, then $n_{i1} = \frac{N}{\ell}$ (minimum dependency).
- If $A(1) = 1$, then a unique n_{i1} exists that $n_{i1} = N$ (maximum dependency).

$C_i \vert L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N
$n_{\cdot j}$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N

Table 2.2: Contingency table at first case ($\ell = 1$)

$C_i \vert L_j$	L_1	$n_{i\cdot}$
C_1	n_{11}	n_{11}
$\vdots$	$\vdots$	$\vdots$
C_i	n_{i1}	n_{i1}
$\vdots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$n_{\ell 1}$
$n_{\cdot j}$	N	N

Table 2.3: Contingency table at second case ($k = 1$)

This new coefficient allows to establish a new methodology to reduce features using the Ameva discretization algorithm.

Given an attribute X_i where $i = 1, 2, \ldots, s$, the Ameva discretization algorithm is applied to this attribute so obtained intervals are considered as a new set of classes. This set of classes is denotes as follows:

$$\mathcal{C}^i = \{C_1^i, C_2^i, \ldots, C_\ell^i\} \tag{2.3}$$

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathcal{C}^i$ where $p = 1, 2, \ldots, \ell$. From Equation 2.3, for each attribute X_j with $j = 1, 2, \ldots, s$, a g_{ijp} value is obtained from $\mathcal{C}^i$ as follows:

- If the X^p data subset all belong to the same class $\mathcal{C}^i$ then $g_{ijp} = A^*(1) = 1$.
- If the subset of data belongs to different classes, then:
  - If values of the attribute X_j are always in the same interval, then $g_{ijp} = A(1)$.

– If values of the attribute X_j are not always in the same interval, then $g_{ijp} = Ameva_N(\ell_i)$, where $Ameva_N(\ell_i)$ is defined as follows[(2)]:

$$Ameva_N(\ell_i) = \frac{\ell_i'}{N_p} Ameva(\ell_i)$$

provided that N_p is the number of instances of the class X^p and l_i' is the number of intervals of the attribute X_i for which there is at least one value in the data subset.

Given $i, j = 1, 2, \ldots, s$, a g_{ijp} value can be obtained applying this methodology for all class $C_p \in \mathcal{C}$ for $p = 1, 2, \ldots, \ell$, and by considering different statistics as follows:

$$g_{ij}^{min} = \min_p g_{ijp}$$

$$g_{ij}^{geo} = \sqrt[\ell]{\prod_{p=1}^{\ell} g_{ijp}}$$

$$g_{ij}^{ari} = \frac{1}{\ell} \sum_{p=1}^{\ell} g_{ijp}$$

$$g_{ij}^{max} = \max_p g_{ijp}$$

Also, it is well-known that $g_{ij}^{min} \leq g_{ij}^{geo} \leq g_{ij}^{ari} \leq g_{ij}^{max}$ is holded.

The main properties of the matrix $G = (g_{ij})$, that is,

$$G = \begin{pmatrix} 1 & g_{12} & \cdots & g_{1s} \\ g_{21} & 1 & \cdots & g_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s1} & g_{s2} & \cdots & 1 \end{pmatrix}$$

are the following: i) it is squared but non symmetric matrix; ii) the values of the main diagonal are 1; iii) $0 \leq g_{ij}, g_{ji} \leq 1$.

(2) This new Ameva coefficient is chosen in order to obtain a normalized value $0 \leq Ameva_N(\ell_i) \leq 1$ as same as $A(1)$. Furthermore, it is straightforward to prove that if $i = j$ for $i = 1, 2, \ldots, s$, then $g_{ijp} = 1$, for all $p = 1, 2, \ldots, \ell$.

From the G matrix, a method of generating rules of dependence between attributes can be defined. For example, a possible rule is the next: given a threshold value, θ, if $max(g_{ij}, g_{ji}) > \theta$ and $i < j$ where $i, j = 1, 2, \ldots, s$ and $i \neq j$, then the X_j attribute is eliminated.

A first approach has been published in [6] and [26] as part of the future work and a new way to use the Ameva discretization algorithm.

CHAPTER 3

ACTIVITY RECOGNITION AND FALL DETECTION

The good thing about science is that it is true whether or not you believe in it - Neil deGrasse Tyson

3.1 Introduction

Activity recognition is an important area of research and applications and its goal is an automated analysis (or interpretation) of ongoing events and their context from data. Its applications include surveillance systems, patient monitoring systems and a variety of systems that involve interactions between persons and electronic devices. Most of these applications require recognition of high-level activities, often composed of multiple simple (or atomic) actions of persons. In this context, one of the main activities that has acquired an special attention in the medical area is falling.

Nevertheless, sensing the activity of the user has been very hard to tackle. Most of the context gathering is achieved through the use of hardware sensors, typically measure a specific phenomenon such as a GPS location or the ambient temperature. However, they could be more complex, for example a camera might track a user's position and through a connection to models can identify whether a user is sitting or standing or even where on a screen he is looking. It should be said that for the latter work, complex setups are needed and hence, a high cost occurs with the deployment of hardware sensors. Although, the calculation of the physical activity of a user, based on data obtained, remains a current research topic, numerous limitations have been identified that make these systems uncomfortable for users in general.

The development of activity recognition applications using smartphones has several advantages such as easy device portability without the need for additional fixed equipment, and comfort to the user due to the unobtrusive sensing. This contrasts with other established approaches which use specific purpose hardware devices or sensor body networks. Although the use of numerous sensors could improve the performance of a recognition algorithm, it is unrealistic to expect that the general public will use them in their daily activities because of the difficulty and the time required to wear them. One drawback of the smartphone-based approach is that energy and services on the mobile phone are shared with other applications and this become critical in devices with limited resources.

Thus, the present section is focused on the recognition of physical activities carried out by users through their mobile devices, and hence special attention must be paid to energy consumption and the computational cost of the methods used. Taking previous works into account, physical activity monitoring through smartphones presents the following challenges:

- To decrease, as far as possible, the risk of forgetting the processing device so that continuous monitoring can be performed anywhere and any time.
- To reduce the drain of energy on the smartphone, by developing an accurate

and efficient system.

- To integrate learning and monitoring on the device itself, in real-time and without sharing server information.

In order to reduce the cost associated to accelerometer and gyroscope signal analysis, this work opts for an approach based on a discretization method that uses only accelerometer sensors since with these good results can be achieved and a lot energy can be saved. Furthermore, the data is processed in the mobile itself and the result is obtained in real time. Thanks to this discretization process, the classification cost is much lower than it would be with continuous variables, and therefore the life of the battery is longer.

3.2 Embedded sensors and activities

Throughout this work, a sensor with a range of sampling frequencies between 25 Hz and 150 Hz was used and it was configured to operate at 50 Hz in order to prevent excessive use of data and to reduce the computational cost. As it will be seen later, it proposes placing the device at the hip, where acceleration forces are lower than at the ankle, and hence frequencies at this position are also lower.

On the other hand, a lower frequency allows the computation cost to be reduced thanks to the fact that each feature is obtained from accelerometry data. Furthermore, by reducing this processing time, the system becomes faster, more efficient and consumes less energy. Indeed, for contextual systems, with their intensive use of sensors, the high-energy consumption required must be taken into account not only in obtaining the data but also in its processing.

Battery life is not a secondary consideration, since, according to a survey performed by some companies like YouGov(1) or LG(2), it stands as one of the most

(1)https://today.yougov.com/news/2016/09/22/smartphone-longer-battery-life-headphone-jack/
(2)http://wccftech.com/lg-us-smartphone-users-survey/

important purchase decision factor for buyers of smartphones. Users acceptance, in the context of aware applications in general and of activity recognition systems in particular, is therefore critical. For this reason, not only has an accurate and fast system been developed, but a low energy consumption model is also presented from the viewpoint of discrete techniques.

In the other hand, even if a large number of personal devices that able to monitor the activity level have been developed by big commercial companies, the limit imposed on the number of activities constitutes the main disadvantage, since they can only detect certain parameters. Although these features are significant, the number of these activities may be insufficient, for example, for users involved in a physical rehabilitation process which must be monitored by doctors. In these cases, activity recognition should have greater granularity to detect activities like fall detection as an specific example. In this work, far from being a static system, the number and type of activities recognized depends on the user since they can carry out hot-training on the system, which may involve adding new activities, and hence other activities can be detected. This is crucial for the accomplishment of a highly customizing environment where the users themselves can determine which activities are important and which remain irrelevant.

3.3 Data collection

The data was obtained from two different sources: from some volunteers through the application that was installed in their smartphones and from some external databases.

The volunteers were a group of 30 people within an age bracket of 19-48 years and was randomly partitioned into two sets where 70% of the volunteers were selected for generating the training data and 30% the test data. They followed a protocol in which the activities were performed wearing a waist-mounted smartphone.

The external datasets that were tested in the system were well-known datasets

(PAMAP2 [101], USC-HAD [131], WISDM [70] and Shoaib [106]) and their details are shown in the Table 3.1.

dataset	|people|	age	|activities|
Own dataset	30	19-48	6
PAMAP2	9	26-31	18
USC-HAD	14	21-49	12
WISDM	29	-	6
Shoaib	10	25-30	7

Table 3.1: Datasets information used during the activity recognition process

Training and recognition sets were obtained using time windows of 5 seconds composed of a set of accelerometer readings from which it is possible to calculate a variety of features. This time was chosen due to the importance of ensuring that, in each time window, there is at least one activity cycle, defined as an complete execution of an activity pattern.

Based on these time windows, a signal module was selected. This eliminates the problem caused by the device rotation and increases user comfort by removing the restriction of maintaining the same orientation during the learning and recognition process.

For each data in a time window size N, $a_i = (a_i^x, a_i^y, a_i^z)$, $i = 1, 2, \ldots, N$ where x, y and z represent the three accelerometer axes, the accelerometer module is defined as follows:

$$|a_i| = \sqrt{(a_i^x)^2 + (a_i^y)^2 + (a_i^z)^2}$$

Once it is defined, then some statistics can be defined as well:

- Mean: $\bar{a} = \frac{1}{N} \sum_{i=1}^{N} |a_i|$
- Minimum: $a_{min} = \min\{|a_1|, |a_2|, \ldots, |a_N|\}$
- Maximum: $a_{max} = \max\{|a_1|, |a_2|, \ldots, |a_N|\}$

- Median: $Me = \left|a_{(N+1)/2}\right|$ if $N\%2 \neq 0$ and $\frac{\left|a_{N/2}\right|+\left|a_{(N+1)/2}\right|}{2}$ otherwise
- Standard deviation: $S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(|a_i| - \bar{a})^2}$
- Signal magnitude area [83]: $SMA = \sum_{i=1}^{N}\left(|a_i^x| + |a_i^y| + |a_i^z|\right)$
- Mean deviation: $D_m = \frac{1}{N}\sum_{i=1}^{N}\left||a_i| - \bar{a}\right|$

In addition to the above variables, a new set of statistics is generated from the frequency domain of the problem. In order to obtain these frequency-domain features, the Fast Fourier Transform (FFT) is applied for each time window and the frequency components associated are defined as follows:

$$y_k = \sum_{j=0}^{F-1} |a_j|\, e^{-\frac{2\pi i}{F}jk}$$

where F is the cardinality of frequency components and $k = 1, 2, \ldots, F$.

From here, the following attributes are identified:

- Min module: $m_{min} = \min\{|y_1|, |y_2|, \ldots, |y_F|\}$
- Max module: $m_{max} = \max\{|y_1|, |y_2|, \ldots, |y_F|\}$
- Min frequency module: $f_{min} = \{f(y_k) \ \mid \ |y_k| = m_{min}\}$
- Max frequency module: $f_{max} = \{f(y_k) \ \mid \ |y_k| = m_{max}\}$

where $|y_k|$ and $f(y_k)$ are the module and frequency associated to the k-th component, respectively.

The previous statistics are the minimum set of statistics used in the evaluation process.

3.4 Methodology

It involves working in the domain of discrete variables to perform learning and recognition of activities using the Ameva discretization algorithm. This concept reduces the high computational cost required for learning algorithms based on continuous variables, which have been used for this purpose over the years. The whole methodology consists on some steps that are described below.

Using the Ameva discretization algorithm that has been deeply described in 2.2, for each statistic $\mathcal{S} \in \{S_1, S_2, \ldots, S_m\}$, where m is the number of statistics, the discretization process is performed.

Given the result of the previous discretization process, a matrix of order $k_p \times 2$ is obtained, where k_p is the number of class intervals and 2 denotes the dimension made by the $inf(L_i^p)$ and $sup(L_i^p)$ interval limits i of the statistic p. Hence, a three-dimensional matrix containing the statistics and the set of interval limits for each statistic is called the Discretization Matrix and is denoted by

$$\mathcal{W} = (w_{pij})$$

where $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$, and $j = 1, 2$. Figure 3.1 shows the contents of the Discretization Matrix obtained during a learning process.

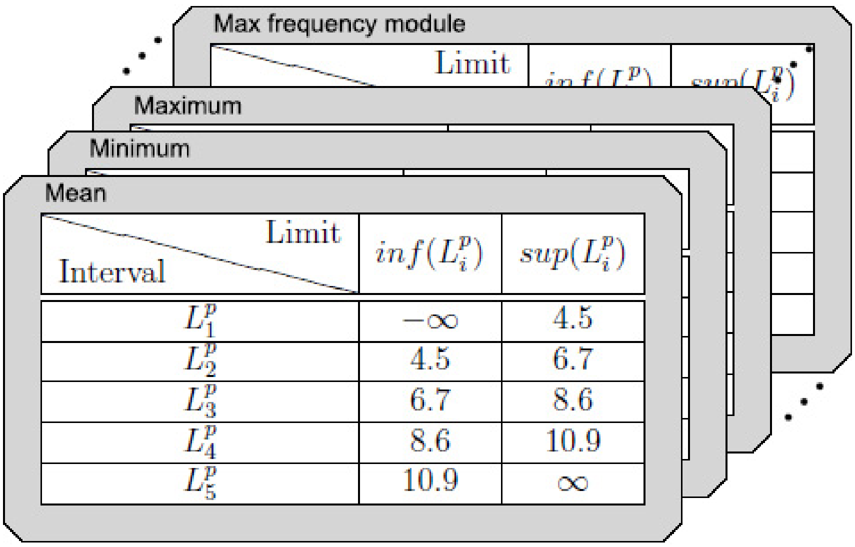


Figure 3.1: Example of Discretization Matrix.

Therefore, the Discretization Matrix determines the interval in which each item of data belongs for the different statistical associated values, by carrying out a simple and fast discretization process.

Based on previously generated intervals, the algorithm has to provide a probability associated with the elements of the training set $x \in \mathcal{X}$.

For that purpose, a Class Matrix, $\mathcal{V}$, is defined as a three-dimensional matrix that contains the number of items from the training set associated with an interval L^p in a activity C_i for each statistic S_p of the system. This matrix is defined as follows:

$$\mathcal{V} = (v_{pij})$$

where $v_{pij} = \#\{x \in X \mid inf(L_i^p) < x \leq sup(L_i^p)\}$, $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$. Hence, each position in the Class Matrix is uniquely associated with a position in the Discretization Matrix, which is determined by its range.

Figure 3.2 shows the contents of the Class Matrix obtained during a learning process for only 3 out of 8 activities recognized by the system. Within these activities, five intervals determined by the Ameva discretization algorithm can be observed in the Class Matrix for the mean statistic.

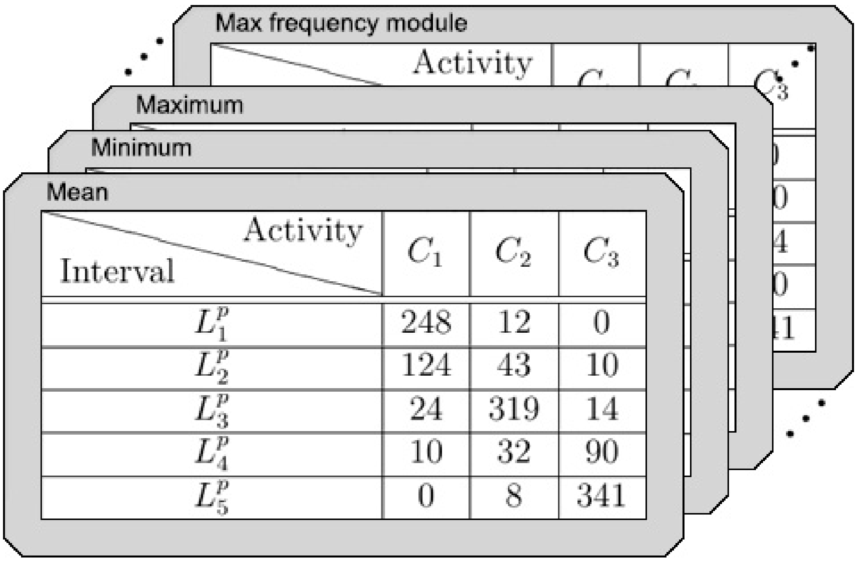


Figure 3.2: Example of Class Matrix.

The next step is to specify the likelihood that a given value x associated to a statistic S_p corresponds to an activity C_i in an interval L^p. This is done by the

Activity-Interval Matrix denoted by $\mathcal{U}$ and defined as follows:

$$\mathcal{U} = (u_{pij})$$

where each value is defined as

$$u_{pij} = \frac{v_{pij}}{v_{p\cdot j}} \frac{\sum_{q=1,q\neq j}^{\ell} \left(1 - \frac{v_{piq}}{v_{p\cdot q}}\right)}{\ell - 1}$$

and $v_{p\cdot j}$ is the total number of time windows of the training process labelled with the activity C_j for the statistic S_p, and $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$, and $j = 1, 2, \ldots, \ell$.

Also, each value u_{pij} can be seen as a degree of belonging for a given x, identified with a activity C_i, to be included in the interval L^p of the statistic S_p. Similarly, they hold the following properties:

- $u_{pij} = 0 \iff v_{pij} = 0 \vee v_{piq} = v_{p\cdot q}, q \neq j$
- $u_{pij} = 1 \iff v_{pij} = v_{p\cdot j} = v_{pi\cdot}$

Figure 3.3 shows a set of values for each of the positions of the Activity-Interval Matrix. These results have been obtained from the training set of the Class Matrix described above in Figure 3.2.

Once the discretization intervals and the probabilities of belonging have been obtained, the process of classification is described. It is divided into two main parts: the way to perform the recognition of physical activity and the task of determining the frequency of a particular activity is presented.

The most likely activity is decided by a majority voting system starting from the Activity-Interval Matrix, $\mathcal{U}$, and a set of data $x \in \mathcal{X}$ for the statistics $\mathcal{S}$.

It consists of finding an activity $C_i \in \mathcal{C}$ that maximizes this likelihood and it is defined in the following expression denoted by mpa (most probable activity):

$$mpa(x) = C_k$$

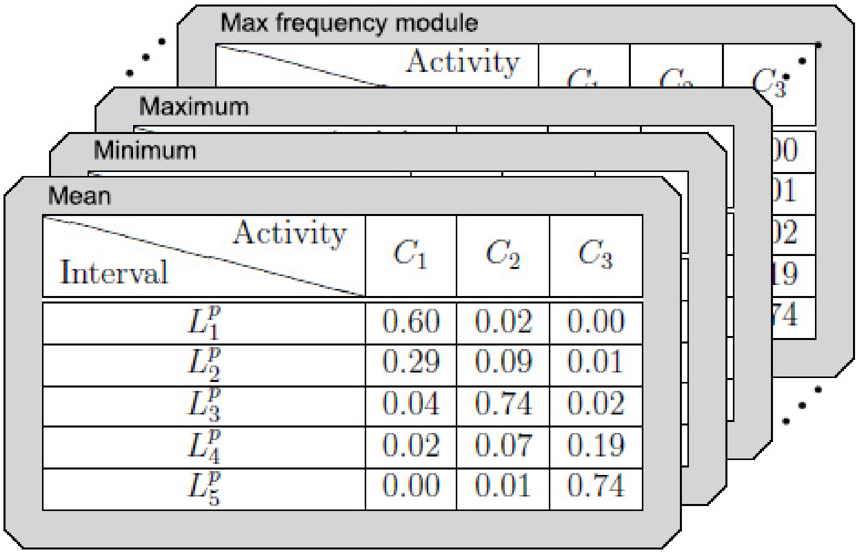


Figure 3.3: Example of Activity-Interval Matrix.

where $k = \max_j \sum_{p=1}^{m} u_{pij} \;\mid\; x \in (inf(L_i^p), sup(L_i^p)]$). This expression assumes that all statistics provide the same information to the system and there is no correlation between them.

A detailed analysis and results have been published in [25], [88] and [28] as part of the main topic of this doctoral thesis.

PART II

Selected research works

CHAPTER 4

DISCRETE TECHNIQUES APPLIED TO LOW-ENERGY MOBILE HUMAN ACTIVITY RECOGNITION. A NEW APPROACH

Overview

This paper addresses a low-energy solution for human activity recognition systems using discrete techniques based on the Ameva algorithm. It is used both in the pre-process step where the continuous values are transformed to discrete values and the classification step where the decision is taken based on a set of intervals that determine the activity where the value belongs to.

Unlike other systems currently in use, this proposal enables recognition of high granularity activities by using accelerometer sensors and remains free of dependence on the features of any recognition activity dataset. Thanks to this discretization process, the classification cost is much lower than it would be with continuous variables and gets the life of the battery to be longer, so the accuracy of activity recognition systems can be increased without sacrificing efficiency. Furthermore, the data is processed in the mobile itself, obtaining the results in real time, and therefore the efficiency is better than if data were sent to a server.

The strengths of the system are both the high success rate and the reduced computational cost associated with the processing of data during the recognition process.

Context

This research was initiated as a part of one of the possible uses of the Ameva algorithm that this PhD candidate carried out with activity recognition systems. The idea of reducing the computational cost in mobile devices using discrete variables was the main goal of the paper. Also, it was the first attempt to use Ameva as a classification algorithm due to its high computational performance. This paper is the result of over 1 year's work in this area.

Journal information

The Journal *Expert Systems with Applications* was selected for the submission of this paper that is indexed in JCR with an Impact Factor of 2.240 and a 5-Year I.F. 2.571. Also, it stands in ranking Q1 in three categories: Operations Research & Management Science (12/81); Engineering, Electrical & Electronic (48/249); Computer Science, Artificial Intelligence (29/123).

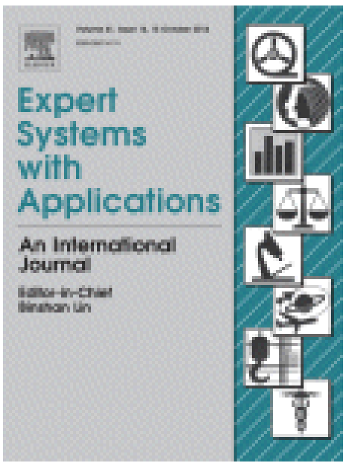


Figure 4.1: Expert Systems with Applications cover.

Expert Systems with Applications 41 (2014) 6138–6146

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Discrete techniques applied to low-energy mobile human activity recognition. A new approach

Álvarez de la Concepción M.A. [a], Soria Morillo L.M. [a,*], Gonzalez-Abril L. [b], Ortega Ramírez J.A. [a]

[a] Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain
[b] Applied Economics I Dept., University of Seville, 41018 Seville, Spain

A R T I C L E   I N F O

Keywords:
Pattern recognition
Discretization method
Qualitative systems
Smart-energy computing
Energy saving

A B S T R A C T

Human activity recognition systems are currently implemented by hundreds of applications and, in recent years, several technology manufacturers have introduced new wearable devices for this purpose. Battery consumption constitutes a critical point in these systems since most are provided with a rechargeable battery. In this paper, by using discrete techniques based on the Ameva algorithm, an innovative approach for human activity recognition systems on mobile devices is presented. Furthermore, unlike other systems in current use, this proposal enables recognition of high granularity activities by using accelerometer sensors. Hence, the accuracy of activity recognition systems can be increased without sacrificing efficiency. A comparative is carried out between the proposed approach and an approach based on the well-known neural networks.

1. Introduction

In recent years, thanks largely to the growing interest in monitoring certain sectors of the population, such as elderly people with dementia and people in rehabilitation, activity recognition systems have experienced an increase in both number and quality of results. However, most of these results incur high computational costs, and hence cannot be applied on a general-purpose mobile device due to their excessive energy consumption.

Although, the calculation of the physical activity of a user, based on data obtained from an accelerometer, remains a current research topic, numerous limitations have been identified that make these systems uncomfortable for users in general.

The first difference observed between the many systems developed is the type of sensor used. There are systems using specific hardware (Ravi, Dandekar, Mysore, & Littman, 2005), while others use general-purpose hardware (Hong, Kim, Ahn, & Kim, 2008). Obviously, the use of generic hardware constitutes a benefit for users, since their availability, low cost, and versatility are points greatly in their favour; not to mention the reduction in the risk of loss, since these devices have already been integrated into everyday objects, such as users' smartphones. However, general purpose devices are used for other purposes, such as making phone calls, surfing the Internet, and listening to music. For this reason, the physical activity recognition system must be executed in background mode and cause the least impact on the system as possible, in terms of complexity and energy consumption.

Another difference found between the proposals surveyed is the number and position of the sensors. In Brezmes, Gorricho, and Cotrina (2009), it can be observed that the accelerometer sensor is placed in a glove and a multitude of activities are recognized depending on the movement of the hand. In contrast, other studies use either various sensors placed on many parts of the body (Bicocchi, Mamei, & Zambonelli, 2010; Lepri, Mana, Cappelletti, Pianesi, & Zancanaro, 2010) or a wearable wireless sensor node with a static wireless non-intrusive sensory infrastructure (Paoli, Fernández-Luque, & Zapata, 2011) to recognize these activities. According to certain comparative studies and previous research based on multiple sensors, these last two types of sensors provide greater accuracy.

However, in studies, such as Hong et al. (2008), a sensor is kept in a user's pocket or worn on the hip, which is more wearable on the monitored person, and requires much lower infrastructure.

Once the most comfortable alternative for users is determined, some device sensors could be chosen to perform the activity monitoring. Some research uses data not only from accelerometers and a gyroscope (Dernbach, Das, Krishnan, Thomas, & Cook, 2012), but also from other sources, such as microphones, light sensors and voice recognition to determine the context of the user (Kwapisz, Weiss, & Moore, 2011) and ECG sensors (Li et al., 2010; Pawar, Chaudhuri, & Duttagupta, 2007; Ward, Lukowicz, Troster, & Starner, 2006) for this purpose.

* Corresponding author.
E-mail addresses: maalvarez@us.es (M.A. Álvarez de la Concepción), lsoria@us.es (L.M. Soria Morillo), luisgon@us.es (L. Gonzalez-Abril), jortega@us.es (J.A. Ortega Ramírez).

http://dx.doi.org/10.1016/j.eswa.2014.04.018
0957-4174/

The first example, which incorporates microphone and bluetooth devices, helps to obtain contextual information about the user environments and would be appropriate to perform a more in-depth analysis of the activity, for instance, whether the user is walking in a disco or at home, or whether s/he is alone or with someone. However, high-level activity recognition (walking, playing, running or standing up) is carried out using other sensors. On the other hand, ECG can help determine high-level activities by means of heart-rate processing. In this sense, certain activities (walking or running) could be discerned based on the effort exerted. However, the problem here is that ECG sensors are both expensive and uncomfortable for the user.

Thus, the present work is focused on the recognition of physical activities carried out by users by means of their mobile devices, and hence special attention must be paid to energy consumption and the computational cost of the methods used.

There are related studies where data for activity recognition is obtained through mobile devices where this data is sent to a server to process the information (Altun, Barshan, & Tunçel, 2010). In these cases, computational cost is no limitation and hence methods of a more complex nature can be used. In contrast, efficiency is a crucial issue when processing is carried out within the mobile device itself (Fuentes, Gonzalez-Abril, Angulo, & Ortega, 2012; Reddy et al., 2010).

Taking previous works into account, physical activity monitoring through smartphones presents the following challenges:

- To decrease, as far as possible, the risk of forgetting the processing device so that continuous monitoring can be performed anywhere and any time.
- To reduce the drain of energy on the smartphone, by developing an accurate and efficient system.
- To integrate learning and monitoring on the device itself, in realtime and without sharing server information.

In order to reduce the cost associated to accelerometer and gyroscope signal analysis, this paper opts for an innovative approach based on a discretization method that uses only accelerometer sensors since with these good results can be achieved and a lot energy can be saved (the more sensors used, the greater the consumption). Furthermore, the data is processed in the mobile itself, and therefore the efficiency is better than if data were sent to a server, and the result is obtained in real time. Thanks to this discretization process, the classification cost is much lower than it would be with continuous variables, and therefore the life of the battery is longer. It is therefore possible not only to eliminate the correlation between variables during the recognition process, but also to minimize the energy consumption of the process.

The remainder of the paper is organized as follows: Section 2 describes the method of data capture through the mobile device using an accelerometer sensor, and outlines all the physical activities that can be recognized by the system described. In Section 3, a new process, that discretizes continuous variables and provides classification, is presented. Section 4 shows a comparison between the new method and other methods used previously in the literature. Finally, in Section 5, a discussion is given on the various advantages of the proposed algorithm as well as on certain challenges and tasks that are currently being developed for the described recognition system.

2. Activity recognition

2.1. Embedded sensors and battery impact

Throughout this work, a KR3DM triaxial accelerometer integrated into a Google Nexus S is used. This sensor has a range of sampling frequencies between 25 Hz and 1500 Hz. Specific frequency is defined by each piece of software by using Android primitives. At these sampling rates, the device used for the testing process is configured to operate at 50 Hz in order to prevent excessive use of data and to reduce the computational cost. Therefore, based on the Nyquist–Shannon theorem, it can be ensured that signals with significant energy components below 25 Hz are liaison free. Depending on where the user takes the device, vibrations and any other kind of noise could be present with frequency components over 25 Hz, but these are of no interest in this work. However, the human-activity frequency range is much lower than the sampling band chosen. By using accelerometers taped to the body while running, Bhattacharya, McCutcheon, Shvartz, and Greenleaf (1980) found the main frequency components between 1–18 Hz at the ankle. As will be seen later, our work proposes placing the device at the hip, where acceleration forces are lower than at the ankle, and hence frequencies at this position are also lower.

On the other hand, a lower frequency allows the computation cost to be reduced thanks to the fact that each feature is obtained from accelerometry data. Furthermore, by reducing this processing time, the system becomes faster, more efficient and consumes less energy. Indeed, for contextual systems, with their intensive use of sensors, the high-energy consumption required must be taken into account not only in obtaining the data but also in its processing. A typical smartphone from the latest generation has a multitude of sensors that are commonly used, such as GPS (Morillo, Ramirez, Garcia, & Gonzalez-Abril, 2012), NFC, and a microphone. This means that, as result of high energy consumption, the useful time between device charges remains very low.

By applying Moore's law, it can be observed that manufacturers increase their processing power at least twice each year, in contrast to battery development, which has failed to double over the last five years. Battery life is not a secondary consideration, since, according to a survey performed by North American Technologies (Forrester, 2011), it stands as the second most important purchase decision factor for buyers of smartphones. Users acceptance, in the context of aware applications in general and of activity recognition systems in particular, is therefore critical. For this reason, in this work, not only has an accurate and fast system been developed, but a low energy consumption model is also presented from the viewpoint of discrete techniques.

There are various solutions using specific hardware (Choudhury et al., 2008) that have a high degree of autonomy. However, the problems faced by these elements are, as outlined earlier, the risk of losing and/or forgetting the device and the discomfort for users. Furthermore, these solutions are usually very expensive. In recent years, a large number of personal devices able to monitor the activity level have been developed by large companies, such as Adidas (Fig. 1), and Nike (Fig. 2). However, the aims of this work, even though related, are quite different. Both of these commercial devices allow the physical activity level to the detected. This is interesting from the point of view of calories burned. The limit imposed on the number of activities constitutes the main disadvantage of these systems, since they can only detect certain parameters, such as the number of sprints or total distance covered. Although these features are significant, for many users the number of these activities may be insufficient, for example, for users involved in a physical rehabilitation process which must be monitored by doctors. In these cases, activity recognition should have greater granularity to detect activities. The activities our system is able to discern will be discussed in the following section.

In short, although our proposal has been tested with smartphone embedded sensors, no limitation whatsoever is envisaged. In this paper, a new independent-sensor technique for activity recognition is presented, which uses and continues the work in Soria Morillo, Ortega Ramirez, and Gonzalez-Abril (2012). Thus, this

Fig. 1. Adidas MyCoach: intalled into Adidas shoes.

Fig. 2. Nike FuelBand: wristband with accelerometer sensor for activity level.

technique can be developed on a smartphone, a commercial wristband or any other appliance which contains an accelerometer and processing unit. In contrast to other proposals, our work is focused on battery consumption, which becomes even more critical on external devices where drains on the battery must be reduced to provide greater usage time.

There are several related studies which centre on the accuracy power trade off, for example in the proposals by Zappi et al. (2008) and Wang et al. (2009). Not only do these proposals operate with smartphones, but also with distributed body sensors, where it is possible to determine which sensors are the most important, and which can be disconnected in order to minimize energy cost.

2.2. Set of activities

In this work, far from being a static system, the number and type of activities recognized depends on the user; users can carry out hot-training on the system, which may involve adding new activities, and hence other activities can be detected. This is crucial for the accomplishment of a highly customizable environment where the users themselves can determine which activities are important and which remain irrelevant.

For a large numbers of users, it could prove more practical to recognize only a few activities, such as walking, sitting, and falling. For other users, however, activities, such as driving and cycling could be more relevant. In a rehabilitation process, by setting out an exercise programme, doctors and relatives could ascertain whether the user is carrying out the prescribed physical exercise correctly.

During the system testing and by performing comparative analysis with other platforms, 8 activities have been taken into account. These activities include standing, walking, running, jumping, cycling, driving, climbing stairs, and descending stairs. Moreover, thanks to this proposal, activities that have not been previously learned by the system can be determined while users carry them out. This learning process is achieved based on the analysis of the probability associated to each pattern while the user is performing each activity. In this sense, the system alerts the user when a non-trained activity is detected and therefore, the opportunity to carry out in-depth training for this activity is presented.

Obviously, the number of activities to be detected has an impact on the accuracy of the system, especially if acceleration patterns between activities are very similar. This aspect will be discussed later.

2.3. Data collection

Several related studies attain results of activity recognition off-line. First, an appropriate set of acceleration readings must be obtained from sensors. Once this set is completed, it must be sent to the server and the readings are then classified into any of the recognized activities. In Duong, Bui, Phung, and Venkatesh (2005) and other model-driven studies into activity recognition, learning is conducted off-line. This means that the user must be connected to the training server so that the model, from which the recognition process can be carried out, can be obtained.

However, training and recognition sets are obtained using time windows of fixed duration. Each time window is composed of a set of accelerometer readings from which it is possible to calculate a variety of features. After having conducted a performance and system accuracy analysis, it has been determined that the optimum length for these windows is 5 s. This time has been chosen due to the importance of ensuring that, in each time window, there is at least one activity cycle. An activity cycle is defined as an complete execution of an activity pattern. For example, two steps are an activity cycle for walking and one pedal stroke is the activity cycle for cycling. If a complete activity cycle is not presented in each time window, then, based on acceleration patterns, it will not be possible to determine the activity performed. The segmentation process and activity cycle are shown in Fig. 3.

Based on these time windows, which contain data for each accelerometer axis and reduce the computational cost of the new solution, a signal module has been chosen. This eliminates the problem caused by the device rotation (He & Jin, 2009). Furthermore, it increases user comfort by removing the restriction of maintaining the same orientation during the learning and recognition process.

For each data in a time window size $N, a_i = (a_i^x, a_i^y, a_i^z)$, $i = 1, 2, \ldots, N$ where x, y and z represent the three accelerometer axes, the accelerometer module is defined as follows:

$$|a_i| = \sqrt{\left(a_i^x\right)^2 + \left(a_i^y\right)^2 + \left(a_i^z\right)^2}$$

Hence, the following statistics are obtained for each time window.

- Mean: $\bar{a} = \frac{1}{N}\sum_{i=1}^{N}|a_i|$.
- Minimum: $a_{min} = \min\{|a_1|, |a_2|, \ldots, |a_N|\}$.

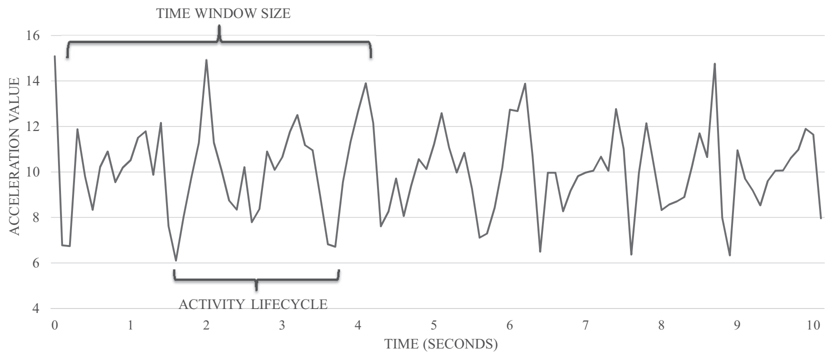


Fig. 3. Time windows split method over accelerometer signal.

- Maximum: $a_{max} = \max\{|a_1|, |a_2|, \ldots, |a_N|\}$.
- Median: $Me = |a_{(N+1)/2}|$ if $N\%2 \neq 0$ and $\frac{|a_{N/2}|+|a_{(N+1)/2}|}{2}$ otherwise.
- Standard deviation: $S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(|a_i| - \bar{a})^2}$.
- Signal magnitude area (Mathie et al., 2004): $SMA = \sum_{i=1}^{N}(|a_i^x| + |a_i^y| + |a_i^z|)$.
- Mean deviation: $D_m = \frac{1}{N}\sum_{i=1}^{N}||a_i| - \bar{a}|$.

In addition to the above variables, hereinafter called temporary variables, a new set of statistics called frequency-domain features are generated from the frequency domain of the problem. In order to obtain these frequency-domain features, the Fast Fourier Transform (FFT) is applied for each time window. Frequency components associated to each time window are defined as follows:

$$y_k = \sum_{j=0}^{F-1}|a_j|e^{-\frac{2\pi i j}{F}k}$$

where F is the cardinality of frequency components and $k = 1, 2, \ldots, F$.

From here, the following attributes are identified:

- Min module: $m_{min} = \min\{|y_1|, |y_2|, \ldots, |y_F|\}$.
- Max module: $m_{max} = \max\{|y_1|, |y_2|, \ldots, |y_F|\}$.
- Min frequency module: $f_{min} = \{f(y_k)|dle||y_k| = m_{min}\}$.
- Max frequency module: $f_{max} = \{f(y_k)|dle||y_k| = m_{max}\}$.

where $|y_k|$ and $f(y_k)$ are the module and frequency associated to the kth component, respectively.

It cannot be forgotten that the entire process is to be executed on the smartphone itself. Accordingly, the number of features must be reduced to minimize the impact of the processing energy used. To reduce the number of features set out above, the PCA method has been applied. In this way, a final set of 48 features was selected. It should be borne in mind that the first ten components deliver more than 98% of the variance in all validations. Therefore, in order to reduce the complexity, henceforth just 10 components will be used to describe the system.

3. Methodology

To the best of our knowledge, our proposal, which involves working in the domain of discrete variables to perform learning and recognition of activities, constitutes a totally innovative approach. This concept arose largely due to the need for a reduction in the high computational cost required for learning algorithms based on continuous variables, which have been used for this purpose over the years.

In Gonzalez-Abril, Velasco, Ortega, and Cuberos (2009), a labelling process, similar to a discretization process, is used to obtain a Qualitative Similarity Index (QSI), and hence it can be said that a transformation of the continuous domain to the discrete domain of values of the variables is beneficial in certain aspects. Therefore, the Ameva discretization (Gonzalez-Abril, Cuberos, Velasco, & Ortega, 2009) is used, which is unsupervised and fast. The most notable feature of these two processes is the small number of intervals generated, which facilitates and reduces the computational cost of the recognition process.

Let us briefly examine these algorithms.

3.1. Ameva algorithm

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a data set of an attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of the ℓ classes whose class variable is denoted by

$$\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \ell \geqslant 2$$

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \to \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in X$ in the domain of the property that is being discretized. Let us consider a discretization $\mathcal{D}$ which discretizes $\mathcal{X}$ into k discrete intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \ldots, L_k\}$$

where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j], j = 2, 3, \ldots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$. The discretization variable $\mathcal{L}(k)$ of $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view.

The main aim of the Ameva method (Gonzalez-Abril et al., 2009) is to maximize the dependency relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k. To this end, the following statistic is used:

$$\text{Ameva}(k) = \frac{\chi^2(k)}{k(\ell - 1)} \quad \text{where } \chi^2(k) = N\left(-1 + \sum_{i=1}^{\ell}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}\right)$$

Here, n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j; $n_{i\cdot}$ is the total number of instances belonging to the class C_i; and $n_{\cdot j}$ is the total number of instances that belong to the interval L_j, for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$, which fulfill the following:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, \quad N = \sum_{i=1}^{\ell}\sum_{j=1}^{k} n_{ij}$$

3.2. Discretization process

For each statistic $S_p \in \{S_1, S_2, \ldots, S_m\}$, the discretization process is performed, and a matrix of order $k_p \times 2$ is obtained, where k_p is the number of class intervals and 2 denotes the $inf(L_i^p)$ and $sup(L_i^p)$ interval limits i of the statistic p. Hence, a three-dimensional matrix containing the statistics and the set of interval limits for each statistic is called the Discretization Matrix and is denoted by

$$\mathcal{W} = (w_{pij})$$

where $p = 1, 2, \ldots, m, i = 1, 2, \ldots, k_p$, and $j = 1, 2$ (see Fig. 4).

Therefore, the Discretization Matrix determines the interval in which each item of data belongs for the different statistical associated values, by carrying out a simple and fast discretization process.

3.2.1. Class integration

The aim in the next step of the algorithm is to provide a probability associated with the statistical data for each of the activities, based on previously generated intervals. For this purpose, the elements of the training set $x \in X$ are processed to associate the label of the specific activity in the training set. In addition, the value of each statistic is calculated based on the time window.

In order to perform this process, a Class Matrix, $\mathcal{V}$, is defined as a three-dimensional matrix that contains the number of item of data from the training set associated with an $\mathcal{L}$ interval in a $\mathcal{C}$ activity for each statistic $\mathcal{S}$ of the system. This matrix is defined as follows:

$$\mathcal{V} = (v_{pij})$$

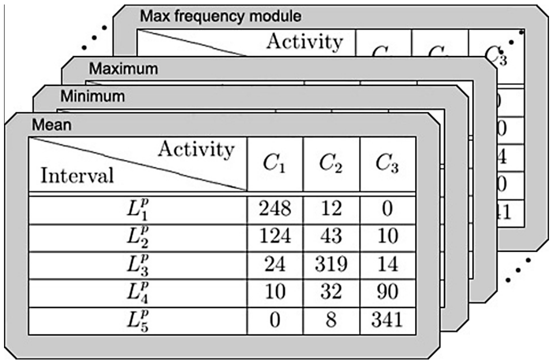
where $v_{pij} = \#\{x \in X | inf(L_i^p) < x \leqslant sup(L_i^p)\}, \mathcal{S} = S_p, \mathcal{C} = C_j, p = 1, 2, \ldots, m, i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$. Hence, each position in the Class Matrix is uniquely associated with a position in the Discretization Matrix, which is determined by its range.

Fig. 5 shows the contents of a real Class Matrix obtained during a learning process from all statistics. In order to simplify the figure for this example, only 3 out of 8 activities recognized by the system have been taken into consideration. Within these activities, five intervals determined by the Ameva algorithm can be observed in the Mean Class Matrix.

At this point, not only is it possible to determine the discretization interval, but the Class Matrix also helps to obtain the probability associated with the discretization process performed with the Ameva algorithm.

3.2.2. Activity-Interval Matrix

In the next step, a three-dimensional matrix, called the Activity-Interval Matrix is defined, denoted by $\mathcal{U}$, which specifies the likelihood that a given value x associated to a statistic S corresponds to



Interval \ Activity	C_1	C_2	C_3
L_1^p	248	12	0
L_2^p	124	43	10
L_3^p	24	319	14
L_4^p	10	32	90
L_5^p	0	8	341

Fig. 5. Class matrices.

an activity $\mathcal{C}$ in an interval $\mathcal{L}$. This ratio is based on the goodness of the Ameva discretization in order to determine the most probable activity from the data and the intervals generated for the training set.

Each value of $\mathcal{U}$ is defined as follows:

$$u_{pij} = \frac{v_{pij}}{v_{p\cdot j}} \frac{\sum_{q=1, q\neq j}^{\ell} \left(1 - \frac{v_{piq}}{v_{p\cdot q}}\right)}{\ell - 1}$$

where $v_{p\cdot j}$ is the total number of time windows of the training process labelled with the j activity for the p statistic, and $p = 1, 2, \ldots, m, i = 1, 2, \ldots, k_p$, and $j = 1, 2, \ldots, \ell$.

Given these values, $\mathcal{U}$ for the p statistic is defined as

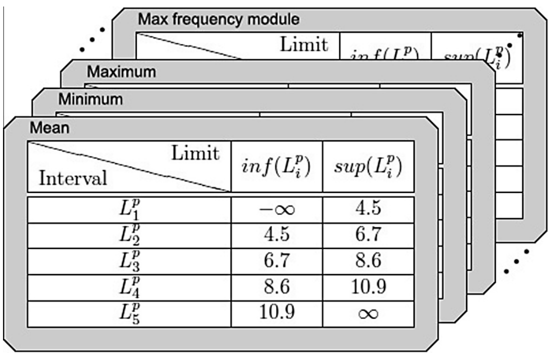
$$\mathcal{U}_p = \begin{pmatrix} u_{p00} & \ldots & u_{p0j} & \ldots & u_{p0\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pi0} & \ldots & u_{pij} & \ldots & u_{pi\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pk_p0} & \ldots & u_{pk_pj} & \ldots & u_{pk_p\ell} \end{pmatrix}$$

As can be seen in the definition of $\mathcal{U}$, the likelihood that data x is associated with the interval L_i corresponding to the activity C_j, depends not only on the data, but on all the elements associated with the interval L_i for the other activities.

Thus, each u_{pij} matrix position can be seen as a degree of belonging for a given x, identified with a C_j activity, to be included in the L_i interval of the S_p statistic.

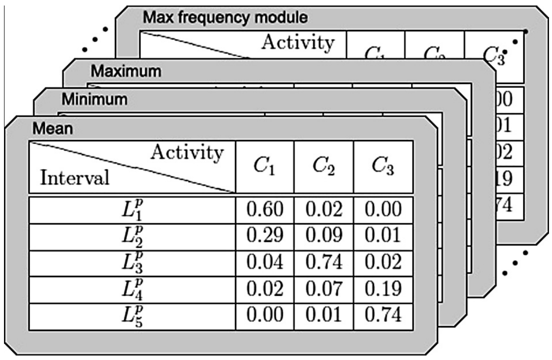
Similarly, the elements of $\mathcal{U}$ hold the following properties:

- $u_{pij} = 0 \iff v_{pij} = 0 \vee v_{piq} = v_{p\cdot q}, q \neq j$



Interval \ Limit	$inf(L_i^p)$	$sup(L_i^p)$
L_1^p	$-\infty$	4.5
L_2^p	4.5	6.7
L_3^p	6.7	8.6
L_4^p	8.6	10.9
L_5^p	10.9	∞

Fig. 4. Discretization matrices.



Interval \ Activity	C_1	C_2	C_3
L_1^p	0.60	0.02	0.00
L_2^p	0.29	0.09	0.01
L_3^p	0.04	0.74	0.02
L_4^p	0.02	0.07	0.19
L_5^p	0.00	0.01	0.74

Fig. 6. Activity-Interval matrices.

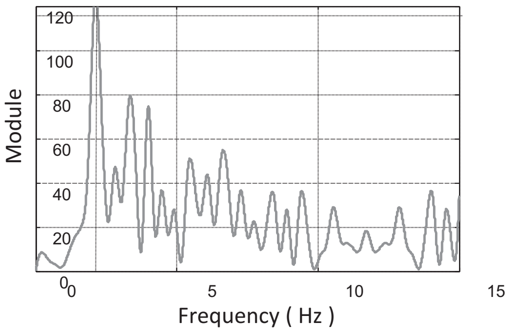


Fig. 7. Frequential study from accelerometer signal on a smartphone.

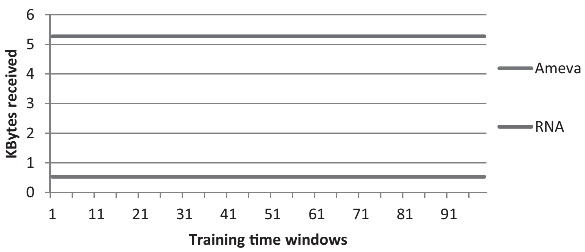


Fig. 8. Information flow between smartphone and server.

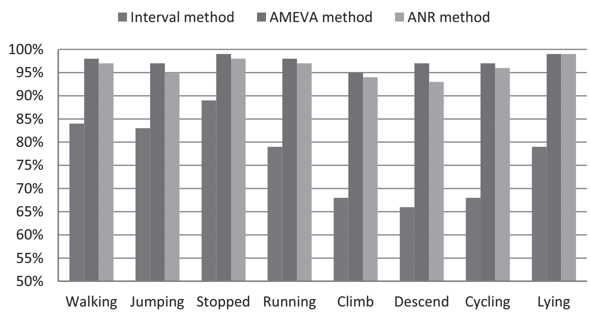


Fig. 9. Processing time of Ameva and neural network methods on the device.

- $u_{pij} = 1 \iff v_{pij} = v_{p\cdot j} = v_{pi\cdot}$

Fig. 6 shows a set of values for each of the positions of the Activity-Interval Matrix. These results have been obtained from the training set of the Class Matrix described above in Fig. 5.

3.3. Classification process

After obtaining the discretization intervals and the various probabilities of belonging, this section describes the process of classification from the data of the analysis time windows. This process is divided into two main parts. First, the way to perform the recognition of physical activity is described, and then the task of determining the frequency of a particular activity is presented.

3.3.1. Classifying data

For the classification process, the most likely activity is decided by a majority voting system. As described above, this process starts from the Activity-Interval Matrix and a set of data $x \in X$ for the set $\mathcal{S}$.

The process therefore consists of finding an activity $C_i \in \mathcal{C}$ that maximizes this likelihood. This criterion is collected in the following expression, denoted by *mpa* (most probable activity):

$$mpa(x) = C_k$$

where $k = arg(\max_j \sum_{p=1}^{m} u_{pij} | x \in (inf(L_i^p), sup(L_i^p))])$. The expression shows that the weight, contributed by each statistic to the probability calculation function, is the same. This expression assumes that all statistics provide the same information to the system and there is no correlation between them.

Thus, the *mpa* represents the activity whose data, obtained within the processing time window, is more suited to the set of values from $\mathcal{U}$. In this way, the proposed algorithm not only determines the *mpa*, but also determines its associated probability.

From this likelihood, certain activities that fail to adapt well to sets of generic classification can be identified, and constitute an indication that the user is carrying out new activities for which the system has not been previously trained.

3.3.2. Frequency activity approach

The proposed system not only determines the activity performed by the user, but the frequency at which it is performed. With this improvement in the activity recognition system, the information obtained can therefore be enriched with information specifying the number of pedal rotations, stairs, or steps that users perform per minute.

To achieve this degree of specification, a study of the accelerometry frequency component is made, whereby the maximum frequency module is used to determine the cadence of a range of activities. Therefore, the maximum frequency of the module could be useful in the analysis of the cadence of different activities, i.e. the max module, and at a second level, of the frequency associated with that value. These two values can be identified in Fig. 7 by the two grey dotted lines.

From a simple expression in conjunction with these values, the frequential value associated with the recognized activity in the current time window can be determined. It is denoted by *af*:

Table 1
Confusion Matrix.

Actual class	Predicted class																	
	Walk		Jump		Stop		Run		Climb		Descend		Cycle		Drive		Total	
	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA
Walk	1036	994	6	6	6	18	6	22	8	16	4	10	8	12	34	10	1108	1088
Jump	2	10	980	974	0	0	4	12	2	8	8	8	2	6	2	6	1000	1024
Immobile	0	0	0	0	1080	1112	0	0	0	0	6	0	8	0	24	8	1118	1120
Run	4	14	12	12	0	0	900	854	8	12	2	8	4	6	4	6	934	912
Climb	2	6	2	6	10	0	2	6	912	868	14	20	6	8	2	6	950	920
Descend	8	20	6	6	10	0	6	8	18	34	754	734	2	2	16	6	820	810
Cycle	2	6	2	4	0	0	2	14	4	6	2	4	844	840	6	2	862	876
Drive	4	8	2	2	36	12	2	6	4	12	2	8	4	4	794	838	848	890
Total	1058	1058	1010	1010	1142	1142	922	922	956	956	792	792	878	878	882	882	7640	7640

Table 2
Performance comparison by using measures of evaluation.

Activity	Measure									
	Accuracy		Recall		Specificity		Precision		F-measure (F_1)	
	Ameva (%)	RNA (%)	Ameva (%)	RNA (%)	Ameva (%)	RNA (%)	Ameva (%)	RNA (%)	Ameva (%)	RNA (%)
Walk	98.77	97.93	97.92	93.95	98.91	98.57	93.50	91.36	95.66	92.64
Jump	99.35	98.87	97.03	96.44	99.70	99.25	98.00	95.12	97.51	95.77
Immobile	98.69	99.50	94.57	97.37	99.42	99.88	96.60	99.29	95.58	98.32
Run	99.27	98.35	97.61	92.62	99.49	99.14	96.36	93.64	96.98	93.13
Climb	98.93	98.17	95.40	90.79	99.43	99.22	96.00	94.35	95.70	92.54
Descend	98.64	98.25	95.20	92.68	99.04	98.89	91.95	90.62	93.55	91.64
Cycle	99.32	99.03	96.13	95.67	99.73	99.47	97.91	95.89	97.01	95.78
Drive	98.14	98.74	90.02	95.01	99.20	99.23	93.63	94.16	91.79	94.58

$$af = Max\ frequency\ module \cdot 60s$$

thus obtaining the cadence of activity per minute. This value is then displayed to the user in the application developed to determine both the intensity of the activity carried out and its repetitions.

3.3.3. Pseudocode

A pseudocode that summarizes the entire procedure more clearly is presented.

```
/* Inputs */
x = x1, x2, ..., xm;
S = S1, S2, ..., Sm
/* Discretization and Class Integration process */
W = {}, V = {}
for each Si in S do
  Wi = Ameva_Discretization (Si)
  Vi = Class_Integration (Si)
  W = W + Wi
  V = V + Vi
end for
/* Activity-Interval process */
U = {}
for each Vi in V do
  Ui = Activity_Interval (Vi)
  U = U + Ui
end for
/* Classification process */
values = {}
for j = 1 to l do
  aux = 0
  for p = 1 to m do
    i = interval (x, p)
    aux = aux + Ui[p][i][j]
    end for
    values = values + aux
end for
/* Output */
return getClass (getIndex (values, max (values)))
```

4. Method analysis

Once the basis of the developed activity recognition algorithm was set out, an analysis of the new proposal was performed. To this end, the new development was compared with a widely used recognition system based on neural networks. In this case, both learning and recognition was performed by continuous methods.

The test process was conducted on a Google Nexus One for a group of 10 users. Notably, the activity habits of these users were radically different, since 5 users were under 30 years old while the rest were older than this age. For this test, a document was delivered to each user so that a description of the activity performed could be reported, together with its start time and end time.

In order to more accurately determine the real activities that the user is carrying out, a specific application on the device has been developed in which the user must enter the activity tracked. Those activities conducted during the test process that had not been previously trained, were dismissed in order to analyze the system accuracy.

Finally, the learning process of each activity recognized by the system consisted of its performance for a time of 6 min. As for the recognition process, users were followed over a period of 72 h.

From these values the real dataset is obtained,[1] which enables the accuracy and delay time, among other indicators, to be determined. However, it has been decided to apply this method to a public human-activity recognition dataset published by the Center for Machine Learning and Intelligent Systems (Anguita, Ghio, Oneto, Parra, & Reyes-Ortiz, 2012). This decision allows this work to be compared with other proposals through the same information repository.

First, after conducting the performance tests, it was found that, in the case of Ameva, the flow of information between the device and the server was much lower. This server is responsible for safeguarding the training data and recognized activities. As can be seen in Fig. 8, the traffic of data necessary was 4.7 KBytes for the system based on neural networks, and 0.6 in the case of Ameva. That is, the traffic of information was reduced by more than 70%. This constitutes an advantage in terms of a reduction in additional costs arising from excessive use of the data network.

Moreover, it is crucial to consider energy consumption and the processing cost of the system when working with the mobile device. In this case, after comparing the above methods, the conclusion is reached that the method based on Ameva reduces the computational cost of the system by about 50%, as can be seen in Fig. 9. The time needed to process a time window by using the Ameva-based method is 0.6 s. However, for methods based on neural networks, this was 1.2 s.

In addition to this benefit with the device, a survey from with a number of questions concerning the system impression was given to users at the end of the testing process. In general, the Ameva-based system scored much higher, due to the fluidity experienced by users while they were working with the device and while the activity recognition service was being executed. Furthermore, the response by the users indicated that the smartphone's temperature was considerably lower than when the neural network solution was employed.

Last but not least, some measures of the methods presented below are analyzed. Accuracy constitutes the most widely used metric for the measurement of the performance of learning systems. Nevertheless, it has been widely demonstrated that, when

[1] Available at http://madeirasic.us.es/idinfor/wp-content/uploads/2014/02/activities.csv.

the prior class probabilities differ greatly, these measures become inappropriate because they fail to consider misclassification costs, are strongly biased towards the majority class, and are sensitive to class skews (Daskalaki, Kopanas, & Avouris, 2006; Huang & Ling, 2005).

Finally, the definition of the measures used are:

- *Accuracy* is taken as the degree of veracity, while in certain contexts precision may mean the degree of reproducibility. It is defined here as $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.
- *Recall* is the fraction of relevant instances that are retrieved. It is defined as: $recall = \frac{TP}{TP+FN}$.
- *Specificity* measures the proportion of negatives which are correctly identified as such. It is defined as: $specificity = \frac{TN}{TN+FP}$.
- *Precision* is the fraction of retrieved instances that are relevant. It is defined as: $precision = \frac{TP}{TP+FP}$.
- *F-measure* (F_1) is a measure of the accuracy of a test. It can be interpreted as a weighted average of the precision and the recall, and is defined as: $F_1 = 2\frac{precision \cdot recall}{precision+recall}$.

Note that TP and TN denote the number of positive and negative cases correctly classified, while FP and FN refer to the number of misclassified positive and negative examples, respectively.

Based on these definitions, Tables 1 and 2 present the test. It was conducted on 8 different activities in order to unify the results for all users. In both tables, differences between the two methods, RNA and Ameva, can be observed. Most values presented for each measure and activity in Table 2 show that the Ameva method performs markedly better than does the RNA, especially as regards precision. That is to say, the number of false positives in the Ameva method is lower than that using the RNA method, as can be observed in Table 1.

Immobile and Drive are controversial activities due to their similar characteristics. Even under observation, it is difficult to differentiate between these two activities. For this reason, and due to the temporal nature of the Immobile activity, results from these two activities present a high level of disturbance in contrast to other activities.

5. Conclusions and future work

As mentioned in Section 1, the number of research studies into activity recognition has increased in numerous domains in recent years. The insightful practical implications include, for example in elderly care the estimation of the quality of self-care and the monitoring of activities of daily living. Furthermore, activity recognition enables obesity to be better fought and improvements to be made toward a more active life in proactive healthcare and can also obtain the correlation of activities with moods, mood swings, and manic depression in psychiatry.

In the workplace, it enables maintenance staff and crowds to be tracked; and accountability in security/workflow monitoring to be managed; activity information to be shared in groups; and diaries and auto-filling journals to be managed for later accounting in memory support.

Efficiency and accuracy are two elements that must be taken into account when any activity recognition system is implemented on a mobile device. In this work, a recognition system based on discrete variables is presented whereby the Ameva discretization algorithm and a new classification system, based on this algorithm, are used.

By using this process to increase the recognition frequency, if has been possible to obtain a physical activity reading every 5 s and to publish these readings in the user activity log (Alvarez-Garcia, Ortega, Gonzalez-Abril, & Velasco, 2010).

This classification algorithm is in a simple unified form for multi-class cases, and in general has equal performance or outperforms RNA in terms of accuracy, recall, specificity, precision, and F-measure (F_1). It is also very fast because it is based on the Ameva discretization algorithm and a majority voting system which both have a very low processing time.

Furthermore, although it has yet to be tested with other datasets, the core of this algorithm remains free of dependence on the features of any recognition activity dataset, and is therefore applicable to any dataset.

The strengths of the system are: the high success rate for which it has been possible to achieve an average accuracy of 98% for the recognition of 8 different types of activities; and, the reduced computational cost associated to the processing of data during the recognition process, thanks to the inclusion of discrete variables.

However, the main problem of this system based on statistical learning lies in the number of activities that can be recognized. Working solely with accelerometer sensors reduces the number of detected activities and there is occasionally a strong correlation between the variables, as can be observed with the Immobile and Drive activities in Table 2. Moreover, the position of the mobile device is a weakness of the developed system since it must be placed at the hip which it is not a natural position (sometimes a belt is necessary). Other positions have been tested, but have not achieved good results.

This system is currently focused on the detection of falls in elderly people, so that it can transmit an alarm signal to the family and/or medical centre. It complements existing telecare services, such as those as offered by the Andalusian Regional Ministry of Equality, Health and Social Policy.

To this end, a new system that can recognize basic activities, such as immobile, walk, run, climb and descend, no matter where the mobile device is located is currently under development. This is expected to provide a major improvement in that the user will not have to worry that the mobile device is placed in the right position for the activities to be detected correctly, and therefore the user can move in a more natural way.

Acknowledgement

This research is partially supported by the project Simon (TIC-8052) of the Andalusian Regional Ministry of Economy.

References

Altun, K., Barshan, B., & Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition, 43*(10), 3605–3620.

Alvarez-Garcia, J., Ortega, J., Gonzalez-Abril, L., & Velasco, F. (2010). Trip destination prediction based on past GPS log using a hidden markov model. *Expert Systems with Applications, 37*(12), 8166–8171<http://dx.doi.org/10.1016/j.eswa.2010.05.070> .

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. (2012). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Ambient assisted living and home care* (pp. 216–223).

Bhattacharya, A., McCutcheon, E., Shvartz, E., & Greenleaf, J. (1980). Body acceleration distribution and O2 uptake in humans during running and jumping. *Journal of Applied Physiology, 49*(5), 881–887.

Bicocchi, N., Mamei, M., & Zambonelli, F. (2010). Detecting activities from body-worn accelerometers via instance-based algorithms. *Pervasive and Mobile Computing, 6*(4), 482–495.

Brezmes, T., Gorricho, J., Cotrina, J. (2009). Activity recognition from accelerometer data on a mobile phone. *Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living* (pp. 796–799).

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., et al. (2008). The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, 7*(2), 32–41.

Daskalaki, S., Kopanas, I., & Avouris, N. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence, 20*(5), 381–417.

Dernbach, S., Das, B., Krishnan, N. C., Thomas, B. L., & Cook, D. J. (2012). Simple and complex activity recognition through smart phones. In *2012 8th International conference on intelligent environments (IE)* (pp. 214–221). IEEE.

Duong, T., Bui, H., Phung, D., & Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE computer society conference on computer vision and pattern recognition* (Vol. 1, pp. 838–845).

Forrester. (2011). Customer technology survey, Tech. rep., North American Technologies.

Fuentes, D., Gonzalez-Abril, L., Angulo, C., & Ortega, J. (2012). Online motion recognition using an accelerometer in a mobile device. *Expert Systems with Applications, 39*(3), 2461–2465.

Gonzalez-Abril, L., Cuberos, F., Velasco, F., & Ortega, J. (2009). Ameva: An autonomous discretization algorithm. *Expert Systems with Applications, 36*(3), 5327–5332.

Gonzalez-Abril, L., Velasco, F., Ortega, J., & Cuberos, F. (2009). A new approach to qualitative learning in time series. *Expert Systems with Applications, 36*(6), 9924–9927.

He, Z., & Jin, L. (2009). Activity recognition from acceleration data based on discrete cosine transform and SVM. In *IEEE international conference on systems, man and cybernetics* (pp. 5041–5044).

Hong, Y., Kim, I., Ahn, S., & Kim, H. (2008). Activity recognition using wearable sensors for elder care. In *Second international conference on future generation communication and networking* (Vol. 2, pp. 302–305).

Huang, J., & Ling, C. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17*(3), 299–310.

Kwapisz, J., Weiss, G., & Moore, S. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter, 12*(2), 74–82.

Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., & Zancanaro, M. (2010). What is happening now? Detection of activities of daily living from simple visual features. *Personal and Ubiquitous Computing, 14*(8), 749–766.

Li, M., Rozgic, V., Thatte, G., Lee, S., Emken, B., Annavaram, M., et al. (2010). Multimodal physical activity recognition by fusing temporal and cepstral information. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18*(4), 369–380.

Mathie, M., Coster, A., Lovell, N., Celler, B., Lord, S., & Tiedemann, A. (2004). A pilot study of long-term monitoring of human movements in the home using accelerometry. *Journal of Telemedicine and Telecare, 10*(3), 144–151.

Morillo, L. S., Ramirez, J. O., Garcia, J. A., & Gonzalez-Abril, L. (2012). Outdoor exit detection using combined techniques to increase GPS efficiency. *Expert Systems with Applications, 39*(15), 12260–12267<http://dx.doi.org/10.1016/j.eswa.2012.04.047> .

Paoli, R., Fernández-Luque, F., & Zapata, J. (2011). A system for ubiquitous fall monitoring at home via a wireless sensor network and a wearable mote. *Expert Systems with Applications, 39*(5), 5566–5575.

Pawar, T., Chaudhuri, S., & Duttagupta, S. P. (2007). Body movement activity recognition for ambulatory cardiac monitoring. *IEEE Transactions on Biomedical Engineering, 54*(5), 874–882.

Ravi, N., Dandekar, N., Mysore, P., & Littman, M. (2005). Activity recognition from accelerometer data. In *Proceedings of the national conference on artificial intelligence* (Vol. 20, p. 1541).

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks, 6*(2), 13.

Soria Morillo, L., Ortega Ramirez, J., Gonzalez-Abril, L. (2012). Aplicaciones contextuales en dispositivos móviles: Arquitectura para la mejora de la eficiencia energética, EAE. Spanish Academic Editorial.

Wang, Y., Lin, J., Annavaram, M., Jacobson, Q., Hong, J., Krishnamachari, B., Sadeh, N. (2009). A framework of energy-efficient mobile sensing for automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services* (pp. 179–192).

Ward, J. A., Lukowicz, P., Troster, G., & Starner, T. E. (2006). Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1553–1567.

Zappi, P., Lombriser, C., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., Tröster, G. (2008). Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In *Wireless sensor networks* (Vol. 4913, pp. 17–33).

CHAPTER 5

LOW ENERGY PHYSICAL ACTIVITY RECOGNITION SYSTEM ON SMARTPHONES

Overview

This paper addresses a low-energy solution for human activity recognition systems using a dynamic sampling rate method for energy reduction based on Ameva as a classification system. In this context, the broadened problem of activity recognition systems based on IMU sensors is the high power consumption caused by keeping the smartphone awake in order to perform all tasks needed, so getting a way to reduce the energy consumption and the processing cost of the system is the focus of this work.

The literature has explored three ways of solving this drawback: selecting the system features depending on the computational cost for their calculation, reducing the sampling rate below the threshold of 50 Hz and implementing a dynamic sampling rate method in the proposed solutions that is based on the demonstration that different activities exhibit differing levels of classification accuracy, depending on the on-body placement of the accelerometers. Through this comparison, the Dynamic

Ameva Classification System has been applied, implementing a dynamic sampling rate method for energy reduction applying a duty cycle optimization.

The accuracy of the system does not vary depending on the user who performs the test and this method has a very slight impact on system accuracy because its auto-reconfiguration makes it possible to increase the sample rate if necessary. As conclusion this strategy brings considerable benefit in terms of the energy savings achieved.

Context

This research topic was based on previous works of human activity recognition systems as a continuation in the same area of knowledge. In this case, the energy reduction was the main goal and the use of a dynamic sampling rate allowed achieve it. This paper is the result of over one year of work in this area and applies usage information (and data from other sources) from real experimentation performed for a previous paper titled *Discrete techniques applied to low-energy mobile human activity recognition. A new approach.*

Journal information

The Journal *Sensors* was selected for the submission of this paper that is indexed in JCR with an Impact Factor of 2.033 and a 5-Year I.F. 2.437. Also, it stands in different rankings in three categories: Instruments & Instrumentation (12/56, Q1); Chemistry, Analytical (36/75, Q2); Electrochemistry (16/27, Q3).

Figure 5.1: Sensors cover.

Sensors **2015**, *15*, 5163-5196; doi:10.3390/s150305163

OPEN ACCESS
sensors
ISSN 1424-8220
www.mdpi.com/journal/sensors

Article

Low Energy Physical Activity Recognition System on Smartphones

Luis Miguel Soria Morillo [1,*], Luis Gonzalez-Abril [2], Juan Antonio Ortega Ramirez [1] and Miguel Angel Alvarez de la Concepcion [1]

[1] Computer Languages and Systems Department, University of Seville, 41012 Seville, Spain; E-Mails: jortega@us.es (J.A.O.R.); maalvarez@us.es (M.A.A.C.)
[2] Applied Economics I Department, University of Seville, 41018 Seville, Spain; E-Mail: luisgon@us.es

* Author to whom correspondence should be addressed; E-Mail: lsoria@us.es; Tel.: +34-954-556-881.

Academic Editor: Leonhard M. Reindl

Received: 17 December 2014 / Accepted: 13 February 2015 / Published: 3 March 2015

Abstract: An innovative approach to physical activity recognition based on the use of discrete variables obtained from accelerometer sensors is presented. The system first performs a discretization process for each variable, which allows efficient recognition of activities performed by users using as little energy as possible. To this end, an innovative discretization and classification technique is presented based on the χ^2 distribution. Furthermore, the entire recognition process is executed on the smartphone, which determines not only the activity performed, but also the frequency at which it is carried out. These techniques and the new classification system presented reduce energy consumption caused by the activity monitoring system. The energy saved increases smartphone usage time to more than 27 h without recharging while maintaining accuracy.

Keywords: contextual information; mobile environment; discretization method; qualitative systems; smart-energy computing

---

1. Introduction

Just 30 min of moderate activity five days a week can improve your health, according to the Centers for Disease Control and Prevention. By enabling activity monitoring on an individual scale over an extended period of time in a ubiquitous way, physical and psychological health and fitness can be improved. Studies performed by certain health institutes [1–4] have shown significant associations between physical activity and reduced risk of incident coronary heart disease and coronary events. Their results can be seen in Figure 1, where the inverse correlation between the risk of cardiovascular incidents and physical activity level is shown through a comparison of four separate studies.

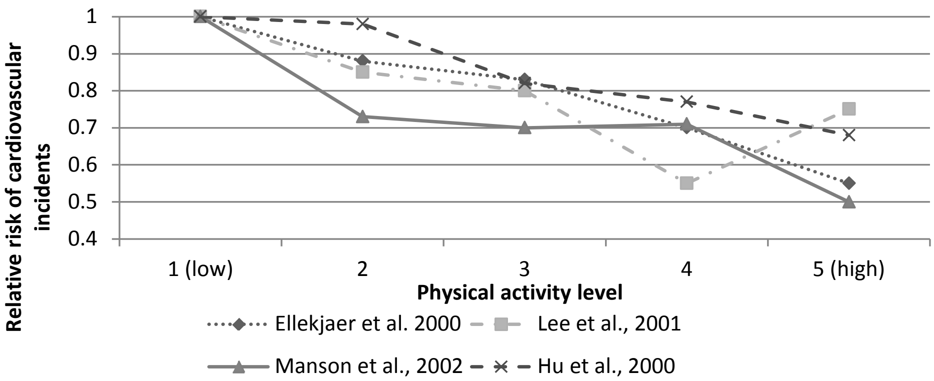


Figure 1. Associations between the risk of cardiovascular incidents and physical activity level.

In recent years, thanks largely to increased interest in monitoring certain sectors of the population, such as elderly people with dementia and people in rehabilitation, activity recognition systems have increased in both number and quality. Furthermore, communication between relatives, friends and professionals can be improved by means of graphs of weekly activity (highly relevant for sportsmen and relatives of elderly people), whereby the doctor can be automatically alerted if any strange activity is detected. In fact, automatic recognition of human activity represents one of the most important research areas in ubiquitous computing [5,6]. For this reason, it is extremely important to ensure that the intrusion level caused by the system is the lowest possible. Some recent works, such as [7–9], attempt to solve this problem by using a variety of sensors, such as accelerometers, gyroscopes, GPS and even radio-frequency identification and near-field communication (NFC) sensors.

As will be seen below, however, the use of these sensors causes a major drain on the energy of autonomous devices running on batteries, such as smartphones. On the other hand, by using data acquired from these sensors and applying certain classification methods, it is possible to perform pervasive physical activity monitoring. Some of these algorithms, such as Bayesian decision (BDM), decision tree algorithms (RBA), least-squares methods (LSM), support vector machines (SVM), K-nearest neighborhood (KNN) and artificial neural networks (ANN), will be analyzed and compared in this work. The results show the main differences between different studies, and certain drawbacks will be determined. These drawbacks, commonly related to energy consumption and computational cost, do not make possible their implementation on real users' smartphones. The first difference observed between the systems developed so far is the type of sensor used.

There are systems that use specific hardware [10–12], whereas others use general purpose hardware [7,13,14]. Obviously, the use of generic hardware, as smartphones, is a benefit to users, because the cost of such devices and their versatility are assets in their favor. The risk of loss, forgetting and disuse is decreased because users' smartphones have already been integrated into the users' daily life. However, general purpose devices are used for other purposes, such as making phone calls, surfing the Internet and listening to music. For this reason, the physical activity recognition system must be executed in background mode and should cause the least impact as possible on the system in terms of complexity and energy consumption.

Another difference found among related studies is the number and position of sensors used. In [15], the accelerometer sensor is placed in a glove, which the user must wear. This sensor can recognize a multitude of activities, depending on the movement of the hand. In contrast, some studies use diverse sensors all over the body to recognize these activities [16–18]. In recent years, as a result of technological progress, it has been possible to build sensors of reduced size that can be installed into the user's clothes [19] or attached to the user's body [20,21]. According to some comparative studies and research based on multiple sensors, this type of sensor gives higher accuracy, although [13] shows that it is more comfortable for the user when the sensor is placed into the user's pocket or at the hip. This increased user acceptance for devices attached at the hip or in the pocket is because the installation on a monitored person's body is easier, not to mention that the infrastructure is much more simple and inexpensive. Once the most comfortable alternative for users is determined, some device sensors can be chosen to perform the activity monitoring.

Some works, close to social computing, make use of microphones [22–26] and electrocardiogram (ECG) sensors [27–29] for this purpose. The former type, which consists of microphone and Bluetooth devices, helps to obtain contextual information about the user's environments and would be appropriate to perform a deeper analysis of the activity, for instance if the user is walking in a disco or at home, if the user is alone or with someone. However, high-level activity recognition (walking, playing, running or standing up) is done using other sensors. ECG can help in determining high-level activities by means of heart rate processing. In this sense, some activities (walking or running) could be discerned based on the effort needed to perform them. The problem here is that ECG sensors are expensive and uncomfortable for the user.

In other works [21], data for activity recognition are obtained through any kind of mobile device (not only mobile phones), although these data are sent to a server, where the information is subsequently processed. Thus, the computational cost is not a handicap, as learning and/or recognition are performed in the server and a more complex processing can be applied. In contrast, when processing is carried out in the mobile device itself [30], efficiency becomes a crucial issue. In this vein, in order to apply a solution based on distributed computing, the device must always be connected to a data network. This does not currently represent a major drawback, since most devices have this kind of connectivity, although there are still users (mostly elderly) whose devices have not been associated with a continuous data connection outside the range of WiFi networks. Finally, decrease the energy cost conflicts with the need to send the data collected in a continuous way between device and server. This means that current strategies of sensor batching (as will be seen hereafter) cannot be applied, and devices must be continually waking up from sleep mode. Furthermore, the intensive use of the data network has a deep impact on the energy use.

The work in [31] shows the increase in energy consumption when 3G and WiFi are used, and in [32], it can also be observed that approximately 44% of battery usage in smartphones occurs by the use of GSM (3G or 2G).

Taking into account previous works, physical activity monitoring through smartphones presents the following challenges:

- To decrease as far as possible the risk of forgetting the processing device, so as to carry out continuous monitoring of users, everywhere and anytime;
- To reduce the energy impact on the smartphone, developing an accurate and efficient system;
- To integrate learning and monitoring on the device itself, in real time and without server information sharing.

Along this work, all of these challenges are addressed, and certain solutions are offered to achieve the proposed objective: to build a complete, accurate and low energy consumption system for pervasive physical activity monitoring using sensors embedded on smartphones.

The remainder of this paper is organized as follows. Section 2 presents the need to reduce the energy consumed by physical activity monitoring systems when they are executed in a smartphone. Section 3 presents the feature extraction model, the dataset obtained from sensors and all physical activities that can be recognized by the described system. In Section 4, discretizing continuous variables and a classification process are presented. Section 5 compares the presented method and other methods used previously in the literature. Finally, Section 6 discusses the advantages of the proposed algorithm, as well as some challenges and future works about the recognition system described.

2. Justification for the Reduction of Energy Consumption

Applying Moore's law, manufacturers increase processing power at least twice each year, in contrast with battery development, which did not even double over the last five years. This is not secondary at all. From a survey performed by North American Technologies [33], battery life is the second most important purchase decision factor for smartphones. Users' acceptance of context-aware applications in general and of activity monitoring systems in particular is critical. For this reason, not only has an accurate and fast system been developed, but a low energy consumption model from the viewpoint of discrete techniques is also presented throughout this work.

To determine the building blocks that promote the least drain of energy, a comparison between the energy consumption of the most frequently-used smartphone sensors in the literature is made. This is critical for choosing the most efficient sensors to be used in the monitoring application. Although some of these sensors cannot be used separately to determine the physical activity performed, some of them could work together.

An application has been developed to measure the battery energy consumed by each sensor in a real environment. For this purpose, Samsung Galaxy S2, Nexus One, Samsung Galaxy S3, Nexus 5 and HTC Tatoo were used. The application was run for four weeks, with battery consumption calculated based on the activated sensors. To prevent problems arising from the use of concrete devices, the application was installed on 20 users' smartphones with different features. The eight most-used sensors (microphone, GPS, WiFi, accelerometer, NFC, Bluetooth, ECG connected by Bluetooth and gyroscope) from the literature in the field of activity recognition were analyzed. It must be taken into account that battery consumption for the microphone depends not only on the microphone sensor itself, but also on the sample rate and the buffer. The last one is useful to keep the audio codec, memory controller and DMA engines awake for the shortest possible time.

To avoid battery capacity and the energy expenditure of the smartphone without performing any action and with no user interaction, a useless energy cost trend line is generated. This line (generated in the absence of normal user usage) is useful as the baseline, regardless of the time and the power consumption of sensors over several kinds of smartphones with different features. Figure 2 shows the result of this comparison (The values are displayed in hours regarding a generic device. However, during the evaluation process, the lifetime deviation compared to the values of useless energy cost has been studied on different devices), with the time represented on the horizontal axis and the battery level at a specific instant of time appearing on the vertical axis. The procedure to represent the results was as follows. The useless energy cost is measured for all devices on which the study was conducted. Once this value is measured (associated with 100% of the battery lifetime), the rest of the battery time runs using different sensors (GPS energy cost, accelerometer energy cost, and so on) was obtained. By a simple ratio, the percentage of reduction in the lifetime of the battery relative to the baseline (useless energy cost) is calculated. Thus, an approximate percentage of the impact that sensors have on the battery lifetime was obtained. Finally, to illustrate the results, these percentages are reflected on a generic device where the unused battery time is about 56 h. It can be seen in the figure that the lowest power consumption is given by the microphone, followed by the accelerometer sensor. Therefore, from these results, it can be deduced that the use of GPS or Bluetooth does not constitute a good choice to develop an energy-efficient physical activity monitoring system, despite their having higher accuracy. In the case of Bluetooth, advances in this sensors have reduced the energy consumption, but this technology still suffers from serious problems when being used in the field of activities recognition. On the one hand, the infrastructure must be installed in each location where it will be used. Currently, there are just a few public Bluetooth access points. Furthermore, dynamic activities, such as walking, running or cycling, can hardly be recognized by Bluetooth, unless additional devices associated with these activities are installed on the objects (bike, skateboard, and so on). It must be noted that nowadays, the smartphone is the only device (together with certain wearables, such as smart-watches) carried continuously for most users. Therefore, the use of Bluetooth devices for activity recognition systems must force the use of these devices, which would not be suitable for user acceptance of AR (*Activity Recognition*) systems. Finally, the cost of infrastructure is also a determining factor. Bluetooth access point networks are more expensive than embedding all of the necessary technology in the smartphone itself.

Other research is based on the use of microphone and voice recognition to determine the context of the user [34], and it could be thought that the result is more energy efficient. However, voice recognition presents problems when the environment is noisy or the user is alone, so sometimes, it is not possible to obtain results from the audio signal classifier. Thus, in the cited work, this process was complemented with other methods based on inertial measurement units (IMUs). It must be noted that an IMU is a device that measures velocity, orientation and gravitational forces, using a combination of accelerometers, gyroscopes and magnetometers.

To reduce the cost related to process accelerometer signals, this paper opts for an innovative technique, through which, the work is performed in the field of discrete variables. Thanks to a discretization process, the classification cost is much lower than that obtained when working with continuous variables. Any dependence between variables during the recognition process is therefore eliminated, and energy consumption from the process itself is reduced.

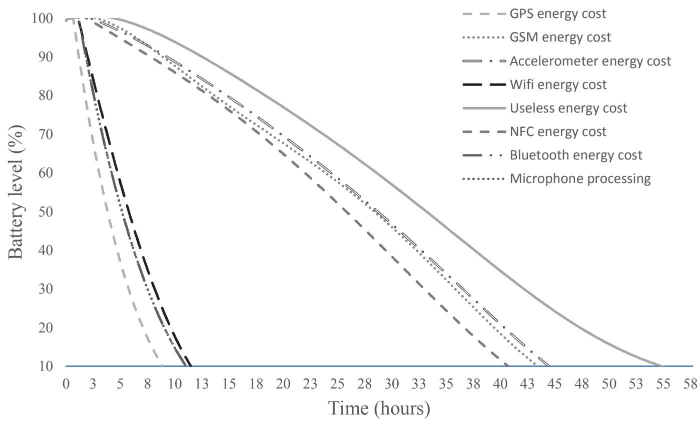


Figure 2. Energy cost comparison between sensors embedded in smartphones.

3. Building the Dataset

3.1. Embedded Sensor Limitations

Throughout this work, a triaxial accelerometer, BMA150 Bosh, integrated into a Google Nexus S, and other similar accelerometers integrated into a Samsung Galaxy S3 and a Nexus 5 were used. These sensors have a range of sampling frequencies between 25 Hz and 1500 Hz; however, human activities have a relatively low frequency, so it is not necessary to fully exploit the capabilities of the sensor. Activities, such as walking, running and jumping, can be determined with small blocks of data. This choice is also supported by other related works using similar devices [35]. However, when working with these elements, not only the high energy consumption required by the data collection, but also the processing of such data must be taken into account.

Mobile devices currently feature a multitude of sensors that are used routinely. This means that, as a result of high energy consumption, the useful time between device recharges is very low. Thus, usage is conditioned by dependence on the electric grid to recharge the mobile device. On the other hand, there are various solutions using specific hardware [36] that have a high degree of autonomy. The problems faced by these elements, however, include the aforementioned risk of loss and the risk of leaving the hardware behind, together with the discomfort for users. Furthermore, these solutions tend to be very expensive and are not oriented towards a wide range of applications.

3.2. Feature Extraction

Certain related studies attain results on activity recognition off-line. A comprehensive training set from the inertial sensor output is first needed before data can be classified into any of the recognized activities. For this purpose, both training and recognition sets are obtained using overlapped time windows of fixed duration. Following the conducting of a performance and system accuracy analysis, it is determined that the optimum length for these windows is 4 s with 1 s overlapping [37,38]. The recognition delay is determined by the following relation:

$$delay = size(v) \cdot 0.75 + k \cdot f(size(v)) \tag{1}$$

where $size(v)$ is the length of the time window, k is a computational constant that depends on the specific device and the f function represents the discretization and classification complexity, which depends on the length of the time window. The length of each time window has been chosen, because it is very important to ensure that each time window contains at least one activity cycle. Figure 3 shows the segmentation process and activity cycle.

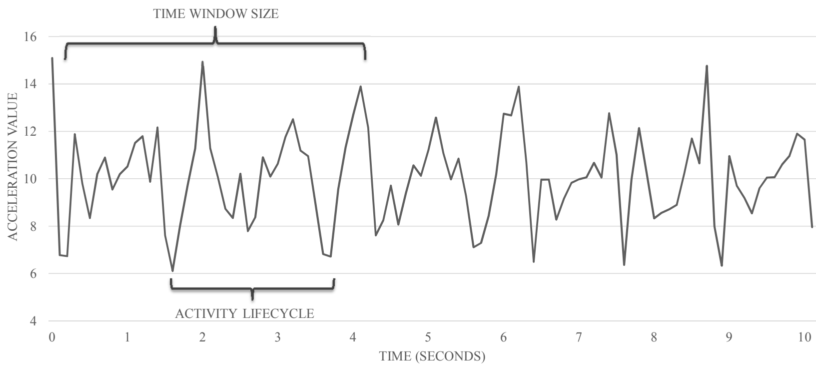


Figure 3. Relation between windows length and activity cycle.

As a premise for the use of time windows as a data collection method, it is assumed that the physical activities recognized by the system must have a duration greater than or equal to the size of this time window. Therefore, certain activities, such as a fall, may not be recognized by the system due to their exceptionality. Once data have been obtained, it is necessary to perform a filtering process before making the classification in order to remove all signal noise. In most cases, the noise of IMUs is negligible, although one kind of noise that can seriously affect the activity classification exists. This noise is produced by vibrations that take place on the device when it is carried by the user. A Butterworth low pass filter is applied to reduce the noise generated. Finally, following this strategy, a total of 561 temporal and frequential variables were derived from the inertial sensors on the devices.

3.3. Set of Activities

To improve and establish a comparative baseline for classification algorithms, which is easy and publicly replicable, it was decided to use two public datasets and one built for this study. The first dataset under study, named the Human Activity Recognition Using Smartphones Data Set, presented in UCI [39], is composed of 10,299 instances of time windows sampled in fixed-width sliding windows of 2.56 s and 50% overlap (128 readings/window) at a constant rate of 50 Hz, 561 variables and six activities labeled, carried out with a group of 30 volunteers with an age bracket of 19–48 years. The second public dataset, named the PAMAP2 Physical Activity Monitoring Data Set and published in UCI, as well [40], consists of 2,211,669 RAW data from three IMUs and 18 activities, performed with nine subjects between 26 and 31 years. RAW data were grouped into 8847 overlapped time windows, and system variables were obtained from them. The structure and content resulting from these datasets are very similar. Both are composed of a set of rows, and each row contains the variable values separated by a comma. Some features, such as the mean, standard deviation, median deviation, maximum, minimum, energy, interquartile range, signal entropy, correlation coefficient, skewness, kurtosis, angle between vectors and the Jerkmean, are processed in order to generate the complete dataset from RAW values of the accelerometer and gyroscope (this being understood as data obtained directly from IMUs embedded in the smartphone, *i.e.*, triaxial acceleration from the accelerometer and triaxial angular velocity from the gyroscope).

The users involved in the experiment followed a protocol in which the activities were performed wearing a waist-mounted smartphone. Each subject performed the protocol twice: in the first trial, the smartphone was fixed on the left side of the belt, and in the second, it was placed by the user himself as preferred. Through visual and sound signals, users were informed about the change of activity. The tasks were performed in laboratory conditions, but volunteers were asked to perform the sequence of activities freely for a more naturalistic dataset. In the laboratory, as users were performing each activity, a researcher was annotating them in a mobile application. The result of this annotation was a dataset for each instance (activity performed by each user) with start time, end time, activity, comments (interesting for further works) and user identification. Thanks to this information, data collected on the user smartphone from IMUs could be split and labeled.

Finally, to carry out a comparative analysis of the accuracy and performance of the discrete recognition method proposed in this paper, a new dataset was built. This file contains 6874 overlapped time windows with 170 variables associated with each one and eight activities supported. These activities are standing, walking, running, jumping, cycling, driving, upstairs and downstairs. A group of 10 users with Samsung Galaxy S3. Samsung Galaxy S4 and LG Nexus 5 smartphones participated in the experiment. Table 1 contains information about the users' smartphone distribution along the experiment. Each row shows the phone model used by the user and the number of time windows obtained throughout the whole experiment.

Table 1. Distribution of users' smartphones during the experiment.

User	Phone Model	Time Windows Collected
1	Samsung Galaxy S3	740
2	Samsung Galaxy S4	683
3	Samsung Galaxy S3	830
4	LG Nexus 5	716
5	Samsung Galaxy S3	519
6	LG Nexus 5	683
7	Samsung Galaxy S4	478
8	Samsung Galaxy S4	854
9	Samsung Galaxy S3	729
10	LG Nexus 5	642

However, far from being a static system, the kind of activities recognized depends on the user. In this line, thanks to the proposed method for new activity detection, introduced later, the system can determine when the users are carrying out activities that had not been learned before. As will be seen later, this is based on the analysis of pattern recognition and identification of low-probability instances.

4. Experimental Section

Working in the domain of discrete variables to perform learning and recognition of activities constitutes the innovative contribution offered by this work. Learning algorithms based on continuous variables, which traditionally have been used for this purpose over the years, lack a high complexity. A main aim in this paper is to use an approach based on discrete variables, which reduces this complexity, as will be shown. Therefore, prior to self-recognition and learning, it is necessary to carry out a process of discretization, which is performed through the application of the Ameva algorithm [41].

This algorithm has a number of advantages, chief among them being the small number of intervals generated, which facilitates and reduces the computational cost of the recognition process. It is worth noting that the Ameva algorithm has always been used as a discretization process [42–44]. In this paper, a new method that allows using the algorithm Ameva as a classification method has been developed.

Figure 4 shows the methodology for the system training. Along this section, each building block is presented, more specifically, the data processing and classification methods. Both make use of discrete techniques for classification, unlike other related works, which usually employ continuous methods [45,46].

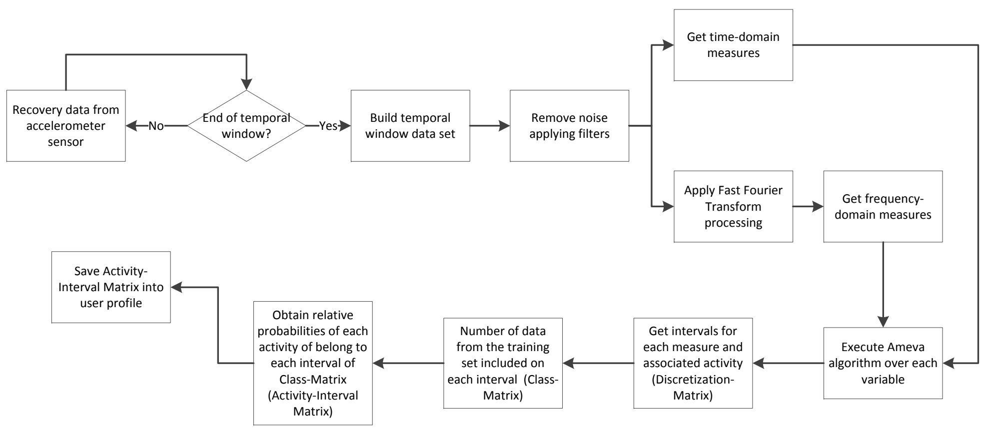


Figure 4. Methodology pipeline for the activity recognition training subsystem.

4.1. Ameva Algorithm

Working in the domain of discrete variables to perform the learning and recognition of activities is a new approach offered by this work. This decision was largely due to the high computational cost required for learning algorithms based on continuous variables that have been used for this purpose over the years.

In [43], a labeling process, like a discretization process, is used to obtain a similarity index, so it can be said that a transformation of the continuous domain to the discrete domain of the values of the variables is beneficial in certain aspects. However, before self-recognition or learning, it is necessary to carry out a process of Ameva discretization from its algorithm [41]. The most notable of these algorithms is the small number of intervals generated, which facilitates and reduces the computational cost of the recognition process.

Let us introduce these algorithms. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset of an attribute $\mathcal{X}$ of mixed-mode data, such that each example x_i belongs to only one of the ℓ classes of class variables denoted by $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}$, $\ell \geq 2$. Table 2 shows a toy dataset with 6 statistics, 10 samples and 3 classes.

Table 2. Example dataset with 6 statistics, 10 samples and 3 different classes.

Statistics						***Class***
Mean	**Std Deviation**	**Maximum**	**Minimum**	**Energy**	**Skewness**	
10.1	3.9	5.1	13.3	9.7	1.9	C_1
12.7	8.6	3.1	16.4	16.2	0.1	C_2
8.3	1.5	8.3	9.5	1.8	−2.4	C_3
11.3	4.1	6.3	14.9	11.2	1.1	C_1
8.6	1.2	8.7	9.1	1.2	1.3	C_3
9.8	2.7	6.5	13.2	9.7	1.7	C_1
14.7	9.2	3.6	15.3	17.1	−0.2	C_2
11.7	8.5	2.9	16.8	14.3	−1.7	C_2
10.6	3.6	5.1	13.8	11.2	0.8	C_1
9.2	0.7	8.9	9.7	0.9	−1.8	C_3

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \to \mathcal{C}$, which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of property that is being discretized. Let us consider a discretization $\mathcal{D}$ that discretizes $\mathcal{X}$ into k discrete intervals: $\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \ldots, L_k\}$, where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2, 3, \ldots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$, which verifies that, for all $x_i \in X$, a unique L_j exists, such that $x_i \in L_j$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$.

The main aim of the Ameva method [41] is to maximize the dependency relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$ and, at the same time, to minimize the number of discrete intervals k. For this, the following statistic is used:

$$Ameva(k) = \frac{\chi^2(k)}{k(\ell - 1)}$$

where:

$$\chi^2(k) = N\left(-1 + \sum_{i=1}^{\ell}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{\cdot i}n_{j\cdot}}\right)$$

and n_{ij} denotes the total number of continuous values belonging to the C_i class that are inside the interval L_j, $n_{i\cdot}$ is the total number of instances belonging to the class C_i and $n_{\cdot j}$ is the total number of instances that belong to the interval L_j, for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$, fulfilling the following:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, \quad N = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij}$$

Table 3 shows this interval grouping process over 12,480 instances and 6 intervals. The number of instances contained in each class for a given interval is shown in each row. Columns contain the number of instances inside each interval for a given class. The last row and the last column contain the sum of instances for each class and each interval, respectively.

Table 3. Number of values of each C_i class contained in each interval L_j.

Interval	Class			$n_{i\cdot}$
	C_1	C_2	C_3	
L_1	3213	65	1	3279
L_2	412	156	4	572
L_3	318	891	86	1295
L_4	136	2178	312	2626
L_5	49	710	813	1572
L_6	0	13	3,123	3136
$n_{\cdot j}$	4128	4013	4339	12,480 (N)

4.2. Discretization Process

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ be a set of m statistics. Hence, for each statistic $S_p \in \mathcal{S}$, the discretization process is performed, obtaining a matrix of order $k_p \times 2$, where k_p is the number of class intervals and 2 denotes the $inf(L_{p,i})$ and $sup(L_{p,i})$ interval limits i of the p statistic. This three-dimensional matrix containing the set of interval limits for each statistic is called the discretization matrix and is denoted by $\mathcal{W} = (w_{pij})$, where $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2$.

Therefore, the discretization matrix determines the interval at which each datum belongs to the different statistical associated values, carrying out a simple and fast discretization process. Table 4 shows an example of this process.

Appendix A shows the distribution for the first 21 statistics. Each distribution contains the intervals generated on the horizontal axis and the number of associated samples. In this figure, it can be noted that samples are not equally distributed, and the number of intervals is not the same for all statistics. These elements depend on the number of samples for each activity in the dataset and the value distribution.

Table 4. Example of a discretization matrix.

Interval	Limit	
	$inf(L_{p,i})$	$sup(L_{p,i})$
$L_{p,1}$	$-\infty$	−0.99
$L_{p,2}$	−0.99	−0.98
$L_{p,3}$	−0.98	−0.66
$L_{p,4}$	−0.66	−0.28
$L_{p,5}$	−0.28	0.02
$L_{p,6}$	0.02	$+\infty$

4.2.1. Class Integration

The aim in the next step of the algorithm is to provide a probability associated with the statistical data for each of the activities based on previously generated intervals. For this purpose, the elements of the training set $x \in \mathcal{X}$ are processed to associate the label of the concrete activity in the training set. In addition, the value of each statistic is calculated based on the time window.

To carry out the previous process, a class matrix, $\mathcal{V}$, is defined as a three-dimensional matrix that contains the number of data from the training set associated with an $\mathcal{L}$ interval in a $\mathcal{C}$ activity for each statistic $\mathcal{S}$ of the system. This matrix is defined as follows: $\mathcal{V} = (v_{pij})$, where $v_{pij} = |\{x \in \mathcal{X} \mid inf(L_{p,i}) < x \leq sup(L_{p,i})\}|$, and $\mathcal{S} = S_p, \mathcal{C} = C_j, p = 1, 2, \ldots, m, i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$.

Thus, each position in $\mathcal{V}$ is uniquely associated with a position in $\mathcal{W}$ determined by its associated interval. Table 5 shows the contents of a real class matrix obtained during a learning process from a standard deviation statistic (S_p). In this table, the six intervals previously calculated by the Ameva algorithm and stored in $\mathcal{W}$ can be observed.

Table 5. Example of class matrix $\mathcal{V}$ for 6 discretization intervals and 6 activities.

Interval	Activity					
	Walking	Upstairs	Downstairs	Sitting	Standing	Lying
$L_{p,1}$	0	0	0	440	524	124
$L_{p,2}$	0	0	0	367	351	388
$L_{p,3}$	3	0	0	349	362	734
$L_{p,4}$	690	375	24	1	0	17
$L_{p,5}$	394	534	226	0	0	3
$L_{p,6}$	17	57	637	0	0	0
Total	1104	966	887	1157	1237	1266

At this point, it is not only possible to determine the discretization interval, but the class matrix also helps to obtain the probability associated with the discretization process performed with the Ameva algorithm.

4.2.2. Activity-Interval Matrix

Now, a matrix of relative probabilities is obtained. This three-dimensional matrix, called the activity-interval matrix and denoted by $\mathcal{U}$, determines the likelihood that a given value x associated with an S statistic corresponds to $\mathcal{C}$ activity in a $\mathcal{L}$ interval. This ratio is based on obtaining the goodness of the Ameva discretization, and the aim is to determine the most probable activity from the data and the intervals generated for the training set.

Each value of $\mathcal{U}$ is defined as follows:

$$u_{pij} = \frac{v_{pij}}{v_{p\cdot j}} \frac{1}{\ell - 1} \sum_{q=1, q\neq j}^{\ell} \left(1 - \frac{v_{piq}}{v_{p\cdot q}}\right)$$

where $v_{p\cdot j}$ is the total number of time windows of the training process labeled with the j activity for the p statistic, and $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$

Given these values, $\mathcal{U}$ for the p statistic is defined as:

$$\mathcal{U}_p = \begin{pmatrix} u_{p00} & \cdots & u_{p0j} & \cdots & u_{p0\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pi0} & \cdots & u_{pij} & \cdots & u_{pi\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pk_p0} & \cdots & u_{pk_pj} & \cdots & u_{pk_p\ell} \end{pmatrix}$$

Then, the following condition must be considered in order for the above definition to be complete and without errors in the training: $v_{p\cdot q} = 0 \rightarrow \frac{v_{piq}}{v_{p\cdot q}} = 0$.

As can be seen in the definition of $\mathcal{U}$, the probability that a datum x is associated with the interval L_i corresponding to the activity C_j depends not only on the data, but on all of the elements associated with the interval L_i for the other activities.

Thus, each u_{pij} matrix position can be considered as the probability that a given x belongs to C_j activity, that it is included in the L_i interval of the S_p statistic.

Similarly, the elements of $\mathcal{U}$ have the following properties:

- $u_{pij} = 0 \iff v_{pij} = 0 \vee v_{piq} = v_{p\cdot q}, q \neq j$
- $u_{pij} = 1 \iff v_{pij} = v_{p\cdot j} = v_{pi\cdot}$

Table 6 shows a set of different values obtained for each of the positions of the activity-interval matrix $\mathcal{U}$. These results were obtained from the training set from the class matrix described in Table 5.

Table 6. Activity-interval matrix $\mathcal{U}$ for the standard deviation with 6 discretization intervals and 6 activities.

Interval	Activity					
	Walking	**Upstairs**	**Downstairs**	**Sitting**	**Standing**	**Lying**
$L_{p,1}$	0.00	0.00	0.00	0.42	0.48	0.10
$L_{p,2}$	0.00	0.00	0.00	0.35	0.31	0.34
$L_{p,3}$	0.00	0.00	0.00	0.25	0.24	0.51
$L_{p,4}$	0.61	0.36	0.02	0.00	0.00	0.01
$L_{p,5}$	0.30	0.49	0.21	0.00	0.00	0.00
$L_{p,6}$	0.02	0.07	0.92	0.00	0.00	0.00

4.3. Classification Process

This section presents the process of classification from the data of the time windows analysis. This process is divided into two main parts. First, the way to perform the recognition of physical activity is described. Later, the task to determine the frequency of a particular activity is exposed.

4.3.1. Classifying Data

For the classification process, the more likely activity is decided by a majority voting system from the activity-interval matrix and a set of data $x \in X$ for the $\mathcal{S}$ set.

Therefore, it consists of finding an activity $C_i \in \mathcal{C}$ that maximizes the likelihood. The above criterion is collected in the following expression, denoted by mpa, $mpa(x) = C_k$, where, $k = \arg\max_j \left(\sum_{p=1}^{m} u_{pij};\, inf(L_{p,i}) < x \leq L_{p,i} \right)$. The expression shows that the weight contributed by each statistic to the likely calculation function is the same. This can be done under the assumption that all statistics provide the same information to the system, and there is no correlation between them. Thus, the most likely activity represents the activity whose data, obtained through the processing time window, are more suited to the value set from the activity-interval matrix.

In this way, the proposed algorithm not only determines the mpa, but its associated probability. Figure 5 shows the total interval probability based on a training set and 100 features (Downstairs and upstairs activities make reference to the dynamic activities of ascending and descending stairs). The peaks presented in this chart correspond to the features giving more information to the system. From this likelihood, certain activities that do not adapt well to sets of generic classification can be identified. It is an indication that the user is carrying out new activities for which the system has not been trained previously.

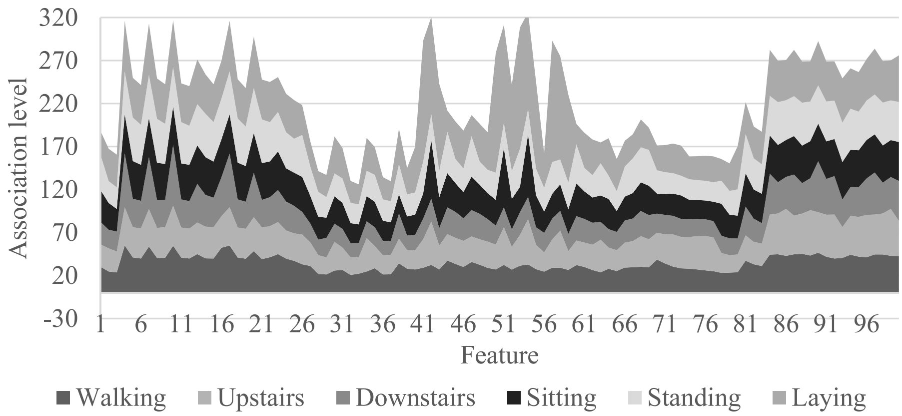


Figure 5. Association level of each activity for the 100 first features.

Figure 6 shows the process flow described in this section for process recognition from the activity-interval matrix calculated in the previous section.

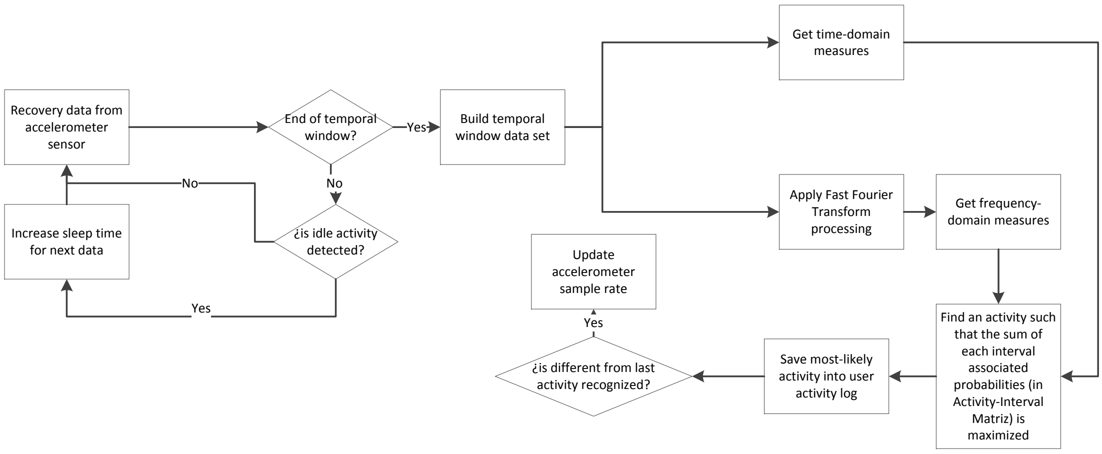


Figure 6. Flow diagram for the recognition process.

4.3.2. Activity Cycle Cadence Approach

The cadence at which any activity is performed is really important in order to get the calories burned and the intensity of the recognized activity. For this reason, although this factor does not make contributions to the field of activity recognition systems, it is crucial to get a system with added value applicable in a real life context. To determine the frequency for each activity cycle, the maximum module of the time window frequencies must first be calculated. Subsequently, the frequency associated with the maximum module (τ) represents the number of cycles per second for the activity related to the time window processed. From this and by using a simple expression, from the frequency value associated with the activity recognized in the current time window, the frequency of the activities per minute ($z(X)$) can be obtained as follows:

$$z(X) = \tau \cdot 60\,sec$$

4.4. Dynamic Sample Rate and Duty Cycle

It is crucial to consider energy consumption and the processing cost of the system when it is working on a mobile device. The broadened problem of activity recognition systems based on IMU sensors is the high power consumption caused by keeping the smartphone awake in order to perform all tasks needed, such as sensor sampling and activity classification. The literature has explored three ways of solving this drawback: selecting the system features depending on the computational cost for their calculation [47,48], reducing the sampling rate below the threshold of 50 Hz [49] and implementing a dynamic sampling rate method in the proposed solutions [35,50].

Dynamic sampling rate approaches, the most extensive over the last few years, is based on the demonstration that different activities exhibit differing levels of classification accuracy, depending on the on-body placement of the accelerometers [51]. Through this comparison, the Dynamic Ameva Classification System has been applied, implementing a dynamic sampling rate method for energy reduction. The frequency varies from 32 Hz for sitting, standing and lying to 50 Hz for walking, upstairs and downstairs. It must be taken into account that these frequencies were obtained experimentally by studying the pattern of the different activities from the point of view of the accelerometery. This study was conducted covering over 600,000 different time windows of activities obtained from 10 different users.

To respect the heterogeneity of the target users of the proposed recognition system, users with different profiles were selected, from those 21 years of age with an athletic profile, to seniors 82 years of age with a sedentary profile. Thanks to this information, it is possible to determine a suitable accelerometer frequency spectrum according to each activity. It should be noted that this frequency will have not only a power impact due to the decrease of data obtained from the accelerometer, but will also cause a reduction of the execution time of the algorithm, due to the smaller size of the time windows.

Algorithm 1 calculates the sampling rate corresponding to a recognized activity at a given time. For the calculation, the current recognized activity and previous detected activity are taken into account. If the recognition process enters into a stable phase, *i.e.*, there is continuity in the last set of recognized activities, the algorithm proceeds to update the sample rate. This update is calculated from the sampling rate associated with the activity detected, which is obtained from the map $SampleRateList$. As mentioned before, the specific values for the list $SampleRateList$ are calculated experimentally for each of the activities recognized. Later, if stabilization occurs for a prolonged period, our proposal proceeds to the regular updating of the sample rate based on the base $\log_2$ of the number of memories used.

Algorithm 1 Dynamic sampling rate algorithm.

$SampleRateList \leftarrow$ InitSampleRate($Frequencies$)
$MaxMemoryList \leftarrow$ InitMemoryList($Memories$)
$AmevaIsRunning \leftarrow$ true
$Count \leftarrow 0$
$WinSize \leftarrow 5$
$WinSamples \leftarrow 50 * WinSize$;
$A_{previous} \leftarrow$ GetAmevaActivity($Statistics$)
while $AmevaIsRunning$ **do**
  $A_{last} \leftarrow$ GetAmevaActivity($Statistics$)
  if $A_{last} == A_{previous}$ **then**
    $Count \leftarrow ActivityMemory + 1$
  else
    $Count \leftarrow 0$
  end if
  if IsCriticalActivity(A_{last}) **and**
  $Count >$ GetMaxMemory(A_{last}) **then**
    $NewFrequency \leftarrow SampleRateList(A_{last})$
    $WinSamples \leftarrow NewFrequency * WinSize$
  end if
  if $Count\%30 == MaxMemoryList(A_{last})$ **then**
    $WinSize \leftarrow WinSize + \log_2(Count\%30)$
    $WinSamples \leftarrow NewFrequency * WinSize$
  end if
  $A_{previous} \leftarrow A_{last}$
end while

The maximum memory limit is defined as the specific period for which an activity is considered stable. This maximum memory is fixed experimentally and depends on the specific activity. In [52] can be seen a related work, in which the sample rate is calculated from the estimated power required for each sample rate. The main problem of this work is that it does not take into account the specific activity to update the sample rate. Thus, errors may occur due to the low sample rate reached in the case of activity stabilization for a long period. Such excessively low sampling frequencies cause a decrease in the accuracy of the system, as can be seen in [52]. In [53], another system of dynamic frequency is proposed. In this case, the sample rate is only addressed from the activity being performed. This means that at the same time that an activity is recognized, the algorithm immediately proceeds to update the sample rate.

This causes two problems. On the one hand, it could be a case of sporadic activities (e.g., a fall) directly affecting the frequency of sampling and having a negative impact on the next set of activities recognized. On the other hand, by failing to update the sample rate periodically, the ability to minimize power consumption when a long-term stabilization occurs (e.g., when the user goes to sleep) is reduced.

In the proposed algorithm, the size of the time windows is increased, thus decreasing the number of times for carrying out the classification process and, therefore, the consumption caused by the main CPU to perform this action. It should be noted that increasing the size of the temporal window has a direct impact on power consumption, especially in those devices having a coprocessor for context purposes [54]. This CPU, separate from the main CPU, allows autonomous acquisition of contextual data, usually from the accelerometer or gyroscope, without activating the main processor. This produces a decrease in power consumption while these data are collected in the form of an asynchronous batch operation. Thus, by increasing the size of the time window, maximizing the use of the contextual coprocessor and delaying the use of the main CPU for the implementation of the classification algorithm, an increase in the lifetime of the battery is achieved while the recognition system is being used. Keep in mind that this improvement in the use of the coprocessor has not been applied in the comparison with other methods, because it is understood that it may be applied to other comparable jobs with no impact on accuracy. However, it has been key for the lifetime data in the comparison shown in Figure 7, where the three versions of the proposed activity recognition system have been subjected to a cross-comparison.

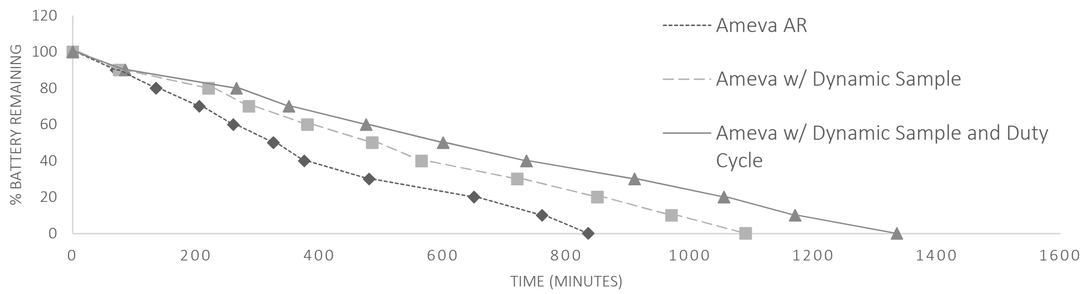


Figure 7. Comparison between simple Ameva method, Ameva with dynamic sample optimization and Ameva with the dynamic sample and duty cycle.

Furthermore, a second step towards energy savings is introduced by applying idle activity detection. It is well known that for some time, users do not wear their smartphones, so activity recognition is not available. However, during this time, the system is running and, thus, consuming power. Thanks to the idle activity recognition, the system can detect this situation and, consequently, reduce the windows processed per unit time. This decision saves energy and reduces the system overload. For this purpose, when no acceleration is detected on the smartphone, idle mode is enabled.

Algorithm 2 presents this process.

Algorithm 2 Duty cycle algorithm.

```
DutyCyclePeriods ← 3,600,000, 1,800,000, 600,000, 300,000, 60,000, 10,000, 3000
AccDataCollection ← ∅
Threshold ← 0.3
A_last ← ∅
AmevaIsRunning ← true
WindowsSleep ← 0
MinWindowsSleep ← 5
DutyIndex ← 0
WinSamples ← WindowLength(WindowSize);
while AmevaIsRunning do
  while size(AccDataCollection) < WinSamples do
    AccDataCollection ← GetLastAccData()
  end while
  WindowVar ← Var(AccDataCollection)
  if WindowVar < Threshold then
    A_last ← No_Device
    WindowsSleep ← WindowsSleep + 1
    if WindowsSleep == MinWindowsSleep then
      if DutyIndex < size(DutyCyclePeriods) then
        DutyIndex ← DutyIndex + 1
      end ifSleep(DutyCyclePeriods[DutyIndex])
    end if
  else
    DutyIndex ← 0
  end if
end while
```

First, a number of periods are defined in milliseconds, which will be the values that lead to sleep mode for the activity recognition system. Once the time window is obtained and the variance over this window is calculated, it is determined whether the variance is less than $Threshold$. This $Threshold$ was set to 0.3 during the experiments carried out. If the variance exceeds this threshold, this indicates that there has been no significant movement on the device. If so, the $WindowsSleep$ varis increased, responsible for counting the number of windows without activity. If $WindowsSleep$ is equal to the number of windows needed to decide that the device is in a period of inactivity, given by $MinWindowsSleep$, the algorithm proceeds to the activation or update of the duty cycle. This update consists of a gradual increase, depending on the values of $DutyCyclePeriods$, of the idle time, while it is determined that there is no movement in the successive time windows.

The impact of such optimization in the battery lifetime can be seen in Figure 7. By using the simple Ameva detection, the battery life time reaches 14 h. This is because the system is always on and the accelerometers are working at 50 Hz. Applying the first optimization criteria, the dynamic sampling rate, a battery life of 18 h is obtained, 3 more than previously, because some activities reduce the sampling rate and, therefore, the battery consumption. The last optimization applied to Ameva classification is the idle activity detection. Because most users expend much time at home or working with the device on the table, idle activity detection tunes up the system, avoiding unnecessary energy drain, and the battery lifetime is up to 26 h.

5. Results and Discussion

Once the basis of the activity recognition algorithm has been laid out, an analysis of the new proposal can be performed. To this end, the new development is compared with widely-used recognition systems based on neural network, decision tree, SVM and the naive Bayes classifier. These systems have been trained using MATLAB R2014b and implemented on the devices through the Android Studio and SDK tools provided by Google. In order to obtain energy consumption results, as will be explained further below, the monitoring software, Trepn, has been used. Trepn Profiler is an on-target power and performance profiling application for mobile devices distributed by Qualcomm for monitoring snapdragon processors. This diagnostic tool allows profiling of the performance and power consumption of Android applications running under this family of processors. Figure 8 shows a screenshot of this tool. As can be observed in the upper side, battery power is being monitored.

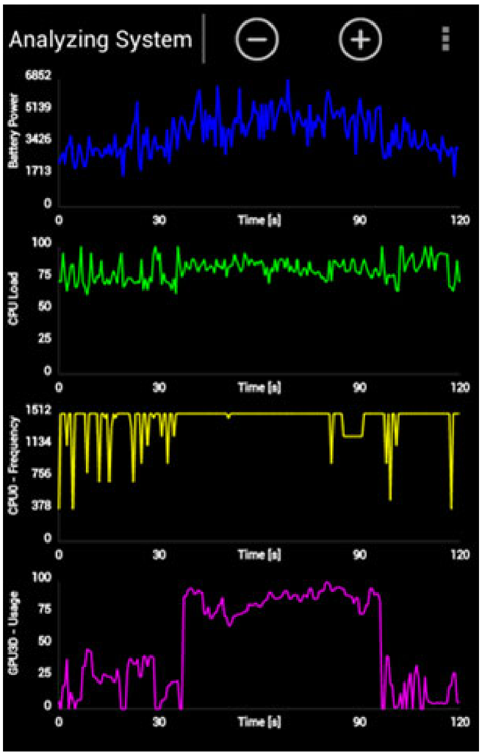


Figure 8. Trepn profiling tool screenshot.

In this case, both learning and recognition are performed by continuous methods. The test process is conducted on Google Nexus S, Samsung Galaxy S3, LG Nexus 5 and Google Nexus One devices for a group of 10 users. Notably, the activity habits of these users are quite different, as two of them are under 25 years old, five users are between 25 and 40 years old and three are over 40. To determine the energy impact of the developed system over the used smartphones, each user was monitored for 15 days. The application ran continuously in the users' smartphones during these 15 days. A total of 3640 h were monitored for all users through the Android application developed.

Furthermore, to perform the accuracy evaluation, each user introduced into the same application the activity that was carried out. In this case, getting users to record their daily activity at all times was very complicated, and hence, a total of 1215 h of registered activities was obtained. From these, a study on the precision was completed. More specifically, the evaluation dataset for each activity is composed as follows: no device (332 h), standing (87 h), walking (198 h), lying (371 h), sitting (215 h), upstairs (5 h) and downstairs (7 h).

5.1. Performance Analysis

In this section, the results of the performance analysis comparison between the method presented in this paper and different related works are presented. These methods are Ameva, ANN, binary tree, Bayesian, KNN, SVM, Mahalanobis distance and discriminant analysis. This test is conducted on eight different activities to unify the results for all users. Based on data from the UCI HAR, UCI PAMAP2 and the dataset built in this paper, a more in-depth comparison is performed. To this end, various measures for the evaluation of the performance of the information retrieval system are used. All measures used below assume a ground truth (gold standard value) contained in the labeled datasets. This comparison was performed from the implementation of the algorithms referenced previously. In Table 7, the differences between the compared methods can be observed (These results have been obtained using the UCI HAR dataset, the PAMAP2 dataset and a custom dataset built in this work). Most values presented for each measure show that the Ameva method gives better classification than the other algorithms, especially regarding precision. That is, the number of false positives in the Ameva method is lower than in the others methods. Furthermore, the training process for Ameva was faster than the others, except for binary tree, where the time was very similar. Appendix B shows more in-depth performance measures obtained by applying different algorithms compared in the table above. These values have been separated according to the specific activity. Thus, the accuracy variation can be observed more in detail between the different methods depending on the activity undertaken. However, as can be seen in these tables, there is no great difference between the values obtained in terms of activity. Therefore, it can be concluded that the compared learning methods are robust to the activities under monitoring. Based on the results for error and classification analysis above, it can be determined that the Ameva method for activity recognition presents better results than the other methods, which are widely used throughout the literature, especially SVM and Bayesian. Furthermore, not only is the execution time of the Ameva algorithm faster than that given by the others, the risk of overloading the system under the Ameva method is also lower. This is because a majority vote brings a dynamism that makes certain statistical values not critical when performing the classification, as for example with classification trees.

Table 7. Performance comparison in % by using measures of evaluation.

Method	Accuracy	Recall	Specificity	Precision	F_1-Measure
Ameva	99.23	96.93	99.56	96.94	96.93
ANN	98.36	93.50	99.07	93.47	93.48
C4.5	97.51	90.08	98.58	90.04	90.03
Bayesian	92.56	70.96	95.75	70.22	70.41
KNN	97.66	90.74	98.66	90.66	90.69
SVM	90.56	63.09	94.61	62.22	62.41
Mahalanobis distance	93.48	87.41	93.87	91.74	91.67
Discriminant analysis	99.15	96.14	97.75	96.37	96.28

5.2. Energy Consumption Results

A comparison with other works in the field of energy consumption is not easy at all. The main problems are the heterogeneity of smartphones on the market with different batteries, consumption, screens and processors and the use of the smartphone for other tasks, such as calling, reading emails or using WhatsApp. A real analysis without any restriction to the users was carried out. Users utilize theirs smartphones normally. Hence, battery consumption depends on this use. Tests were executed on LG Nexus 5, Samsung Galaxy S3 and Samsung Galaxy S4 devices. The devices were restored to their original configuration after each test to avoid interference from external application consumption.

Figure 9 shows the battery lifetime for 10 users by applying the Ameva method with the optimizations described above. As can be seen, battery life depends largely on the users' habits. Whereas, for User 3, the usage time is up to 23 h, for User 4, it is close to 18 h. It must be noted that all users are related to computer science environments, and the use of their devices is quite high.

The current work is now compared with KNN, binary decision tree (C4.5), SVM, neural networks and the naive Bayes classifier. In all cases, the process for obtaining the needed data and calculating the features is the same. This allows comparison of just the computational cost of the classification algorithms.

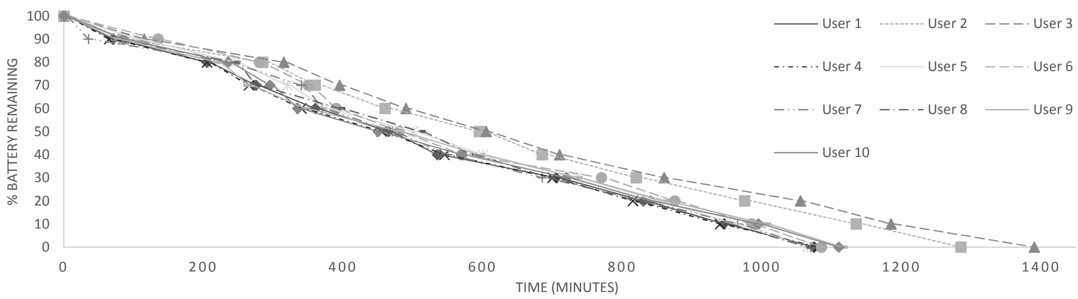


Figure 9. Analysis of battery lifetime through different user habits.

Table 8 shows the battery lifetime (in minutes) for each classifier, with four tests conducted for each one. As can be observed, the Ameva classifier extends the battery time by three hours compared with the next most efficient method, the binary decision tree (C4.5). In this section, learning times are not taken into account; only the recognition process was evaluated.

Table 8. Battery lifetime in minutes for the execution with different classifiers.

Algorithm	Test				Mean
	1	2	3	4	
Ameva	1130	1150	1070	1205	1138
KNN	750	730	760	700	731
Binary decision tree (C4.5)	915	925	898	907	905
SVM	820	870	780	830	822
Neural Networks	515	597	578	510	548
Naive Bayes	780	750	820	870	814

Finally, this work is compared with other AR systems from the literature [30,50,55,56] implemented and run on the same user's smartphone. Figure 10 shows the results.

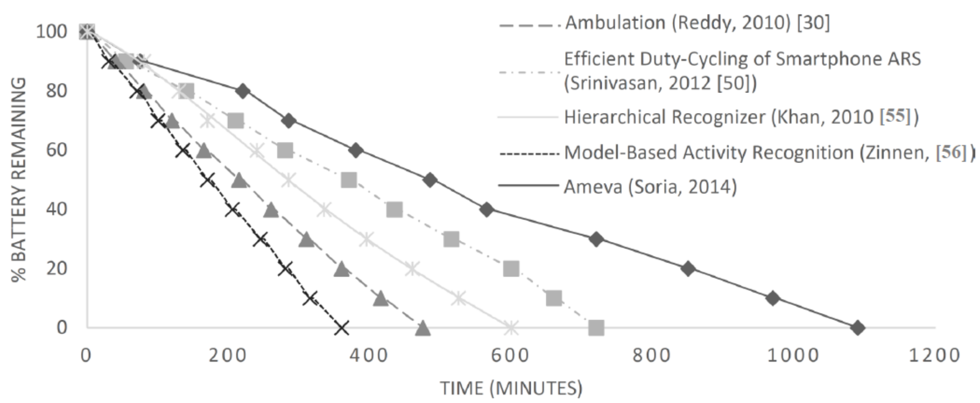


Figure 10. Battery lifetime analysis over other methods.

It can be observed in Figure 10 that the Ameva system increases the battery lifetime by up to 4.5 h. As mentioned before, all methods were executed in the same smartphone used by the same user, each for five days. Later, other works will be compared in terms of accuracy, but for now, we can see that the proposed system significantly reduces the energy consumed. Figure 11 shows a power consumption comparison based on W/h (The watt-hour (W/h) is a unit of energy equivalent to one watt of power expended for one hour) for each compared method. To reduce deviations from the power consumption given by Trepn software, the device was previously calibrated using a spectrum analyzer, which measured the real consumption. Thereby, the correctness of the Trepn Qualcomm software readings was checked. Moreover, because of the impact that CPU consumption obviously has on processing cost, it was decided to keep a wake lock on the device. This ensured that the processor would not enter into a low-power mode during the performance comparison. Thus, the baseline for comparison is one in which the device is not used, but the processors remain awake. It should be noted that this wake lock was deactivated for the battery life tests, in order to perform testing of each algorithm behavior in a real application scenario. As can be seen in Figure 11, the energy consumption of the Ameva algorithm is significantly less than the other alternatives. Specifically, it can be seen that the consumption is about 50% that of the most efficient alternative among those compared. This is mainly due to two reasons: first, the possibility of carrying out a selection of variables at run time, depending on the coefficient Ameva for each statistic; and second, the reduced computational cost of the proposed classification method.

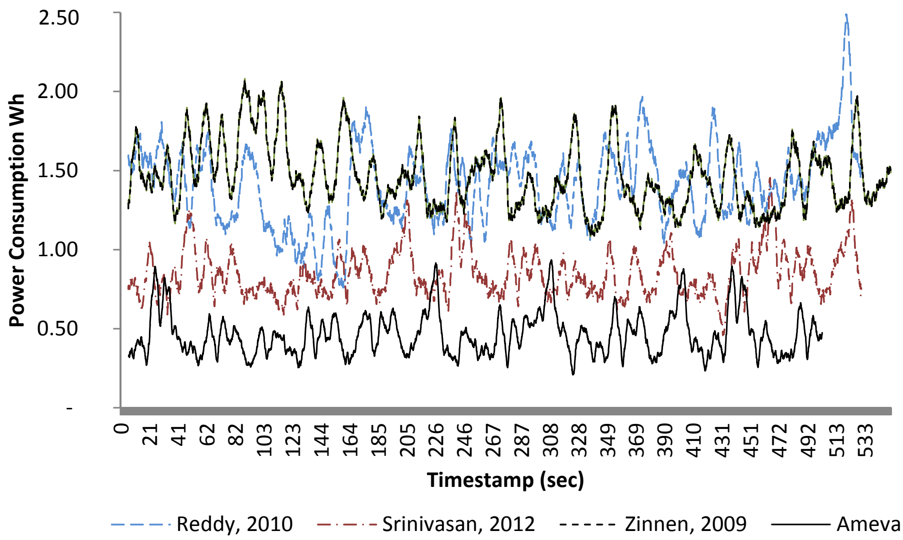


Figure 11. Power consumption analysis.

To clarify the differences in the power consumption characteristics of different classifiers, these are compared with the proposed algorithm. The C4.5 algorithm will specifically be discussed, which was *a priori* more efficient from the point of view of complexity and which has been used by many studies in the literature. Regarding accuracy, there is no big difference between Ameva and the algorithm based on C4.5, as was shown previously. However, there is a difference from the point of view of energy consumption. Specifically, Ameva is about 50% below the average consumption of the algorithm C4.5. This is mainly due to the capability of automatic selection that is made based on the characteristics of the Ameva coefficient. Those intervals generated by Ameva whose coefficient is less than the threshold (at this moment, the threshold is defined in a static way) are removed, and the associated statistics are not taken into account when computing the result. This makes the used statistic, on average, 40% of the pre-established attributes of each temporal window. This process, applied to the UCI HAR dataset [39], makes the total statistics considered go from 561 to 63, while C4.5 should consider 117 attributes for classification. The tree generated by the C4.5 algorithm can be seen in Figure 12. Consequently, this decrease in the processed attributes makes it unnecessary to process the data on the time window in order to get them and, as a result, reduces the complexity of the process of collecting statistics regarding the C4.5 algorithm.

In conclusion, the power consumption aim of this work has been accomplished. It is noted in [57] that users recharge their smartphones once a day, mostly happening at 8 pm, when users are at home. Because our system allows the user to maintain the device battery for more than 18 hours in all cases, they can retain their recharging habits.

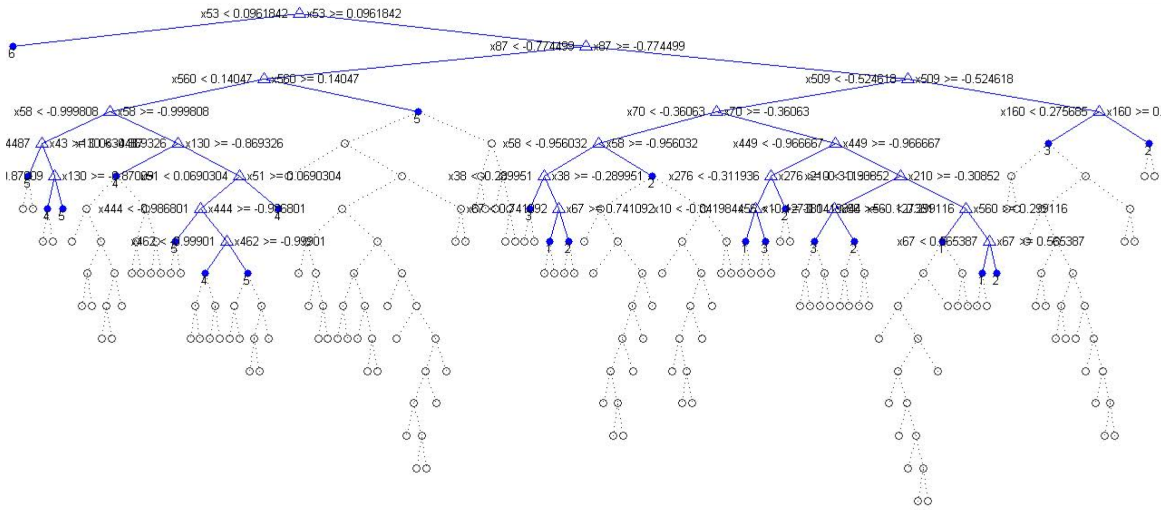


Figure 12. C4.5 bounded decision tree.

5.3. Data Traffic Reduction

Regarding the flow of information between device and server, it has been determined that it is much lower in the case of the Ameva method. This server is responsible for safeguarding the training data and recognized activities. The received data flow necessary for the maintenance of such data is 4.7 Kbytes for the system based on neural networks and 500 bytes in the case of Ameva. That is, the data flow between device and server is reduced by more than 70%. This reduction prevents additional costs arising from excessive use of the data network, because, with the Ameva method, it is only necessary to send the bounds of the intervals. However, using neural networks to train the system, every network parameter and weight must be sent, thereby resulting in a much greater size. The same occurs for all continuous learning methods.

5.4. Comparing with Other Works

Once the proposal has been analyzed and optimization has been applied to the original system, a comparison with other proposals for activity monitoring is made in this section. The comparison is based on the following attributes: number of activities, average accuracy, number of sensors, execution environment (the device on which to carry out the recognition), average processing time and battery lifetime. Based on previous data, an analysis of the latest work in the activity recognition field was developed. From all related works, four studies were chosen to carry out the analysis. All four studies are recent and present a large number of citations.

Table 9 shows the results of the comparison. An analysis of the table shows that the number of activities recognized by our proposal is higher than those in the other proposals, except for the Kerem Altun study. However, the Kerem Altun proposal uses five specific sensors, whereas our proposal uses only one sensor embedded in the user's own device.

Table 9. Comparison with other activity recognition systems. ND (not determined). * These results are a little confusing. Most smartphones' batteries rarely reach up to 120 h making intensive use of the accelerometer. Furthermore, only two activities (stable or moving) are recognized. ** Making intensive use of Bluetooth.

Method	No. of Activities	Average Accuracy (%)	No. of Sensors	Execution Environment	Average Process Time (s)	Battery Life (h)
Our proposal 2014	9	98	1 general	Smartphone	0.3	18 (measured)
Kerem Altun [21] 2010	16	97	5 specific	AMD Athlon 64 X2	0.1	Not applicable
Chang Goo Han [11] 2010	6	92	1 specific	PC	ND	Not applicable
Tanzeem Choudhury [36] 2008	8	84	2 specific	Specific device	ND	10–20 (experimental)
Sasank Reddy [30] 2010	5	93	8 general	Phone	15.0	8.2 (measured)
Hong Lu [49] 2010	5	94	1 general	Smartphone	ND	14 (experimental)
Vijay Srinivasan [50] 2012	6	91	1 general	Smartphone	0.6	12.5 (measured)
Khan [55] 2010	7	97	1 specific	Specific device	2.3	10 (measured)
Yi Wang [58] 2012	2	90	1 general	Phone	ND	150 (experimental) (*)
Jia [59] 2013	7	98	2 specific	Smartphone	1.5	7 (measured) (**)
Andreas Zinnen [56] 2009	21	85	5 general	Smartphone	ND	6.5 (measured) (**)

The Chang Goo and Tanzeem Choudour proposals use just one sensor, but the accuracy is lower than that presented in this paper. Furthermore, the possibility of integrating the whole recognition system inside a mobile phone renders the device more convenient for users.

Another aspect to consider in the comparison is the efficiency of the methods at performing the whole process. In this sense, it is necessary to differentiate between two types of proposals: smartphone-embedded methods and server methods. In the former, the process is executed entirely in the mobile phone, whereas in the latter, a computer is required to execute the solution and to process the data. For this reason, in the Kerem Altun and Chang Goo proposals, the battery life is longer than that in our proposal, which collects and processes the data in the device itself.

Andreas Zinnen marked a new point of view of activity recognition, called model-oriented methods. In that work, some accelerometers are placed on the user's arm, the aim being to recognize the movements made by the body like a three-dimensional model of the user. This technique is often used in animation, but the main drawback is the number of sensors needed. Furthermore, one of the aims of this work is to develop the entire system in the user's smartphone, without external sensors. However, as can be seen in [56], the number of activities recognized is quite high.

Finally, Jia's work introduces other external sensors, such as the ECG meter, which improves the accuracy of the whole system. However, this kind of system has a drawback: the power consumption caused by the Bluetooth connection between external sensors and the smartphone.

6. Conclusions and Future Work

This work presents a highly accurate recognition system, based on discrete variables, that uses the Ameva discretization algorithm and a new Ameva-based classification system. It has therefore been possible to achieve an average accuracy of 98% for the recognition of eight types of activities. Furthermore, working with discrete variables significantly reduces the computational cost associated with data processing during the recognition process. By using this process to increase recognition frequency, it has been possible to obtain a physical activity reading every four seconds and to save this contextual information in the user activity live log.

The main problem detected in the system based on statistical learning is the limitation of the number of activities that can be recognized. Actually, the problem is not provoked by the method itself, but by the accelerometer sensors. The number of system features is limited, thus leading to a strong correlation between these variables. This problem could be solved by including new sensors (NFC, Bluetooth, and so on), which provide more information to the system.

Based on the studies performed and the conclusions reached in the Dynamic Sample Rate and Duty Cycle section, the accuracy of the system, once duty cycle optimization is applied, does not vary depending on the user who performs the test. As was mentioned before, this method has a very slight impact on system accuracy. This is because its auto-reconfiguration makes it possible to increase the sample rate if necessary. However, this strategy brings considerable benefit in terms of the energy savings achieved.

In this way, a system that extends the number of recognized activities is currently being developed. It is based on the data presented in this work combined with the help of GPS and NFC sensors embedded in the device. The system involves the analysis of labels installed in smart items that, in addition to providing information about the item itself, inform the system about the activities supported. Therefore, if a user is sitting near the television remote control, then the new activity recognition would be watching TV. Similarly, if a user is walking and GPS information indicates that the user is in the park, the activity would be walking through the park.

Acknowledgments

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness, HERMES (TIN2013-46801-C4-1-r) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

Author Contributions

Luis Gonzalez-Abril together with Miguel Angel Alvarez de la Concepcion designed the Ameva discretization algorithm and performed the tests in a controlled environment. Luis Miguel Soria Morillo developed the adaptation of the Ameva algorithm in the field of the activity recognition environment and designed the Adaptive Duty Cycle and Sample Rate algorithms. Juan Antonio Ortega Ramirez conducted and designed the experimentation. Luis Gonzalez-Abril and Luis Miguel Soria Morillo validated the results and drafted the manuscript. All authors read and approved the final manuscript.

Appendix

Appendix A

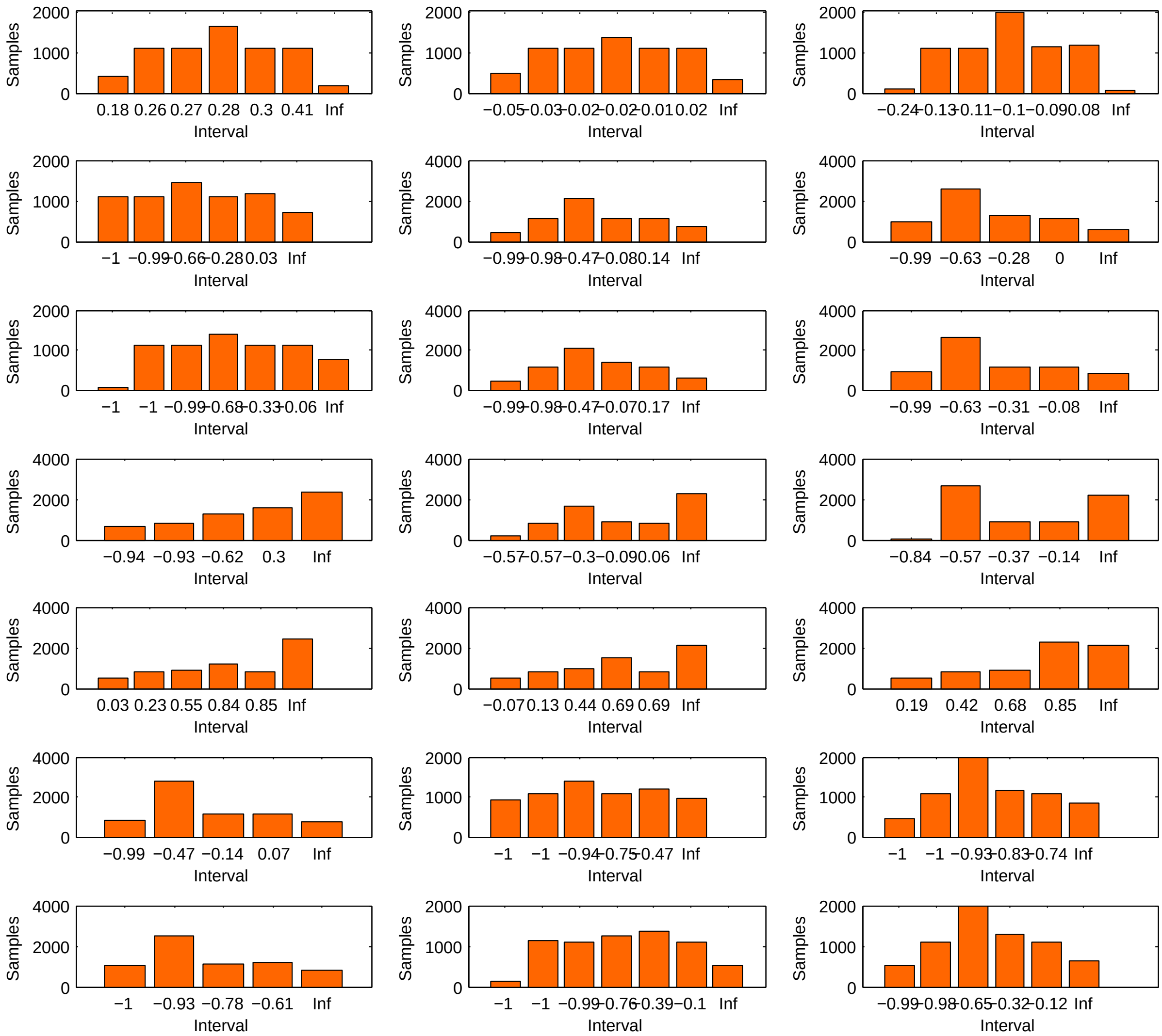


Figure A1. Distribution of instances in each interval for the first 21 statistics.

Appendix B

Table B1. Confusion matrix and performance values for the Ameva-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-Measure
Walk	7200	10	10	61	80	40	37	89	7527	7200	52,135	327	291	98.97%	96.12%	99.38%	95.66%	95.88%
Jump	10	7304	0	152	0	0	21	8	7495	7304	52,180	191	278	99.22%	96.33%	99.64%	97.45%	96.89%
Immobile	0	0	7256	0	0	0	0	142	7398	7256	52,529	142	26	99.72%	99.64%	99.73%	98.08%	98.86%
Run	16	128	0	7298	0	7	40	8	7497	7298	52,205	199	251	99.25%	96.68%	99.62%	97.35%	97.01%
Up	89	70	0	8	7273	195	2	2	7639	7273	51,918	366	396	98.73%	94.84%	99.30%	95.21%	95.02%
Down	96	0	0	23	316	7252	4	5	7696	7252	52,015	444	242	98.86%	96.77%	99.15%	94.23%	95.48%
Cycle	0	0	0	0	0	0	7302	1	7303	7302	52,536	1	114	99.81%	98.46%	100.00%	99.99%	99.22%
Drive	80	70	16	7	0	0	10	7215	7398	7215	52,300	183	255	99.27%	96.59%	99.65%	97.53%	97.05%
Accumulated	7491	7582	7282	7549	7669	7494	7416	7470	59,953	58,100	417,818	1853	1853	99.23%	96.93%	99.56%	96.94%	96.93%

Table B2. Confusion matrix and performance values for the ANN-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-Measure
Walk	6930	46	41	117	107	79	86	121	7527	6930	51,807	597	619	97.97%	91.80%	98.86%	92.07%	91.93%
Jump	76	7124	16	187	13	8	46	25	7495	7124	51,892	371	566	98.44%	92.64%	99.29%	95.05%	93.,83%
Immobile	47	9	6970	4	63	46	13	246	7398	6970	52,307	428	248	98.87%	96.56%	99.19%	94.21%	95.37%
Run	44	221	3	7060	39	43	59	28	7497	7060	52,016	437	440	98.54%	94.13%	99.17%	94.17%	94.15%
Up	126	89	17	25	7091	256	19	16	7639	7091	51,587	548	727	97.87%	90.70%	98.95%	92.83%	91.75%
Down	103	4	7	27	415	7103	16	21	7696	7103	51,754	593	503	98.17%	93.39%	98.87%	92.29%	92.84%
Cycle	36	32	43	35	75	56	6928	98	7303	6928	52,386	375	264	98.93%	96.33%	99.29%	94.87%	95.59%
Drive	187	165	121	45	15	15	25	6825	7398	6825	52,000	573	555	98.12%	92.48%	98.91%	92.25%	92.37%
Accumulated	7549	7690	7218	7500	7818	7606	7192	7380	59,953	56,031	415,749	3922	3922	98.36%	93.50%	99.07%	93.47%	93.48%

Table B3. Confusion matrix and performance values for the C4.5-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-Measure
Walk	6210	78	98	318	298	96	92	337	7527	6210	51,570	1317	856	96.38%	87.89%	97.51%	82.50%	85.11%
Jump	113	6932	54	245	36	17	67	31	7495	6932	51,645	563	813	97.70%	89.50%	98.92%	92.49%	90.97%
Immobile	59	15	6790	23	83	72	59	297	7398	6790	52,073	608	482	98.18%	93.37%	98.85%	91.78%	92.57%
Run	54	294	23	6896	64	57	71	38	7497	6896	51,649	601	807	97.65%	89.52%	98.85%	91.98%	90.74%
Up	160	113	51	34	6943	298	21	19	7639	6943	51,272	696	1042	97.10%	86.95%	98.66%	90.89%	88.88%
Down	187	18	15	53	427	6917	43	36	7696	6917	51,600	779	657	97.60%	91.33%	98.51%	89.88%	90.60%
Cycle	64	79	54	66	96	74	6716	154	7303	6716	52,241	587	409	98.34%	94.26%	98.89%	91.96%	93.10%
Drive	219	216	187	68	38	43	56	6571	7398	6571	51,643	827	912	97.10%	87.81%	98.42%	88.82%	88.31%
Accumulated	7066	7745	7272	7703	7985	7574	7125	7483	59,953	53,975	413,693	5978	5978	97.51%	90.08%	98.58%	90.04%	90.03%

Table B4. Confusion matrix and performance values for the Bayesian-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-measure
Walk	5067	187	218	573	421	329	198	534	7527	5067	49,701	2460	2725	91.35%	65.03%	95.28%	67.32%	66.15%
Jump	741	5180	113	1023	57	124	216	41	7495	5180	49,239	2315	3219	90.77%	61.67%	95.51%	69.11%	65.18%
Immobile	115	87	5014	288	356	298	243	997	7398	5014	51,473	2384	1082	94.22%	82.25%	95.57%	67.78%	74.31%
Run	78	1520	46	5290	155	96	245	67	7497	5290	50,047	2207	2409	92.30%	68.71%	95.78%	70.56%	69.62%
Up	613	509	79	77	5521	708	56	76	7639	5521	49,478	2118	2836	91.74%	66.06%	95.90%	72.27%	69.03%
Down	470	27	43	76	1309	5641	76	54	7696	5641	50,323	2055	1934	93.35%	74.47%	96.08%	73.30%	73.88%
Cycle	105	127	145	227	451	300	5216	732	7303	5216	51,511	2087	1139	94.62%	82.08%	96.11%	71.42%	76.38%
Drive	603	762	438	145	87	79	105	5179	7398	5179	50,054	2219	2501	92.13%	67.43%	95.75%	70.01%	68.70%
Accumulated	7792	8399	6096	7699	8357	7575	6355	7680	59,953	42,108	401,826	17,845	17,845	92.56%	70.96%	95.75%	70.22%	70.41%

Table B5. Confusion matrix and performance values for the SVM-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-Measure
Walk	4562	210	289	658	513	401	201	693	7527	4562	48,751	2965	3675	88.92%	55.38%	94.27%	60.61%	57.88%
Jump	819	4716	214	1142	64	176	275	89	7495	4716	48,621	2779	3837	88.96%	55.14%	94.59%	62.92%	58.77%
Immobile	156	115	4663	342	398	341	296	1087	7398	4663	51,018	2735	1537	92.87%	75.21%	94.91%	63.03%	68.58%
Run	89	1598	85	4982	174	158	335	76	7497	4982	49,516	2515	2940	90.90%	62.89%	95.17%	66.45%	64.62%
Up	769	654	97	85	4728	1127	82	97	7639	4728	48,875	2911	3439	89.41%	57.89%	94.38%	61.89%	59.83%
Down	636	53	68	84	1614	5092	87	62	7696	5092	49,504	2604	2753	91.06%	64.91%	95.00%	66.16%	65.53%
Cycle	354	297	241	361	580	467	4189	814	7303	4189	51,120	3114	1530	92.25%	73.25%	94.26%	57.36%	64.34%
Drive	852	910	543	268	96	83	254	4392	7398	4392	49,637	3006	2918	90.12%	60.08%	94.29%	59.37%	59.72%
Accumulated	8237	8553	6200	7922	8167	7845	5719	7310	59,953	37,324	397,042	22,629	22,629	90.56%	63.09%	94.61%	62.22%	62.41%

Table B6. Confusion matrix and performance values for the KNN-based classification system.

Test/Real	Walk	Jump	Immobile	Run	Up	Down	Cycle	Drive	Instances	TP	TN	FP	FN	Accuracy	Recall	Specificity	Precision	F_1-Measure
Walk	6726	65	52	143	137	87	121	196	7527	6726	51,490	801	936	97.10%	87.78%	98.47%	89.36%	88.56%
Jump	161	6879	26	265	36	27	62	39	7495	6879	51,768	616	690	97.82%	90.88%	98.82%	91.78%	91.33%
Immobile	65	26	6719	21	95	64	26	382	7398	6719	52,193	679	362	98.26%	94.89%	98.72%	90.82%	92.81%
Run	86	244	36	6854	65	79	84	49	7497	6854	51,834	643	622	97.89%	91.68%	98.77%	91.42%	91.55%
Up	149	98	32	41	6881	375	36	27	7639	6881	51,317	758	997	97.07%	87.34%	98.54%	90.08%	88.69%
Down	187	21	16	31	504	6870	26	41	7696	6870	51,522	826	735	97.40%	90.34%	98.42%	89.27%	89.80%
Cycle	65	51	58	47	124	78	6748	132	7303	6748	52,254	555	396	98.41%	94.46%	98.95%	92.40%	93.42%
Drive	223	185	142	74	36	25	41	6672	7398	6672	51,689	726	866	97.34%	88.51%	98.61%	90.19%	89.34%
Accumulated	7662	7569	7081	7476	7878	7605	7144	7538	59,953	54,349	414,067	5604	5604	97.66%	90.74%	98.66%	90.66%	90.69%

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Manson, J.E.; Greenland, P.; LaCroix, A.Z.; Stefanick, M.L.; Mouton, C.P.; Oberman, A.; Perri, M.G.; Sheps, D.S.; Pettinger, M.B.; Siscovick, D.S. Walking compared with vigorous exercise for the prevention of cardiovascular events in women. *N. Engl. J. Med.* **2002**, *347*, 716–725.
2. Ellekjaer, H.; Holmen, J.; Vatten, L. Physical activity and stroke mortality in women. *Stroke* **2000**, *31*, 14–18.
3. Sattelmair, J.R.; Kurth, T.; Buring, J.E.; Lee, I.M. Physical Activity and Risk of Stroke in Women. *Stroke* **2010**, *41*, 1243–1250.
4. Lee, I.M.; Rexrode, K.M.; Cook, N.R.; Manson, J.E.; Buring, J.E. Physical activity and coronary heart disease in women. *JAMA* **2001**, *285*, 1447–1454.
5. Weiser, M. The computer for the 21st century. In *Human-Computer Interaction*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 933–940.
6. Ordoñez, F.J.; de Toledo, P.; Sanchis, A. Activity Recognition Using Hybrid Generative/Discriminative Models on Home Environments Using Binary Sensors. *Sensors* **2013**, *13*, 5460–5477.
7. Hong, Y.J.; Kim, I.J.; Ahn, S.C.; Kim, H.G. Mobile health monitoring system based on activity recognition using accelerometer. *Simul. Model. Pract. Theory* **2010**, *18*, 446–455.
8. Khan, A.M.; Lee, Y.K.; Lee, S.; Kim, T.S. Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly. *Med. Biol. Eng. Comput.* **2010**, *48*, 1271–1279.
9. Preece, S.J.; Goulermas, J.Y.; Kenney, L.P.J.; Howard, D.; Meijer, K.; Crompton, R. Activity identification using body-mounted sensors—A review of classification techniques. *Physiol. Measur.* **2009**, *30*, R1–R33.
10. Ravi, N.; Nikhil, D.; Mysore, P.; Littman, M.L. Activity recognition from accelerometer data. In Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence (IAAI), Pittsburgh, PA, USA, 9–13 July 2005; pp. 1541–1546.
11. Han, C.W.; Kang, S.J.; Kim, N.S. Implementation of HMM-Based Human Activity Recognition Using Single Triaxial Accelerometer. *IEICE Trans.* **2010**, *93-A*, 1379–1383.
12. Paiyarom, S.; Tungamchit, P.; Keinprasit, R.; Kayasith, P. Activity monitoring system using Dynamic Time Warping for the elderly and disabled people. In Proceedings of the 2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 February 2009; pp. 1–4.
13. Hong, Y.J.; Kim, I.J.; Ahn, S.C.; Kim, H.G. Activity Recognition Using Wearable Sensors for Elder Care. *Future Gener. Commun. Netw.* **2008**, *2*, 302–305.

14. Musolesi, M.; Piraccini, M.; Fodor, K.; Corradi, A.; Campbell, A.T. Supporting Energy-Efficient Uploading Strategies for Continuous Sensing Applications on Mobile Phones Pervasive Computing. In Proceedings of the 8th International Conference on Pervasive Computing, Helsinki, Finland, 17–20 May 2010; Floréen, P., Krüger, A., Spasojevic, M., Eds.; Springer: Berlin/Heidelberg, Germany; pp. 355–372.
15. Brezmes, T.; Gorricho, J.-L.; Cotrina, J. Activity Recognition from Accelerometer Data on a Mobile Phone. *Test* **2009**, *5518*, 796–799.
16. Lepri, B.; Mana, N.; Cappelletti, A.; Pianesi, F.; Zancanaro, M. What is happening now? Detection of activities of daily living from simple visual features. *Pers. Ubiquitous Comput.* **2010**, *14*, 749–766.
17. Bicocchi, N.; Mamei, M.; Zambonelli, F. Detecting activities from body-worn accelerometers via instance-based algorithms. *Pervasive Mob. Comput.* **2010**, *6*, 482–495.
18. Intille, S.S.; Bao, L. *Physical Activity Recognition from Acceleration Data under SemiNaturalistic Conditions.* Technical Report; Massachusetts Institute of Technology: Cambridge, MA, USA, 2003.
19. Laerhoven, K.V. ISWC 2010: The Latest in Wearable Computing Research. *IEEE Pervasive Comput.* **2011**, *10*, 8–10.
20. Khan, A.M.; Lee, Y.K.; Lee, S.Y.; Kim, T.S. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *Trans. Inf. Tech. Biomed.* **2010**, *14*, 1166–1172.
21. Altun, K.; Barshan, B.; Tunçel, O. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recogn.* **2010**, *43*, 3605–3620.
22. Choudhury, T.; LaMarca, A.; LeGrand, L.; Rahimi, A.; Rea, A.; Borriello, G.; Hemingway, B.; Koscher, K.; Landay, J.A.; Lester, J.; *et al.* The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Comput.* **2008**, *7*, 32–41.
23. Liang, G.; Cao, J.; Zhu, W. CircleSense: A Pervasive Computing System for Recognizing Social Activities. In Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communication, San Diego, CA, USA, 18–22 March 2013.
24. Fogarty, J.; Au, C.; Hudson, S.E. Sensing from the basement: A feasibility study of unobtrusive and low-cost home activity recognition. In Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, Montreux, Switzerland, 15–18 October 2006; ACM: New York, NY, USA; pp. 91–100.
25. Stager, M.; Lukowicz, P.; Troster, G. Implementation and evaluation of a low-power sound-based user activity recognition system. In Proceedings of the Eighth International Symposium on Wearable Computers, Arlington, VA, USA, 31 October–3 November 2004; pp. 138–141.
26. Wojek, C.; Nickel, K.; Stiefelhagen, R. Activity Recognition and Room-Level Tracking in an Office Environment. In Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, Heidelberg, Germany, 3–6 September 2006; pp. 25–30.
27. Li, M.; Rozgic, V.; Thatte, G.; Lee, S.; Emken, B.; Annavaram, M.; Mitra, U.; Spruijt-Metz, D.; Narayanan, S. Multimodal Physical Activity Recognition by Fusing Temporal and Cepstral Information. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2010**, *18*, 369–380.

28. Pawar, T.; Chaudhuri, S.; Duttagupta, S.P. Body Movement Activity Recognition for Ambulatory Cardiac Monitoring. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 874–882.
29. Ward, J.; Lukowicz, P.; Troster, G.; Starner, T. Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers. *IEEE Trans. Pattern Anal. Machine Intell.* **2006**, *28*, 1553–1567.
30. Reddy, S.; Mun, M.; Burke, J.; Estrin, D.; Hansen, M.; Srivastava, M. Using mobile phones to determine transportation modes. *ACM Trans. Sen. Netw.* **2010**, *6*, 1–27.
31. Sharkey, J. Coding for life—Battery life, that is. In Proceedings of the Google IO Developer Conference, San Francisco, CA, USA, 27–28 May 2009.
32. Maloney, S.; Boci, I. Survey: Techniques for Efficient energy consumption in Mobile Architectures. *Power (mW)* **2012**, *16*, 7–35.
33. Forrester. *North American Technographics Consumer Technology Online Survey, Q1 2011 (US)*; Technical Report; Forrester: Cambridge, MA, USA, 2011.
34. Miluzzo, E.; Lane, N.D.; Fodor, K.; Peterson, R.; Lu, H.; Musolesi, M.; Eisenman, S.B.; Zheng, X.; Campbell, A.T. Sensing meets mobile social networks: The design, implementation and evaluation of the CenceMe application. In Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, New York, NY, USA, 5–7 November 2008; ACM: New York, NY, USA; pp. 337–350.
35. Yan, Z.; Subbaraju, V.; Chakraborty, D.; Misra, A.; Aberer, K. Energy-Efficient Continuous Activity Recognition on Mobile Phones: An Activity-Adaptive Approach. In Proceedings of the 2012 16th International Symposium on Wearable Computers (ISWC), Newcastle, CA, USA, 18–22 June 2012; pp. 17–24.
36. Choudhury, T.; Borriello, G.; Consolvo, S.; Haehnel, D.; Harrison, B.; Hemingway, B.; Hightower, J.; Klasnja, P.; Koscher, K.; LaMarca, A.; *et al.* The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Comput.* **2008**, *7*, 32–41.
37. Ravi, N.; Dandekar, N.; Mysore, P.; Littman, M.L. Activity recognition from accelerometer data. *AAAI* **2005**, *5*, 1541–1546.
38. Tapia, E.M.; Intille, S.S.; Haskell, W.; Larson, K.; Wright, J.; King, A.; Friedman, R. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers, Boston, MA, USA, 11–13 October 2007; pp. 37–40.
39. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. In Proceedings of the International Workshop of Ambient Assited Living (IWAAL 2012), Vitoria-Gasteiz, Spain, 3–5 December 2012.
40. Reiscs, A.; Stricker, D. Introducing a New Benchmarked Dataset for Activity Monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers (ISWC), Newcastle, CA, USA, 18–22 June 2012; pp. 108–109.
41. Gonzalez-Abril, L.; Cuberos, F.J.; Velasco, F.; Ortega, J.A. Ameva: An autonomous discretization algorithm. *Expert Syst. Appl.* **2009**, *36*, 5327–5332.

42. Falomir Llansola, Z.; Museros Cabedo, L.; González Abril, L.; Escrig Monferrer, M.T.; Ortega, J.A. A model for the qualitative description of images based on visual and spatial features. *Comput. Vis. Image Underst.* **2012**, *116*, 698–714.
43. Gonzalez-Abril, L.; Velasco, F.; Ortega, J.; Cuberos, F. A new approach to qualitative learning in time series. *Expert Syst. Appl.* **2009**, *36*, 9924–9927.
44. Falomir, Z.; Gonzalez-Abril, L.; Museros, L.; Ortega, J.A. Measures of Similarity Between Objects Based on Qualitative Shape Descriptions. *Spat. Cogn. Comput.* **2013**, *13*, 181–218.
45. Kwapisz, J.R.; Weiss, G.M.; Moore, S.A. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.* **2011**, *12*, 74–82.
46. Ward, J.A.; Lukowicz, P.; Troster, G.; Starner, T.E. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1553–1567.
47. Chu, D.; Lane, N.D.; Lai, T.T.T.; Pang, C.; Meng, X.; Guo, Q.; Li, F.; Zhao, F. Balancing energy, latency and accuracy for mobile sensor data classification. In Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, Seattle, WA, USA, 1–4 November 2011; ACM: New York, NY, USA; pp. 54–67.
48. Wang, Y.; Krishnamachari, B.; Zhao, Q.; Annavaram, M. The tradeoff between energy efficiency and user state estimation accuracy in mobile sensing. In Proceedings of MobiCase, San Diego, CA, USA, 26–29 October 2009.
49. Lu, H.; Yang, J.; Liu, Z.; Lane, N.D.; Choudhury, T.; Campbell, A.T. The Jigsaw continuous sensing engine for mobile phone applications. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, Zurich, Switzerland, 3–5 November 2010; ACM: New York, NY, USA, pp. 71–84.
50. Srinivasan, V.; Phan, T. An accurate two-tier classifier for efficient duty-cycling of smartphone activity recognition systems. In Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones, Toronto, ON, Canada, 6 November 2012; ACM: New York, NY, USA; pp. 11:1–11:5.
51. Krause, A.; Ihmig, M.; Rankin, E.; Leong, D.; Gupta, S.; Siewiorek, D.; Smailagic, A.; Deisher, M.; Sengupta, U. Trading off prediction accuracy and power consumption for context-aware wearable computing. In Proceedings of the Ninth IEEE International Symposium on Wearable Computers, Osaka, Japan, 18–21 October 2005; pp. 20–26.
52. Yurur, O.; Liu, C.H.; Liu, X.; Moreno, W. Adaptive Sampling and Duty Cycling for Smartphone Accelerometer. In Proceedings of the 2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS), Hangzhou, China, 14–16 October 2013; pp. 511–518.
53. Yan, Z.; Subbaraju, V.; Chakraborty, D.; Misra, A.; Aberer, K. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In Proceedings of the 2012 16th International Symposium on Wearable Computers (ISWC), Newcastle, CA, USA, 18–22 June 2012; pp. 17–24.
54. Shen, C.; Chakraborty, S.; Raghavan, K.R.; Choi, H.; Srivastava, M.B. Exploiting processor heterogeneity for energy efficient context inference on mobile phones. In Proceedings of the Workshop on Power-Aware Computing and Systems, Farmington, PA, USA, 3–6 November 2013.

55. Khan, A.; Lee, Y.K.; Lee, S.; Kim, T.S. A Triaxial Accelerometer-Based Physical-Activity Recognition via Augmented-Signal Features and a Hierarchical Recognizer. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 1166–1172.
56. Zinnen, A.; Blanke, U.; Schiele, B. An Analysis of Sensor-Oriented *vs.* Model-Based Activity Recognition. In Proceedings of the International Symposium on Wearable Computers, Linz, Austria, 4–7 September 2009; pp. 93–100.
57. Ferreira, D.; Dey, A.K.; Kostakos, V. Understanding human-smartphone concerns: A study of battery life. In Proceedings of the 9th International Conference on Pervasive Computing, San Francisco, CA, USA, 12–15 June 2011; Springer-Verlag: Berlin, Heidelberg, Germany; pp. 19–33.
58. Wang, Y.; Krishnamachari, B.; Annavaram, M. Semi-Markov state estimation and policy optimization for energy efficient mobile sensing. In Proceedings of the 2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), Seoul , Korea, 18–21 June 2012; pp. 533–541.
59. Jia, R.; Liu, B. Human daily activity recognition by fusing accelerometer and multi-lead ECG data. In Proceedings of the 2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC), Kunming, China, 5–8 August 2013; pp. 1–4.

CHAPTER 6

MOBILE ACTIVITY RECOGNITION AND FALL DETECTION SYSTEM FOR ELDERLY PEOPLE USING AMEVA ALGORITHM

Overview

This paper addresses the human activities recognition problem for elderly people where the most important problem with classification models is that a good training process is needed to get the best results. Therefore, it is important to detect correctly the lifestyle of elderly people, but the difficulty is that they will waste their time training a device. This aspect makes the aged population a singular group for activity recognition systems.

All common learning processes have a test phase that normally is carried out in a laboratory setting and, in activity recognition, this stage is performed in a controlled environment, throughout public generated datasets and also by the developed team of the system itself. The presented approach differs from the traditional system because it has been proved effective also in a non-controlled environment.

The developed system was accurate, comfortable and efficient and monitors the physical activity carried out by the user using the core of the Ameva algorithm. Moreover, with the purpose of reducing the cost and increasing user acceptance and usability, the entire system uses only a smartphone to recover all the information required.

Context

This research topic was initiated as an extension to an awarded activity recognition system that participated in the EvAAL 2012 and EvAAL 2013 competitions where the fall detection was the main goal. This paper is the result of several years of acquiring a wide range of knowledge in these fields.

Journal information

The Journal *Pervasive and Mobile Computing* was selected for the submission of this paper that is indexed in JCR with an Impact Factor of 2.079 in 2014 and 1.719 in 2015; and a 5-Year I.F. 2.366. Also, it stands in ranking Q2 in two categories in 2016: Telecommunications (34/89); Computer Science, Information Systems (53/146).

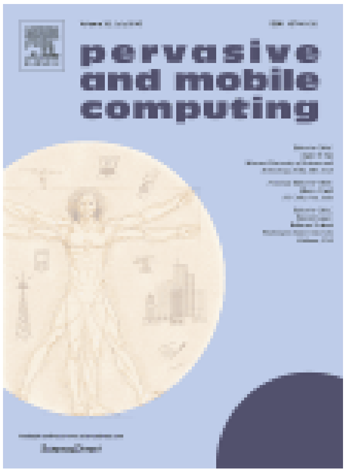


Figure 6.1: Pervasive and Mobile Computing cover.

Pervasive and Mobile Computing 34 (2017) 3–13

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Mobile activity recognition and fall detection system for elderly people using Ameva algorithm

Miguel Ángel Álvarez de la Concepción [a,*], Luis Miguel Soria Morillo [a], Juan Antonio Álvarez García [a], Luis González-Abril [b]

[a] Computer Languages and Systems Department, University of Seville, 41012 Seville, Spain
[b] Applied Economics I Department, University of Seville, 41012 Seville, Spain

ARTICLE INFO

Article history:
Available online 21 May 2016

Keywords:
Activity recognition
Artificial intelligent
Cognitive computing
Contextual information
Mobile environment
Smart-energy computing

ABSTRACT

Currently, the lifestyle of elderly people is regularly monitored in order to establish guidelines for rehabilitation processes or ensure the welfare of this segment of the population. In this sense, activity recognition is essential to detect an objective set of behaviors throughout the day. This paper describes an accurate, comfortable and efficient system, which monitors the physical activity carried out by the user. An extension to an awarded activity recognition system that participated in the EvAAL 2012 and EvAAL 2013 competitions is presented. This approach uses data retrieved from accelerometer sensors to generate discrete variables and it is tested in a non-controlled environment. In order to achieve the goal, the core of the algorithm Ameva is used to develop an innovative selection, discretization and classification technique for activity recognition. Moreover, with the purpose of reducing the cost and increasing user acceptance and usability, the entire system uses only a smartphone to recover all the information required.

1. Introduction

According to Taipale [1], in 2014, there were more than 800 million people 60 years of age and over and the global number is increasing. The United Nations [2] indicated that 64 countries are expected to have an elderly population of more than 30% by 2050. Also, the Global Age Watch Index [3] shows that the number of people living alone is increasing and older people emerge as a growing market for consumption.

One of the aims of gerontechnology [4] is to extend the time during which elderly people can live independently in their preferred environment with the support of information and communication technologies [5], thus maximizing their vital and productive years and reducing the cost of care in later life. To achieve this goal, activity recognition is one of the main facilities of gerontechnology: real-time monitoring of human activities represents a useful tool for many purposes and applications such as daily activities assistance, health, and activity monitoring or safety and security enhancement [6]. Although activities of daily living (ADL) are useful to analyze user behavior, falls are the most important events that need to be detected. According to the World Health Organization [7], more than 28% of people aged 65 and over fall each year, increasing to more than 32% for those over 70 years. If preventive measures are not taken in the near future, the number of injuries caused by falls is projected to double by 2030. In this environment, assistive devices that contribute to reduce the

* Corresponding author.
E-mail addresses: maalvarez@us.es (M. Álvarez de la Concepción), lsoria@us.es (L.M. Soria Morillo), jaalvarez@us.es (J.A. Álvarez García), luisgon@us.es (L. González-Abril).

http://dx.doi.org/10.1016/j.pmcj.2016.05.002
1574-1192/

incidence of this kind of events are a social need. The automatic and unobtrusive identification of user's activities, including falls, is one of the challenging goals of context-aware computing [8,9] and it is a fast-growing field in ubiquitous computing domain. Indeed, it is expected that activity recognition systems will be a practical solution to monitor elderly people in the coming years. Although there are many mobile activity recognition systems, most of them lack battery draining [10], and in the last few years, developers have focused their efforts to tackle this problem. Liang [11] used a low sampling frequency with a hierarchical scheme methodology in order to improve the battery consumption. Weng [12] described a similar approach but the hierarchical support vector machine is supported by an additional strategy that reduces the sensor data sampling rates. This addition allowed the authors to reduce the computational complexity. On the other hand, Rault [13] proposed a new decision metric in order to evaluate these systems. It takes into account latency, accuracy, and energy consumption requirements in order to select the best execution configuration.

As an addition, some of the developed systems include automatic customization of the mobile device's behavior. For example, Kozina's system [14] sends calls directly to voicemail if a user is jogging and generates a daily/weekly activity profile to determine if a user is performing a healthy amount of exercise. Lu [15] implemented two simple proof-of-concept applications using a continuous sensing engine, JigMe, that automatically records a user's daily diary and, GreenSaw, that gives users awareness of their daily calorie expenditure and carbon footprint. These add-ons increase the value of the research and evidence that they could be used out of the laboratory by common users.

In this paper, we present a low-consuming battery system using the core of the Ameva algorithm [16] validated in the activity recognition track of EvAAL competition[1] in the editions of 2012 and 2013. The proposed system supports the inclusion of new activities once the training stage is completed, a feature that makes the system feasible for rehabilitation exercises recommended to elderly people by physicians. Furthermore, these new activities are automatically detected, inviting the user to perform a more exhaustive training on them if necessary. This last feature is especially interesting for unsupervised systems where users are free to perform any kind of activity, and not only for the elderly, whose set of activities tend to be more limited.

The remainder of this paper is structured as follows. Section 2 presents a review of activity recognition systems targeting elderly people. Section 3 introduces the EvAAL Activity Recognition competition and the results of our system. Section 4 presents the AMEVA recognition system and its evolution. A comparative analysis is described in Section 5. Finally, Section 6 concludes the paper and discusses future extensions.

2. Activity recognition systems for elders

Activity recognition systems (AR systems) have experienced an increase in both number and quality, mainly due to the growing interest in monitoring elderly people with dementia or people in rehabilitation. AR systems are classified into two categories: external sensor-based and wearable sensor-based. In external sensor-based systems, the devices are fixed in predetermined places. In wearable sensor based systems, the devices are attached to the user. Smart-home projects [17–19] include all kinds of sensors (temperature, smoke, humidity, presence, light and bed presence, NFC or RFID labels, etc.) but these systems have a pervasiveness issue: the only place where the activity is recognized is in the user's home or where the sensors are located. Another kind of research venue focuses on the usage of cameras for the recognition of gestures [20–22]. This is especially suitable for security (e.g. intrusion detection), but privacy issues [23] make this option unfeasible to recognize ADL. On the other hand, these systems can only be used in controlled environments. Robots are another kind of external sensor [24,25] that can assist the elderly, but the cost of deployment and maintenance of these systems is currently a big disadvantage. Furthermore, this kind of system presents a common drawback: people are not always monitored and hence some activities and events like falls could be unnoticed.

Wearable sensors are the preferred option for the latest generation of AR systems. Most solutions in this area employ various sensors placed in the body. Accelerometers are the most suitable option to detect movement, but accuracy improves when gyroscopes, magnetometers, and barometers are included in the system [26,27]. Smartphones, which embed all these sensors, can be considered a type of wearable due to their pervasiveness. Furthermore, the low adoption barrier on healthcare applications [28] through application markets such as Google Play or AppStore makes them the best option to target the mass market. Some of them are focused on fall detection [29,30], but normally do not cover both ADL and falls [31], so a classification system must be designed to consider them.

In general, the most important problem with classification models is that a good training process is needed to get the best results. Therefore, in AR systems, it is important to detect correctly the lifestyle of elderly people, but the difficulty is that they will waste their time training a device. This aspect makes the aged population a singular group for AR systems. Abdallah [32] developed a framework that incorporates incremental and active learning for real-time recognition and adaptation in streaming settings. However, the majority of existing solutions detect only a few activities. A major step forward would be the possibility for the system to recognize additional activities after the initial training [33,34].

All common learning processes have a test phase that normally is carried out in a laboratory setting. In AR, this stage is performed in a controlled environment, throughout public generated datasets and also by the developed team of the system itself. The presented approach differs from the traditional system because it has been proved effective also in a non-controlled environment as it is described in Section 3.

[1] Available from: http://evaal.aaloa.org.

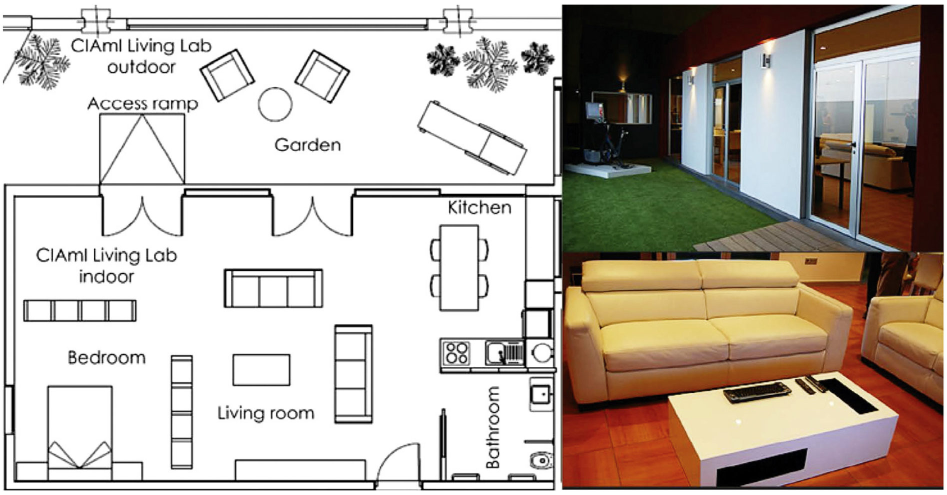


Fig. 1. Map from CIAmI Living Lab and pictures of garden and living room.

3. EvAAL activity recognition competition

The competition EvAAL (Evaluation of Ambient Assisted Living Systems through Competitive Benchmarking) was created with the idea of comparing and validating AAL solutions and platforms. It is an annual international contest that helps to measure the state of the art of AAL solutions by assessing the participants, level of autonomy, independent living, and quality of life they deliver to elders.

The developed algorithm was presented in the Activity Recognition track, which has as main objective the measurement of hardware and software performance through the established set of benchmarks and evaluation metrics. In order to participate in the AR track, the system must be capable of detecting the following activities: lie, sit, stand, walk, bend, fall (any kind of fall) and cycle (using a stationary bike). There is no limitation to the number of devices that can be used and competing solutions can be based on a variety of sensors and technologies.

The two editions of the track took place at the CIAmI Living Lab [35] in Valencia (Spain). Fig. 1 shows the CIAmI Living Lab.

3.1. Benchmark description

Once the participant team installs the solution, an actor (an evaluation committee member) performs a predefined physical activity trip across the smart home wearing an elderly simulation kit (Fig. 2). Álvarez-García [36] explains the complete process.

3.2. Evaluation criteria

The evaluation of the competing systems is carried out using five-evaluation criteria:

- **Accuracy**. Represents the confidence of the system under evaluation recognizing the activities performed by the actor. Accuracy is computed from the recognized activity instances using F-measure. $\frac{2*precision*recall}{precision+recall}$.
- **User acceptance**. Expresses how much the system is invasive in the user's daily life and hence the perceived impact. This parameter is estimated with a questionnaire that considers aspects of usability like the invasiveness, visibility of the installation within the environment and the complexity of maintenance procedures.
- **Recognition delay**. Measures the elapsed time between the instant when the user begins an activity and the time the system recognizes it.
- **Installation complexity**. It measures the effort required to install the recognition system in a home. It is measured as a function of the person-minutes of work needed to complete the installation.
- **Interoperability**. Measures how easy is to integrate the system with other systems. Interoperability is measured with a questionnaire that takes into account aspects like the availability of APIs and documentation, the licensing scheme, the presence of testing tools and the portability among different operating systems.

Fig. 2. Elderly simulation kit where elements 7, 8 and 9 were used to simulate the elderly's movements.

Table 1
Best performance results 2012.

Team	Accuracy	Delay	Installation	User acceptance	Interoperability	**Final score**
USS[a]	4.33	9.00	10.00	7.47	7.63	**7.39**
CMUU[b]	7.17	9.00	0.00	7.93	6.15	**6.50**
CUJ[c]	1.44	5.00	0.00	5.60	5.09	**3.52**

[a] University of Seville from Spain.
[b] Carnegie Mellon and Utah Universities from USA.
[c] Chiba University from Japan.

Table 2
Best performance results 2013.

Team	Accuracy	Delay	Installation	User acceptance	Interoperability	**Final score**
JSIS[a]	6.94	10.00	10.00	8.55	7.20	**8.36**
CNRI[b]	4.04	10.00	10.00	7.04	6.15	**6.94**
USS	4.68	9.00	10.00	6.99	5.54	**6.89**

[a] Jožef Stefan Institute from Slovenia.
[b] Consiglio Nazionale delle Ricerche from Italy.

3.3. Results

After peer review, only four competitors participated in the challenge in both editions where University of Seville from Spain and Chiba University from Japan teams repeated with improved versions of their solutions in 2013.

Table 1 shows the scores of the top three marked teams on the scale of 0–10 for the 2012 edition and Table 2 the ones for 2013 edition.

The winner team of 2012 edition (USS team) [37], composed of three of the authors of this paper, obtained acceptable results in accuracy (it was below that of the CMUU team), but its simplicity (although it uses multiple mathematical methods it only relies on accelerometers) and interoperability allowed it to achieve good marks in all the evaluated criteria.

The winner of the 2013 edition (JSIS team) [14,38] obtained very good results in all the evaluated criteria. The poor performance of the USS system in the 2013 edition is related to the lower priority associated to the activity "bend", penalized to give higher importance to ADL and fall. In addition, the Android mobile device was not properly secured on the right hip of the actor and it fell to the ground during the cycling activity penalizing the final result. In order to avoid this problem, which could be present in a daily use of this approach, next generations of the AR system will be installed into smart-watches. These devices allow to run the application and extract information about accelerometry and at the same time, increase the user acceptance and reduce the risk of forgetting. In the next section, the AR system and its improvements will be described.

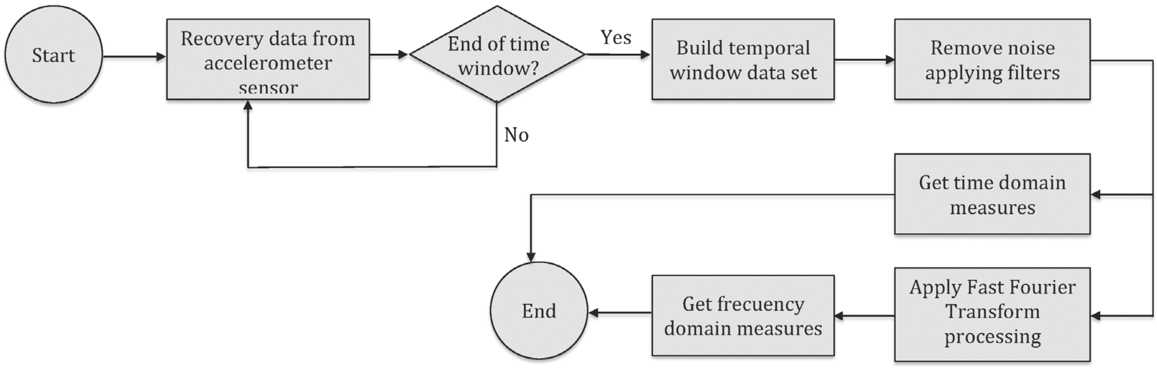


Fig. 3. Common steps for training and recognition process.

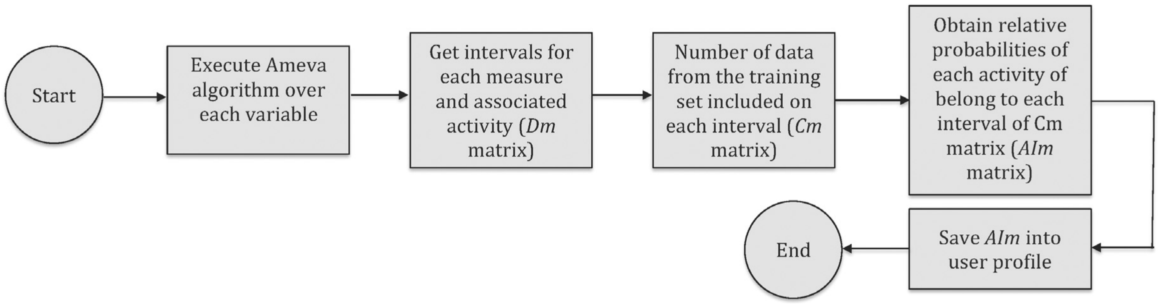


Fig. 4. Overall process of data analysis and interval determination.

4. The activity recognition algorithm

The initial steps of the activity recognition and fall detection system (from now on activity recognition) are depicted in Fig. 3. These steps are common in training and classification processes, and they are always executed before the recognition stage.

The process starts with collecting data from the accelerometer of the mobile device at an average frequency of 25 Hz. Time windows of fixed duration are used to get training data (and later to get recognition data). Each time window is composed of a set of accelerometer readings from which it is possible to calculate a variety of features. After performance and system accuracy analysis, it has been determined in an empirical way that the optimum length for these windows is five seconds.

Based on these time windows the module of the signal has been chosen in order to reduce the computational cost of the new solution and solve some problems related to the device's orientation. Using it, the accelerometer module measures acceleration values in three spatial directions (x, y, z) in the form $a_i = (a_{xi}, a_{yi}, a_{zi})$. The norm of the vector is computed as follows.

$$|a_i| = \sqrt{(a_{xi})^2 + (a_{yi})^2 + (a_{zi})^2}.$$

Finally, arithmetic mean, minimum, maximum, median, standard deviation, geometric mean and measures from frequency domain are generated. An important feature of the system is that the user can decide what activities must be recognized. This feature is critical for the application of the system to a specific scenario of interest. In our case the classes of the activities we wanted to recognize were immobile, walk, run, fall, drive, walk-upstairs, walk-downstairs and ride a bicycle.

Once the activities are defined, the training phase is required in order to recognize them. To get a training set, the user wears the smartphone doing repetitions for each activity to get personal information and training data. The number of examples for each activity must be balanced in order to avoid over training. All the activities are trained by the research team but in order to adapt the system to the user they must be trained again by said user.

Falls constitute an exception to this process and their training is performed via simulated falls. Walking, running, walking-upstairs and downstairs require only 20 s of training. Driving is an activity in which accelerations do not occur at a specific frequency so a small trip of 15 min is necessary to train the system. Finally riding a bicycle can be trained in three minutes.

Once the training set is prepared, the statistic process can be carried out using data analysis and interval determination. Fig. 4 shows the steps.

The first step of this process is the discretization of each variable in order to reduce the computational cost of the algorithm. This discretization process is based on the Ameva algorithm [16]. The aim of it is to maximize the dependency

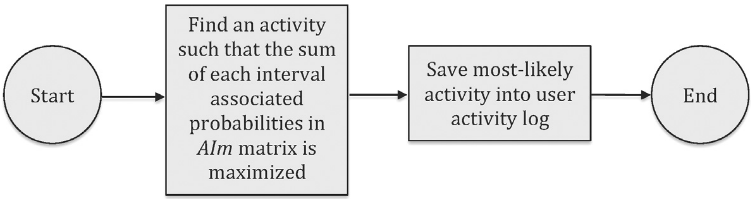


Fig. 5. Overall recognition process from data sensors.

relationship between the class labels $\mathcal{C}$ and the continuous attribute $\mathcal{L}(k)$. Furthermore, the lower the number of discrete intervals k, the faster the classification will be.

The application of the algorithm to each statistical value of the system allows obtaining a set of intervals associated with a particular $\mathcal{C}$ tag. Thus, after processing all system statistics, a matrix denoted by $Dm\{\mathcal{C}, L, S\}$ is produced as output.

The dimensions of the matrix are in the order:

1. the label $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_\ell\}, \ell \geq 2$ of the activity;
2. the interval $L = \{L_1, L_2, \dots, L_k\}$ where $L_i = (L_i^{low}, L_i^{up}], i = 0, 1, \dots, k$ defining lower L_i^{low} and upper limit L_i^{up} of the range;
3. the statistics measurement of the data (arithmetic mean, minimum, maximum, etc.).

After the data collection $\mathcal{X}$, the probability associated with the statistical data for each activity is computed. In order to carry out this process, a matrix denoted by $Cm\{x, L_i, S\}$ is defined as follows:

$$Cm_{x,i,s} = |x \in \mathcal{X}|, \quad x \geq L_i^{low} \wedge x < L_i^{up} \wedge x_{\mathcal{C}} = C_s.$$

It means that each entry of the matrix Cm contains the number of instances $x \in \mathcal{X}$ belongs to a specific interval of the range of a statistical S.

After Cm is computed, the relative probability matrix is carried out. This matrix is denoted by $AIm\{x, L_i, \mathcal{S}\}$ and it determines the likelihood that a given value x associated to an S statistic corresponds to a specific C_i activity. This ratio is based on the quality of the discretization performed by Ameva. The contents of the array AIm are defined as follows

$$AIm_{c,i,s} = \frac{Cm_{c,i,s}}{total_{c,s}} \cdot \frac{1}{l-1} \sum_{j=1, j\neq c}^{l} \left(1 - \frac{Cm_{j,i,s}}{total_{j,s}}\right)$$

where $total_{c,s}$ is the total number of time windows of the training process labeled with the c activity for the S statistic.

Finally, Fig. 5 shows the overall process described on this section for recognition process from matrix AIm where it is based on a majority voting system.

This process starts from the matrix AIm and uses a set of data $x \in \mathcal{X}$ for each of the statistics belonging to the set $\mathcal{S}$. The process consists of finding an activity $mpa \in \mathcal{C}$ such that the likelihood is maximized. The above rule is included in the expression

$$mpa(\mathcal{X}) = \max \sum_{s=1}^{s} AIm_{c,i,s} | x_s \in (L_i^{low}, L_i^{up}]. \tag{1}$$

The expression shows that the weight of each statistical metric to the calculation of the probability is equal. This can be carried out under the assumption that all statistics provide the same amount of information to the system, and that there is no correlation between them. Thus, the mpa represents those activities whose data is more fitted to the AIm set values using a majority voting system.

The final system with an innovative algorithm is developed and deployed in a smartphone to get the necessary data from the accelerometer and to identify the activity that user is doing.

4.1. Improvements of the system

The main advantage of the approach is the reduced battery consumption caused by the usage of discrete variables instead of continuous ones. Also, the dependencies between them are eliminated from the system to get only the information and to reduce the noise.

After the 2013 edition when the system achieved third position, several improvements over the original algorithm have been applied. The discretization process performed in the original algorithm needed to evaluate the cuts criterion using an iterative algorithm. Hence, its goal was to find the cut, which minimizes the variance of the class labels belonging to instances of each interval. The application of the algorithm to a multivariate dataset with a high amount of samples caused the time complexity of Ameva to explode and, therefore, the low performances.

The optimization introduced allowed significant simplification. The main advantage consists in the need to compute the sample variance of the class label associated with each instance only once. This optimization generates a higher number

Table 3
Comparison between Ameva original and Ameva optimized algorithms.

# of statistics	Ameva original time (s)	Ameva optimized time (s)	Ameva original # of intervals	Ameva optimized # of intervals
1	0.53	0.13	6	12
5	2.80	0.75	36	63
10	4.31	1.22	60	101
20	7.77	2.56	120	195
50	22.13	6.30	300	555
100	43.08	12.69	600	1097
200	90.13	25.13	1200	2168
400	167.84	50.00	2400	4074

of intervals than the original algorithm. Moreover, the intervals generated by the original algorithm are a subset of those generated by the optimization, which ensures that the results of the classification process are theoretically a superset of the ones carried out by the originals, but in practice are really close to them. In terms of runtime duration, a significant reduction was expected by the application of the improvement described. These expectations were met in the testing process, where execution time was reduced by 70%. Table 3 shows the comparison between both algorithm (Ameva original and Ameva optimized) in terms of execution time and it has obtained from a dataset of 10,299 instances and 561 statistical.

4.2. Definition of the used datasets

The gathering data process was carried out with a group of 30 volunteers within an age bracket of 19–48 years wearing a smartphone on the waist and was randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data [39]. Also, there were other datasets that were tested in the system:

- USC-HAD [40]: the data was collected from a set of human subjects that have a mean of 30.1 years old, 170 cm height and 64.6 kg weight, doing 12 activities. To overcome the limitations of the existing datasets, thus serving as a standard benchmark for researchers in the ubiquitous computing community to compare performance of their human activity recognition algorithms.
- WISDM [41]: it contains data from 36 users and six activities. Test data was collected from an extensive variety of Android smartphones (1.3–2.1 version), like the system presented here.
- Shoaib [42]: this dataset contains smartphone sensor data for six physical activities collected using four participants. It was useful because the data was collected from four smartphones on four body positions, allowing for comparison to approaches using only one smartphone.

Furthermore, the validation of results was not only tested on these datasets by cross-validation, but competition data were used to check their suitability for a real environment with elderly people.

4.3. Learning process

Another problem in AR systems is the learning process for new activities. There are some scenarios, such as inexperienced users or elderly people usage, in which training is not performed in an accurate way, so learning process is reduced or non-existing. The learning process is thus reduced or non-existent. Although there are expert systems that do not require learning processes, they suffer from low accuracy and reduced user adaptation. This was the reason why the accuracy values during the competition were lower than those obtained during the validation. On the other hand, the lack of training causes the set of recognized activities to be the same for every user, taking out the customization options of the system. This is the case of most commercial applications such as Moves [43] or Fitbit [44] among others.

The presented approach allows not just the training phase, critical for getting an acceptable accuracy and high customization capabilities, but also the recognition of new activities without user supervision. Thanks to the far distance activities detector (this is the name given to the module within the recognition system), the algorithm is able to automatically recognize new activities based on the total score of the current time window obtained by the recognition process.

The learning is performed for each particular user, so it is adapted to the way in which the user carries out the activities. For elderly, it is also personalized, but the classification of the accelerometric profile and all statistics generated are fitted to the user.

In this stage, the system obtains the probability that a value belongs to each of the recognized activities maximizing the value computed in (1). However, if the value of the *mpa* is too low, it would be an indicator that the activity is not taken into account when conducting the training. This fact allows the far distance activities detector to generate an alert of new activity detected. The threshold value is set empirically, although experiments had determined that the value, which maximized the accuracy of the system, is equal to the number of statistics divided by four. This entails that, overall, the joint probability that a given time window is classified correctly as the activity is below 25%.

Falls are considered a special kind of activity. This characterization is because the training of the system under these circumstances is not possible. Although there are studies that identify these events dynamically [45], the learning process is

Table 4
AIm matrix for standard deviation with six discretization intervals and six activities.

Interval	Activity					
	Walking	Upstairs	Downstairs	Sitting	Standing	Lying
$L_{p,1}$	0.00	0.00	0.00	0.42	0.48	0.10
$L_{p,2}$	0.00	0.00	0.00	0.35	0.31	0.34
$L_{p,3}$	0.00	0.00	0.00	0.25	0.24	0.51
$L_{p,4}$	0.61	0.36	0.02	0.00	0.00	0.01
$L_{p,5}$	0.30	0.49	0.21	0.00	0.00	0.00
$L_{p,6}$	0.02	0.07	0.92	0.00	0.00	0.00

Table 5
Confusion matrix of the complete set of activities (percentage).

		Predicted class							
		Walk	Fall	Stop	Run	Up	Down	Cycle	Drive
Actual class	Walk	**93.50**	0.54	0.54	0.54	0.72	0.36	0.72	3.07
	Fall	0.20	**98.00**	0.00	0.40	0.20	0.80	0.20	0.20
	Stop	0.00	0.00	**96.60**	0.00	0.00	0.54	0.72	2.15
	Run	0.43	1.28	0.00	**96.36**	0.86	0.21	0.43	0.43
	Up	0.21	0.21	1.05	0.21	**96.00**	1.47	0.63	0.21
	Down	0.98	0.73	1.22	0.73	2.20	**91.95**	0.24	1.95
	Cycle	0.23	0.23	0.00	0.23	0.46	0.23	**97.91**	0.70
	Drive	0.47	0.24	4.25	0.24	0.47	0.24	0.47	**93.63**

usually performed on a synthetic dataset, due to the great handicap of generating data of real falls. Therefore, in this paper the problem of falling is addressed in parallel, defining an accelerometric profile criteria and a period of no accelerations. By this method, the system detects a fall when the accelerometry profile shows acceleration peaks followed by a long period of inactivity. During the execution, this period was set to five seconds.

Table 4 shows a set of different values obtained for each of the positions of the *AIm* matrix.

In order to show the quality of the system with the considered set of activities, a confusion matrix with the complete set of activities is presented in Table 5. These data have been collected from a test case where a 31 year old researcher simulated the movements and two falls (forward and backward) of an elderly person during 15 min.

These scores were obtained during the training phase from the datasets described above and mean that the likelihood for a given value x associated to an S statistic corresponds to the right activity C_i; i.e. this indicator reflects the probability that a given configuration of the input parameters belongs to the activity C_i based on the intervals L_i generated from the *AIm* matrix for all variables obtained in the dataset using the maximum likelihood method.

5. Comparison with other systems

As mentioned in Section 1, the number of research studies and applications of the activity recognition has increased in several contexts in recent years. The insightful practical implications include elderly care, quality of self-care estimation and monitoring daily live activities among others.

Lara [8] studied a deep evaluation of some existing activity recognition systems taking into account the type of sensors and the measured attributes, the integration device, the level of obtrusiveness, the type of data collection protocol, the level of energy consumption, the classifier flexibility level, the feature extraction method(s), the learning algorithm(s) and the overall accuracy for all activities. Also, they were divided into online and supervised offline systems.

Shoaib [46] used five positions using a four motion sensors smartphone collecting data in some statistics in order to detect seven physical activities through a time window approach. The study consists in which are the best places for the smartphone using a combination of them and analyzing the accuracy obtained by the most widely used classification algorithms.

Based on their evaluations, they show that both the accelerometer and the gyroscope are capable of taking the lead roles in the activity recognition process, depending on the type of activity being recognized, the body position, the classification method and the feature set being used. However, the authors do not offer a new approach, but a general analysis.

Plasqui's main goal [47] was to review all recent validation studies of accelerometers against doubly labeled water in order to guide researchers in their selection of an appropriate accelerometer for a specified research goal.

They concluded that the best-wearing position for an accelerometer to assess daily life physical activity is as close as possible to the center of mass, hence the lower back or hip. Only the differentiation between standing and sitting could not be achieved with a single accelerometer at this position.

Finally, Ellis [48] used two accelerometers (right hip and non-dominant wrist) and a GPS in order to monitor four activities of 40 overweight and obese breast cancer survivors. They use a two-step process to determine the activity. The first step is performed by a low-level classifier using a random forest classifier over the combination of GPS and accelerometer features. The output of each decision tree in the forest is combined using majority voting to obtain a prediction. The second

Table 6
Comparison with other activity recognition systems. ND (not determined). (*) This value of battery life has been obtained directly from the author's paper assertions.

Method	Number of activities	Average accuracy (%)	# of sensors	Execution environment	Battery life (h) (*)
Our proposal (2015)	9 (extensible)	95	1 general	Smartphone	18.0 (measured)
Antos et al. [10] (2014)	5	88	1 general	Smartphone	ND
Liang et al. [11] (2014)	11	85	1 general	Smartphone	3.2 (measured)
Weng et al. [12] (2014)	4	98	1 general	Smartphone	ND
Kozina et al. [14] (2013)	7	ND	2 general	Smartphone	ND
Shoaib et al. [46] (2014)	7	ND	5 general	Smartphone	ND
Ellis et al. [48] (2014)	4	85	2 general	ND	ND
Sasank Reddy [52] (2010)	5	93	8 general	Phone	8.2 (measured)
Hong Lu [15] (2010)	5	94	1 general	Smartphone	14.0 (experimental)
Vijay Srinivasan [53] (2012)	6	91	1 general	Smartphone	12.5 (measured)
Yi Wang [54] (2012)	2	90	1 general	Phone	150.0 (experimental)
Jia [55] (2013)	7	98	2 specific	Smartphone	7.0 (measured)

level classifier is a Hidden Markov model. Each hidden state belongs to one of the activities. The results of their activity classification system used leave-one-subject-out cross validation and the overall accuracy was 85.6%. Although the accuracy is very close to the approach presented here, it uses two devices and only detects four activities.

For the online systems (which are the focus of our approach), three of them get a high accuracy (over 94%) but have some disadvantages. Ermes [49] only applied a subject-dependent evaluation. Besides, their data were collected from only three subjects, which inhibits flexibility to support new users. eWatch [50], which embeds four sensors and a microcontroller within a device that can be worn as a watch for sport uses, is very energy efficient. The execution time for the feature extraction and the classification stage is lower than 0.3 ms. However, data was collected under controlled conditions, i.e., a lead experimenter supervised and gave specific guidelines to the subjects on how to perform the activities. Kao [51] presented a triaxial accelerometer placed on the user's dominant wrist, sampling at 100 Hz. The system reports an average response time of less than 10 ms. However, given the nature of the recognized activities, the excess of granularity causes confusion, among others, between swinging, knocking and running.

Table 6 shows different related works in the field of activity recognition. In this comparison, five features were evaluated. Number of activities shows the total amount of physical activities recognized by the system. The average accuracy indicates the performance of recognition in terms of activities properly classified compared with total instances. The number of devices (smartphones, sensors, motes, etc.) used during the recognition and training process is collected in # of sensors. Hardware configuration where the system is executed is presented under the execution environments feature. Finally, battery life shows the total time during which the execution environment is working until it runs out of battery.

It should be noted that the studies analyzed are minimal when compared to the amount of research found on activity recognition. This is because we only want to take into account studies that have been tested or have the computational requirements to be tested in mobile devices.

6. Conclusions and future work

Efficiency and accuracy are two elements that must be taken into account when any AR system is implemented on a mobile device and, more importantly, when this mobile device is a smartphone. In this work, a recognition system based on discrete variables is presented whereby the discretization algorithm Ameva and a new classification system are used. It has a low complexity and both the runtime and energy consumption have been reduced in comparison to other related works. The system has been validated in an international competition (1st and 3rd positions). Although the accuracy was not very good in the 2012 and the 2013 EvAAL competitions, the system is very usable and easy to introduce in lifestyle of elderly people. Several improvements have been carried out and described in this paper.

The classification algorithm has been developed for multi-class datasets and it obtains a good accuracy when there is approximately the same number of examples for each class in the training phase. It is also fast because it is based on the discretization algorithm Ameva and a majority voting system which both have a very low processing time. This makes it possible to embed the system into tiny pervasive hardware such as smartwatches or specific devices attached to the user's clothing or body. Furthermore, although it will be tested with other datasets, the core of this algorithm remains free of dependence on the features of any recognition activity dataset and is, therefore, applicable to any dataset that contains activities with different behavior patterns (for example, walk and stand).

The advantages of the system are the high accuracy rate and the reduced computational cost as has been demonstrated in the experimental results. Regarding the success rate, it has been possible to achieve an average accuracy of 95% in the recognition of eight different types of activities with a group of 30 volunteers. On the other hand, the complexity associated with the data processing during the recognition process has been optimized due to the inclusion of discrete variables and 24 h of continuous monitoring have been reached without recharging the device.

Based on the results obtained during the process of experimentation, the most likely system confusions occur when the acceleration associated with the activity is high. In the case of falls, it could give false positives if the device is thrown on a surface or when the user performs actions involving high accelerations followed by an inactivity period (let yourself fall on a chair or a sofa). This approach is not really bad because it reduces the number of the false negatives especially when the most critical activity is the falling [56,57].

Finally, the case of picking something up or similar activities, would not be a problem for the system. These activities could lead to a high acceleration during its performance (not as great as those generated during a fall, although similar), but there is no inactivity period after its execution.

As for interference caused by the frequency of carrying out the activities, their consequences are mild. In the case of walking at a high enough rate, the system could interpret the user is running. Or if you jump fast enough and with a not too high altitude, you might consider running is the most likely activity. However, even looking directly at the user, these activities can lead to confusion because of their similarity.

This system is currently focused on ADL activities and fall detection of elderly people. Due to that, the application developed to execute this AR system can transmit an alarm signal to the relatives and/or medical center. Besides this, relatives and medical staff can examine the activities performed by the elderly under their monitoring along the day in a website, improving the system functionality beyond an alarm system. It complements existing telecare services, such as those as offered by the Andalusian Regional Ministry of Equality, Health and Social Policy among others.

Finally, in order to improve the accuracy and to track rehabilitation process by means of a fine-grained AR system, the Myo device[2] that is a gesture control armband will be included in the solution. It detects the hand and wrist motions throughout high sensitivity electromyogram sensors, and it could help caregivers to obtain more information about the lifestyle of their patients. For instance, it could be possible to retrieve information about the use of the walking stick in daily life or during rehabilitation process. This hardware will also allow the detection of important aspects such as dyskinesia (for Parkinson's patients), fatigue or other muscular pathologies.

Acknowledgments

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness HERMES (TIN2013-46801-C4-1-r), Simon (P11-TIC-8052) and M-Learning (P11-TIC-7124) of the Andalusian Regional Ministry of Economy, Innovation and Science.

References

[1] V. Taipale, Global trends, policies and gerontechnology, Gerontechnology 12 (4) (2014) 187–193.
[2] Fund UNP, International HA, Ageing in the twenty-first century: a celebration and a challenge. United Nations Population FundHelp Age New York, London, 2012.
[3] Index GA, Insight Report, London, UK: HelpAge International. Retrieved from http://www.helpage.org, 2014.
[4] J.E. van Bronswijk, H. Bouma, J.L. Fozard, W.D. Kearns, G.C. Davison, P.C. Tuan, Defining gerontechnology for R&D purposes, Gerontechnology 8 (1) (2009) 3–10.
[5] N. Agarwal, M. Sebastian, Wireless infrastructure setup strategies for healthcare, in: Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, ACM, 2014, p. 66.
[6] O. Yurur, C. Liu, W. Moreno, A survey of context-aware middleware designs for human activity recognition, IEEE Commun. Mag. 52 (6) (2014) 24–31.
[7] Ageing WHO, Unit LC, WHO Global Report on Falls Prevention in Older Age, World Health Organization, 2008.
[8] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2013) 1192–1209.
[9] H. Gjoreski, Context-based reasoning in ambient intelligence (Ph.D. thesis), IPS Jožef Stefan, Ljubljana, Slovenia, 2015.
[10] S.A. Antos, M.V. Albert, K.P. Kording, Hand, belt, pocket or bag: Practical activity tracking with mobile phones, J. Neurosci. Methods 231 (2014) 22–30.
[11] Y. Liang, X. Zhou, Z. Yu, B. Guo, Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare, Mob. Netw. Appl. 19 (3) (2014) 303–317.
[12] S. Weng, L. Xiang, W. Tang, H. Yang, L. Zheng, H. Lu, et al., A low power and high accuracy MEMS sensor based activity recognition algorithm, in: Bioinformatics and Biomedicine, BIBM, 2014 IEEE International Conference on, IEEE, 2014, pp. 33–38.
[13] T. Rault, A. Bouabdallah, Y. Challal, F. Marin, Context-aware energy-efficient wireless sensor architecture for body activity recognition, in: Pervasive Computing and Communications Workshops, PERCOM Workshops, 2014 IEEE International Conference on, IEEE, 2014, pp. 203–206.
[14] S. Kozina, H. Gjoreski, M. Gams, M. Luštrek, Efficient activity recognition and fall detection using accelerometers, in: Evaluating AAL Systems Through Competitive Benchmarking, Springer, 2013, pp. 13–23.
[15] H. Lu, J. Yang, Z. Liu, N.D. Lane, T. Choudhury, A.T. Campbell, The Jigsaw continuous sensing engine for mobile phone applications, in: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, ACM, 2010, pp. 71–84.
[16] L. Gonzalez-Abril, F.J. Cuberos, F. Velasco, J.A. Ortega, Ameva: An autonomous discretization algorithm, Expert Syst. Appl. 36 (3) (2009) 5327–5332.
[17] N. Costa, P. Domingues, F. Fdez-Riverola, A. Pereira, A mobile virtual butler to bridge the gap between users and ambient assisted living: a smart home case study, Sensors 14 (8) (2014) 14302–14329.
[18] S. Azzi, C. Dallaire, A. Bouzouane, B. Bouchard, S. Giroux, Human activity recognition in big data smart home context, in: Big Data (Big Data), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–8.
[19] Q. Zhang, Y. Su, P. Yu, Assisting an elderly with early dementia using wireless sensors data in smarter safer home, in: Service Science and Knowledge Innovation, Springer, 2014, pp. 398–404.
[20] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comput. Vis. Image Underst. 117 (6) (2013) 633–659.
[21] C. Rougier, J. Meunier, A. St-Arnaud, J. Rousseau, 3D head tracking for fall detection using a single calibrated camera, Image Vis. Comput. 31 (3) (2013) 246–254.
[22] G. Srivastava, J. Park, A.C. Kak, B. Tamersoy, J. Aggarwal, Multi-camera human action recognition, in: Computer Vision, Springer, 2014, pp. 501–511.

[2] Available from https://www.thalmic.com/en/myo/.

[23] M. Kanis, S. Robben, J. Hagen, A. Bimmerman, N. Wagelaar, B. Krose, Sensor monitoring in the home: giving voice to elderly people, in: Pervasive Computing Technologies for Healthcare, PervasiveHealth, 2013 7th International Conference on, IEEE, 2013, pp. 97–100.
[24] Yh Wu, J. Wrobel, M. Cornuet, H. Kerhervé, S. Damnée, A.S. Rigaud, Acceptance of an assistive robot in older adults: a mixed-method study of human–robot interaction over a 1-month period in the living lab setting, Clin. Intervent. Aging (2014) 9.
[25] Y.H. Wu, J. Wrobel, V. Cristancho-Lacroix, L. Kamali, M. Chetouani, D. Duhaut, et al., Designing an assistive robot for older adults: The ROBADOM project, IRBM 34 (2) (2013) 119–123.
[26] P. Dell'Acqua, L.V. Klompstra, T. Jaarsma, A. Samini, An assistive tool for monitoring physical activities in older adults, in: Serious Games and Applications for Health, SeGAH, 2013 IEEE 2nd International Conference on, IEEE, 2013, pp. 1–6.
[27] L. Hu, Y. Chen, S. Wang, Z. Chen, b-COELM: A fast, lightweight and accurate activity recognition model for mini-wearable devices, Pervasive Mob. Comput. 15 (2014) 200–214.
[28] R. Rana, M. Hume, J. Reilly, R. Jurdak, J. Soar, Transforming knowledge capture in healthcare: Opportunistic and Context-aware affect Sensing on Smartphones. arXiv preprint arXiv:150202796, 2015.
[29] M.A. Habib, M.S. Mohktar, S.B. Kamaruzzaman, K.S. Lim, T.M. Pin, F. Ibrahim, Smartphone-based solutions for fall detection and prevention: challenges and open issues, Sensors 14 (4) (2014) 7181–7208.
[30] P.S. Jeong, YH. Cho, Fall detection system using smartphone for mobile healthcare, J. Korea Soc. IT Serv. 12 (4) (2013) 435–447.
[31] J.A. Álvarez-García, L.M.S. Morillo, M.Á.Á. de La Concepción, A. Fernández-Montes, J.A.O. Ramírez, Evaluating wearable activity recognition and fall detection systems, in: 6th European Conference of the International Federation for Medical and Biological Engineering, Springer, 2015, pp. 653–656.
[32] Z.S. Abdallah, M.M. Gaber, B. Srinivasan, S. Krishnaswamy, Adaptive mobile activity recognition system with evolving data streams, Neurocomputing 150 (2015) 304–317.
[33] Z. Zhao, Z. Chen, Y. Chen, S. Wang, H. Wang, A class incremental extreme learning machine for activity recognition, Cogn. Comput. 6 (3) (2014) 423–431.
[34] M.Á.Á. de la Concepción, L.S. Morillo, L. Gonzalez-Abril, J.O. Ramírez, Discrete techniques applied to low-energy mobile human activity recognition. A new approach, Expert. Syst. Appl. 41 (14) (2014) 6138–6146.
[35] A. Martínez, M. Llorente, J.P. Lázaro, Ciami living lab: an economically sustainable technological tool for open innovation, I ENoLL Living Lab Summer School, 2010.
[36] J.A. Álvarez-García, P. Barsocchi, S. Chessa, D. Salvi, Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 EvAAL competition, J. Ambient Intelli. Smart Environ. 5 (1) (2013) 119–132.
[37] M.Á.Á. de la Concepción, L.M.S. Morillo, L.G. Abril, J.A.O. Ramírez, Activity recognition system using non-intrusive devices through a complementary technique based on discrete methods, in: Evaluating AAL Systems Through Competitive Benchmarking, Springer, 2013, pp. 36–47.
[38] H. Gjoreski, S. Kozina, M. Gams, M. Lustrek, J.A. Álvarez-García, J.H. Hong, et al., Competitive live evaluations of activity-recognition systems, IEEE Pervasive Comput. 14 (1) (2015) 70–77.
[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, A Public Domain Dataset for Human Activity Recognition using Smartphones, in: ESANN, 2013.
[40] M. Zhang, A.A. Sawchuk, USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors, in: ACM International Conference on Ubiquitous Computing, Ubicomp, Workshop on Situation, Activity and Goal Awareness (SAGAware), Pittsburgh, Pennsylvania, USA, 2012.
[41] G.M. Weiss, J.W. Lockhart, The impact of personalization on smartphone-based activity recognition, in: AAAI Workshop on Activity Context Representation: Techniques and Languages, 2012.
[42] M. Shoaib, H. Scholten, P.J. Havinga, Towards physical activity recognition using smartphone sensors, in: Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing, UIC/ATC, IEEE, 2013, pp. 80–87.
[43] F. Guo, Y. Li, M.S. Kankanhalli, M.S. Brown, An evaluation of wearable activity monitoring devices, in: Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia, ACM, 2013, pp. 31–34.
[44] M. Carman, N. Tam, M. Woodward, Why Fitbit has succeeded when other pedometers have failed, 2013.
[45] F. Bagalà, C. Becker, A. Cappello, L. Chiari, K. Aminian, J.M. Hausdorff, et al., Evaluation of accelerometer-based fall detection algorithms on real-world falls, PLoS One 7 (5) (2012) e37062.
[46] M. Shoaib, S. Bosch, O.D. Incel, H. Scholten, PJ. Havinga, Fusion of smartphone motion sensors for physical activity recognition, Sensors 14 (6) (2014) 10146–10176.
[47] G. Plasqui, A. Bonomi, K. Westerterp, Daily physical activity assessment with accelerometers: new insights and validation studies, Obes. Rev. 14 (6) (2013) 451–462.
[48] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, Multi-sensor physical activity recognition in free-living, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM, 2014, pp. 431–440.
[49] M. Ermes, J. PÄrkkÄ, J. MÄntyjÄrvi, I. Korhonen, Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions, IEEE Trans. Inf. Technol. Biomed. 12 (1) (2008) 20–26.
[50] U. Maurer, A. Rowe, A. Smailagic, D. Siewiorek, Location and activity recognition using eWatch: A wearable sensor platform, in: Ambient Intelligence in Everyday Life: Foreword by Emile Aarts, Springer, Berlin, Heidelberg, 2006, pp. 86–102.
[51] T.P. Kao, C.W. Lin, J.S. Wang, Development of a portable activity detector for daily activity recognition, in: 2009 IEEE International Symposium on Industrial Electronics, 2009, pp. 115–120.
[52] B. Longstaff, S. Reddy, D. Estrin, Improving activity classification for health applications on mobile devices using active and semi-supervised learning, in: Pervasive Computing Technologies for Healthcare, PervasiveHealth, 2010 4th International Conference on-NO PERMISSIONS, IEEE, 2010, pp. 1–7.
[53] V. Srinivasan, T. Phan, An accurate two-tier classifier for efficient duty-cycling of smartphone activity recognition systems, in: Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones, ACM, 2012, p. 11.
[54] Y. Wang, B. Krishnamachari, M. Annavaram, Semi-Markov state estimation and policy optimization for energy efficient mobile sensing, in: Sensor, Mesh and Ad Hoc Communications and Networks, SECON, 2012 9th Annual IEEE Communications Society Conference on, IEEE, 2012, pp. 533–541.
[55] R. Jia, B. Liu, Human daily activity recognition by fusing accelerometer and multi-lead ECG data, in: Signal Processing, Communication and Computing, ICSPCC, 2013 IEEE International Conference on, IEEE, 2013, pp. 1–4.
[56] Q. Li, J.A. Stankovic, M.A. Hanson, A.T. Barth, J. Lach, G. Zhou, Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information, in: Wearable and Implantable Body Sensor Networks, 2009. BSN 2009, Sixth International Workshop on, IEEE, 2009, pp. 138–143.
[57] A. Sixsmith, N. Johnson, A smart sensor to detect the falls of the elderly, IEEE Pervasive Comput. 3 (2) (2004) 42–47.

PART III

Other relevant research works

CHAPTER 7

A QUALITATIVE METHODOLOGY TO REDUCE FEATURES IN CLASSIFICATION PROBLEMS

Overview

This paper develops a preliminary feature reduction technique using the Ameva algorithm that has been confirmed as one of the most promising algorithms due to its reduced execution time and the smaller number of intervals provided.

When data is obtained experimentally, is not considered what features are relevant for the studied system and the application of several techniques that determines which are relevant for the system is a common step. The methodology presented in this paper uses the Ameva algorithm because it is based on the statistic χ^2 that determines the relationship between features and classes. For this reason, it is possible to use this algorithm to determine the relationship between features.

The core of the algorithm is a new coefficient called entropy that has been developed to determine the dependence between features. This approach can be satisfactorily apply in the feature reduction area when the data set has a lot of instances and features, and one of these features determines the class which each instance belongs.

Context

This research was originally started in the feature reduction area as a new possible use of the Ameva algorithm. The idea took as a principle the advantages of the χ^2 statistic that determines the correlation between two features and proposes the ones with the highest Ameva coefficient value as the features that could be removed. This paper is the result of over 8 months' work and currently is one of the areas under investigation.

Research work information

The *International Workshop on Qualitative Reasoning* is focus on Qualitative Reasoning that is a research area at the interface of Artificial Intelligence, Cognitive Science and Engineering Sciences. This workshop is always co-located with the International Joint Conference on Artificial Intelligence (IJCAI).

Link: http://www.qrg.northwestern.edu/papers/Files/qr-workshops/QR2011/2011_Proceedings/qr11_paper1.pdf

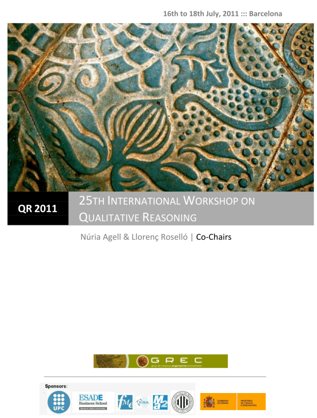


Figure 7.1: Qualitative Reasoning cover.

A qualitative methodology to reduce features in classification problems

Alvarez, M.A.[a]**; Gonzalez-Abril, L.**[b]**; Ortega, J.A.**[a]**; Soria, L.M.**[a]
[a]Department of Computer Science, University of Seville
[b]Department of Applied Economics I, University of Seville
Seville, Spain
{maalvarez, luisgon, jortega, lsoria}@us.es

Abstract

In this paper, a preliminary methodology which quantifies the dependence between features in a data set by using the Ameva discretization algorithm and the advantages of a qualitative model is developed. Thus, different matrices of interdependence are built providing a grade of dependence between two features. This methodology is applied to a well-known data set, obtaining promising results for the carried out system.

1 Introduction

The problem of classification is one of the main problems in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of features. The induction of classifiers from data sets of classified instances is a central problem in machine learning. For that purpose, a large number of methodologies based on SVM [1], Naive Bayesian [2], C5.0 [3], etc. have been developed.

Additionally, qualitative modeling and reasoning is a very interesting area for applying and experimenting with machine learning techniques. Qualitative reasoning has special interest to systems where machine learning can be applied as modeling, diagnosis, control, discovery, design, and knowledge compilation.

One of the most important preprocess in classification is the discretization. This process establishes a relationship between continuous variables and their discrete transformation through functions. Therefore, it is possible to model qualitatively a series of continuous values if a label is assigned to them. Some studies [4] have shown that execute a prior process to discretize continuous features is more efficient than work directly with the continuous values. This process reduces the computation time and memory usage in the application of classification algorithms and it is used to manage the set of values of a feature more effectively. Some relevant discretization methods are Ameva [5], Khiops [6], CAIM [7] and others [8; 9].

The Ameva discretization method has been confirmed as one of the most promising algorithms due to its reduced execution time and the smaller number of intervals provided. This behavior is outstanding when the data set has a large number of classes, although it has a slight reduction in the capacity of identification [5; 10].

Another problem in the classification process is the existence of irrelevant features [11]. When data is obtained experimentally, is not considered what features are relevant for the studied system. Several techniques [12; 13; 14] have been developed to reduce the number of features and to determine which are relevant for the system. Some of these techniques are based on principals components analysis [15] or factorial analysis [16].

The Ameva discretization algorithm [10] performs the discretization process effectively and quickly, so the set of values of a feature is greatly reduced, but do not reduce the number of features. Because Ameva uses the statistic χ^2 to determine the relationship between features and classes, it is possible to use this algorithm to determine the relationship between features.

In this paper, a new methodology based on Ameva algorithm is developed in order to reduce the number of features of a data set. This method exploits the advantages of Ameva in runtime and brings a different approach which was developed on.

The rest of this paper is organized as follows: first, the definition of the problem is presented in Section 2 to establish the notation of the rest of the paper. Also, the Ameva discretization algorithm and the Entropy coefficient are presented. Section 3 presents the new methodology to determine the dependence between features using the Ameva algorithm and the entropy coefficient. Section 4 reports the obtained results of applying the methodology in a toy example. The paper is finally concluded with a summary of the most important points and future works.

2 Discretization

Let $X = \{x_1, x_2, \dots, x_N\}$ be a data set of a continuous attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of ℓ classes of the variable denoted by

$$\mathcal{C} = \{C_1, C_2, \dots, C_\ell\}, \quad \ell \geq 2 \tag{1}$$

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \rightarrow \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of the property that is being discretized.

Let us consider a discretization $\mathcal{D}$ which discretizes $\mathcal{X}$ into k discrete intervals:

$$\mathcal{L}(k; X; \mathcal{C}) = \{L_1, L_2, \cdots, L_k\}$$

where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2, 3, \ldots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; X; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, k$. The discretization variable $\mathcal{L}(k)$ of attribute $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view. Having two discrete attributes, a two-dimensional frequency table (called contingency table) as shown in the Table 1 can be built.

$C_i \vert L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell k}$	$n_{\ell\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

Table 1: Contingency table

In Table 1, n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j. $n_{i\cdot}$ is the total number of instances belonging to the class C_i, and n_j is the total number of instances that belong to the interval L_j, for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$. So that:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, \quad N = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij}$$

2.1 The Ameva discretization

Given discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$, the contingency coefficient, denoted by $\chi^2(k) \stackrel{def}{=} \chi^2(\mathcal{L}(k), \mathcal{C}|X)$, defined as

$$\chi^2(k) = N \left(-1 + \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} \right) \tag{2}$$

is considered. It is straightforward to prove that

$$\max_{X, \mathcal{L}(k), \mathcal{C}} \chi^2(k) = N(\min\{\ell, k\} - 1) \tag{3}$$

Hence, the Ameva coefficient, $Ameva(k) \stackrel{def}{=} Ameva(\mathcal{L}(k), \mathcal{C}|X)$, is defined as follows:

$$Ameva(k) = \frac{\chi^2(k)}{k(\ell - 1)} \tag{4}$$

for $k, \ell \geq 2$. The Ameva criterion has the following properties:

- The minimum value of $Ameva(k)$ is 0 and when this value is achieved then both discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are statistically independent and viceversa.
- The maximum value of $Ameva(k)$ indicates the best correlation between class labels and discrete intervals. If $k \geq \ell$ then, for all $x \in C_i$ a unique j_0 exists such that $x \in L_{j0}$ (remaining intervals $(k - \ell)$ have no elements); and if $k < \ell$ then, for all $x \in L_j$, a unique i_0 exists such that $x \in C_{i0}$ (remaining classes have no elements) i.e. the highest value of the Ameva coefficient is achieved when all values within a particular interval belong to the same associated class for each interval.
- The aggregated value is divided by the number of intervals k, hence the criterion favors discretization schemes with the lowest number of intervals.
- From (3), it is followed that $Ameva_{max}(k) \stackrel{def}{=} \max_{X, \mathcal{L}(k), \mathcal{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ if $k < \ell$ and $\frac{N}{k}$ otherwise. Hence, $Ameva_{max}(k)$ is an increasing function of k if $k \leq \ell$, and a decreasing function of k if $k > \ell$. Therefore, $\max_{k \geq 2} Ameva_{max}(k) = Ameva_{max}(\ell)$ i.e. the maximum of the Ameva coefficient is achieved in the optimal situation, it is to say, when all values of C_i are in a unique interval L_j and viceversa.

Therefore, the aim of the Ameva method is to maximize the dependence relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k.

2.2 The entropy

If $\ell = 1$ or $k = 1$ then it is not possible to use the Ameva method. Let us see these two cases (see Table 2 and Table 3): Equation (2) can not be calculated using Table 2 because it

$C_i \vert L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N
$n_{\cdot j}$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N

Table 2: Contingency table at first case ($\ell = 1$)

$C_i \vert L_j$	L_1	$n_{i\cdot}$
C_1	n_{11}	n_{11}
$\vdots$	$\vdots$	$\vdots$
C_i	n_{i1}	n_{i1}
$\vdots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$n_{\ell 1}$
$n_{\cdot j}$	N	N

Table 3: Contingency table at second case ($k = 1$)

is not possible to divide by 0. Nevertheless, all the instances belong to the same class, therefore can be concluded that the dependence is maximum. In this case, let us indicate that $A^*(1) = 1$.

Regarding to Table 3, Ameva method can not be used because $\chi^2(k) = 0$ and the Ameva coefficient does not give any information about the dependence. However, the dependence is not minimum and a new coefficient is necessary. By taking into account that if all instances are distributed equally in all classes, the dependence is minimum, and if exists i such that $n_{i1} = N$, the dependence is maximum. Hence the following coefficient, called Entropy, is considered:

$$A(1) = 1 + \frac{1}{N \ln \ell} \sum_{i=1}^{\ell} n_{i1} \ln\left(\frac{n_{i1}}{N}\right)$$

It holds that $0 \leq A(1) \leq 1$, and:

- If $A(1) = 0$, then $n_{i1} = \frac{N}{\ell}$ (minimum dependence).
- If $A(1) = 1$, then a unique n_{i1} exists that $n_{i1} = N$ (maximum dependence).

Note 2.1 *Let us indicate these pathologic cases do not happen in a standard discretization, but it will be necessary taking into account in the presented methodology in the next section.*

3 The methodology

Given an attribute X_i where $i = 1, 2, \ldots, s$, the Ameva discretization algorithm is applied to this attribute so obtained intervals are considered as a new set of classes. This set of classes is denoted as follows:

$$\mathcal{C}^i = \{C_1^i, C_2^i, \ldots, C_{\ell_i}^i\} \tag{5}$$

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathcal{C}^i$ where $p = 1, 2, \ldots, \ell$. From (5), for each attribute X_j with $j = 1, 2, \ldots, s$, a g_{ijp} value is obtained from $\mathcal{C}^i$ as follows:

- If the X^p data subset all belong to the same class C^i, then $g_{ijp} = A^*(1) = 1$.
- If the subset of data belongs to different classes, then:
  - If values of the attribute X_j are always in the same interval, then $g_{ijp} = A(1)$.
  - If values of the attribute X_j are not always in the same interval, then $g_{ijp} = Ameva_N(\ell_i)$, where $Ameva_N(\ell_i)$ is defined as follows:

    $$Ameva_N(\ell_i) = \frac{\ell_i'}{N_p} Ameva(\ell_i)$$

    provide that N_p is the number of instances of the class X^p and ℓ_i' is the number of intervals of the attribute X_i for which there is at least one value in the data subset.

Note 3.1 *This new Ameva coefficient is chosen in order to obtain a normalized value $0 \leq Ameva_N(\ell_i) \leq 1$ as same as $A(1)$.*

Furthermore, it is straightforward to prove that if $i = j$ for $i = 1, 2, \cdots, s$, then $g_{iip} = 1$, for all $p = 1, 2, \cdots, \ell$.

Given $i, j = 1, 2, \cdots, s$, a g_{ij} value can be obtained applying this methodology for all class $C_p \in \mathcal{C}$ ($p = 1, 2, \cdots, \ell$), and by considering different statistics as follows:

- The minimum $g_{ij}^{min} = \min_p g_{ijp}$.
- The geometric mean $g_{ij}^{geo} = \sqrt[\ell]{\prod_{p=1}^{\ell} g_{ijp}}$.
- The arithmetic mean $g_{ij}^{arit} = \frac{1}{\ell} \sum_{p=1}^{\ell} g_{ijp}$.
- The maximum $g_{ij}^{max} = \max_p g_{ijp}$.

It is well-known that the following relationship is holded:

$$g_{ij}^{min} \leq g_{ij}^{geo} \leq g_{ij}^{arit} \leq g_{ij}^{max}$$

The main properties of the matrix $G = (g_{ij})$, that is,

$$G = \begin{pmatrix} 1 & g_{12} & \cdots & g_{1s} \\ g_{21} & 1 & \cdots & g_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s1} & g_{s2} & \cdots & 1 \end{pmatrix}$$

are the following: i) it is square but non symmetric matrix; ii) the values of the main diagonal are 1; and iii) $0 \leq g_{ij}$, $g_{ji} \leq 1$.

From the G matrix, a method of generating rules of dependence between attributes can be defined. For example, a possible rule is the next: given a threshold value, U, if $maxG_{ij}, G_{ji} > U$ and $i < j$, then the X_j variable is eliminated. Let us illustrate it with an example in the next section.

4 A toy example

Let us consider the Iris Plants Database[1] from UCI Repository which is perhaps the best known database to be found in the pattern recognition literature. This data set is considered due to its simplicity since this methodology is not completely defined yet.

The data set contains four attributes (sepal length, sepal width, petal length and petal width) and three classes (Setosa, Versicolour and Virginica) of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are not linearly separable from each other.

The matrices generated by the presented methodology in this paper are:

$$G_{Iris}^{min} = \begin{pmatrix} 1 & 0.4898 & 0.6667 & 0.0998 \\ 0.3265 & 1 & 0.035 & 0.093 \\ 0.028 & 0.0586 & 1 & 0.0303 \\ 0.0545 & 0.0998 & 0.0836 & 1 \end{pmatrix} \tag{6}$$

$$G_{Iris}^{geo} = \begin{pmatrix} 1 & 0.7883 & 0.8736 & 0.4638 \\ 0.6886 & 1 & 0.3271 & 0.453 \\ 0.1727 & 0.2674 & 1 & 0.1293 \\ 0.1573 & 0.3222 & 0.2244 & 1 \end{pmatrix} \tag{7}$$

$$G_{Iris}^{arit} = \begin{pmatrix} 1 & 0.8299 & 0.8889 & 0.6999 \\ 0.7755 & 1 & 0.6783 & 0.6977 \\ 0.4039 & 0.4617 & 1 & 0.3672 \\ 0.3753 & 0.4783 & 0.4063 & 1 \end{pmatrix} \tag{8}$$

$$G_{Iris}^{max} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{9}$$

This result shows that it is possible to determine the dependence of attributes of a data set from the Ameva discretization

[1]Available at http://archive.ics.uci.edu/ml/datasets/Iris

algorithm and the adjustments to resolve the inconsistencies outlined above with the entropy.

The coefficients in the minimum matrix (6) determine the lowest coefficients of dependence between two attributes. These coefficients provide information about there is a class for which the two attributes have less dependency. If these values are high, it is possible to conclude that the dependence between two attributes is high. Therefore, these coefficients are a minimum threshold for each pair of attributes.

A similar conclusion can be obtained from the maximum matrix (9). The coefficients provide information about there is a class for which the two attributes have a high dependence. In this case, these coefficients are the maximum threshold values for each pair of attributes.

Given a data set, the best result is achieved when the maximum and minimum matrix are the same. In this case, all the attributes are the same dependence with other regardless of the original class. Thus, there is only one matrix for generate the discrimination rules.

The arithmetic mean (8) and the geometric mean matrix (7) represent a global value of dependency. While the geometric mean matrix rewards the worst situations about a class, leading to a low value on the global coefficient, the arithmetic mean matrix balances the values of the coefficients.

A possible interpretation to determine which attributes are dependent of each other is to establish a threshold value. From this limit, two attributes are dependent if the average of the coefficients g_{ij} and g_{ji} of the arithmetic mean matrix is greater than or equal to this value.

In this case, the threshold value of 0.75 is established to check which attributes are dependents. The pair g_{ij}, g_{ji} that reaches this threshold is G_{12}, G_{21} because the arithmetic mean of G_{12} and G_{21} is greater than 0.75. It is necessary indicate that the sepal length and the sepal width features are the first and second attributes in the experiment.

Thus, in order to carry out a classification problem can be declared that the X_1 and X_2 features are similar. Let us see this affirmation by using as classification algorithm the Support Vector Machine (SVM) [1].

A performance for the 1-v-r SVM, in the form of accuracy rate, has been evaluated on models using the Gaussian kernel with $\sigma = 1$, and $C = 1$. The criteria employed to estimate the generalized accuracy is the 5-fold cross-validation on the whole set of training data. This procedure is repeated 120 times in order to ensure good statistical behavior. The obtained results are:

- With all features, the accuracy rate is 0.9320.
- Without the sepal length feature, the accuracy rate is 0.9341.
- Without the sepal width feature, the accuracy rate is 0.9667.

Furthermore to check that the accuracy rate is not less when a feature is eliminated, the methodology has discovered that these features introduce noise in the classification problem when both are used at the same time because the results are improved without the second feature.

5 Conclusions and future work

We have studied a method of discretization, Ameva, whose objective is to maximize the dependence between the intervals that divide the values of an attribute and the classes to which they belong, providing at the same time the minimum number of intervals.

After that, we have developed a methodology to reduce the number of features of a data set based on the dependence between them. To the best of knowledge, there are not existing researches that directly address the problem to reduce the number of features using a similar approach to ours.

This development is based on taking advantage of Ameva discretization algorithm. Thus, a new coefficient has been developed to determine the dependence between features. Hence, we have reduced the number of values of features and the number of features from a qualitative reasoning.

To test the development of the methodology, it has been applied to a well-known data set for obtain the dependent relationship between their features. Nevertheless, we think that this approach can be satisfactorily apply in this area when the data set has a lot of instances and features, and one of these features determines the class which each instance belongs. Another data sets must fulfill these characteristics.

Finally, after applying the discrimination of features obtained in the methodology, the modified data set has been carried out for the classification tests to verify the effectiveness of the methodology.

The next step to complement this development is the design of an automatic method of creation of feature discrimination rules. Subsequently, we must define some improvements in this methodology to automatically know the dependence between features without setting manually a threshold value.

Acknowledgments

This research is supported by the Spanish Ministry of Science and Innovation R&D project ARTEMISA (TIN2009-14378-C02-01).

References

[1] L. González, C. Angulo, F. Velasco, and A. Catala. Dual unification of bi-class support vector machine formulations. *Pattern recognition*, 39(7):1325–1332, 2006.

[2] Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[3] M. Govindarajan. Text Mining Technique for Data Mining Application. *Proceedings of World Academy of Science, Engineering and Technology*, 26:544–549, 2007.

[4] R. Entezari-Maleki, S.M. Iranmanesh, and B. Minaei-Bidgoli. An Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification Methods. *World Applied Sciences Journal*, 7:216–223, 2009.

[5] L. Gonzalez-Abril, FJ Cuberos, F. Velasco, and JA Ortega. Ameva: An autonomous discretization algorithm. *Expert Systems with Applications*, 36(3):5327–5332, 2009.

[6] M. Boulle. Khiops: A statistical discretization method of continuous attributes. *Machine Learning*, 55(1):53–69, 2004.

[7] L.A. Kurgan and K.J. Cios. CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.

[8] F.J. Ruiz, C. Angulo, N. Agell, X. Rovira, M. Sánchez, and F. Prats. A Discretization Process in Accordance with a Qualitative Ordered Output. *Proceeding of the 2005 conference on Artificial Intelligence Research and Development*, 131:273–280, 2005.

[9] R.P. Li and Z.O. Wang. An entropy-based discretization method for classification rules with inconsistency checking. 1:243–246, 2002.

[10] L. Gonzalez-Abril, F. Velasco, JA Ortega, and FJ Cuberos. A new approach to qualitative learning in time series. *Expert Systems with Applications*, 36(6):9924–9927, 2009.

[11] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[12] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. *Machine Learning Conference Proceedings*, pages 121–129, 1994.

[13] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *Intelligent Systems and Their Applications, IEEE*, 13(2):44–49, 1998.

[14] KM Faraoun and A. Rabhi. Data dimensionality reduction based on genetic selection of feature subsets. *INFOCOM - Journal of Computer Science*, 6(2):9–19, 2007.

[15] L. Rocchi, L. Chiari, and A. Cappello. Feature selection of stabilometric parameters based on principal component analysis. *Medical and Biological Engineering and Computing*, 42(1):71–79, 2004.

[16] Nitin Khosla. *Dimensionality Reduction Using Factor Analysis*. PhD thesis, 2006.

CHAPTER 8

THE CICA GRID - A CLOUD COMPUTING INFRASTRUCTURE ON DEMAND WITH OPEN SOURCE TECHNOLOGIES

Overview

This paper describes the implementation of a technology solution called CICA GRID that allows the expansion or replication of resources depending on research demands. A cluster is supported and the applied model is basically IaaS inside a private cloud which is accessed only by users from a private network.

The main motivation is to increase the requirements for access to excess computational resources of a working scheme in a cluster with an Local Resource Manager where authenticated and authorized users could design their own computational infrastructure and then use and manage it across a comfortable and simple interface.

From the point of view of the energy-saving involved in virtualized environments, it renders research advances with less cost by releasing electrical consumption and the advantage of increasing and decreasing the resources according to the needs.

Context

This research started as a result of a research project where researchers could increase or decrease resources in a cloud environment. It was applied in a real professional context as a result of one year of researching in the Cloud Computing area in the Scientific Computing Center of Andalusia where this PhD student worked.

Research work information

The purpose of the *International Conference on Enterprise Information Systems* is to bring together researchers, engineers and practitioners interested in the advances and business applications of information systems. This conference is held in conjunction with Evaluation of Novel Approaches to Software Engineering.

It is indexed in SJR with a Impact Factor of 0.111 and H Index of 4 in one category: Information Systems and Management. Also, it is indexed in CORE with a Rank C.

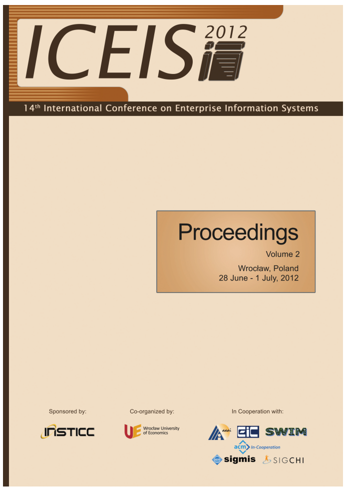


Figure 8.1: International Conference on Enterprise Information Systems cover.

DOI: https://doi.org/10.5220/0003992603010304

The CICA GRID
A Cloud Computing Infrastructure on Demand with Open Source Technologies

M. A. Alvarez[1], A. Fernandez-Montes[1], J. A. Ortega[1] and L. Gonzalez-Abril[2]
[1]*Computer Languages and Systems Department, University of Seville, Seville, Spain*
[2]*Applied Economics I Department, University of Seville, Seville, Spain*

Keywords: Cloud Computing, Cluster, Cobbler, IaaS, OpenNebula, Profile, Puppet, ReCarta, Virtual Machine.

Abstract: A new approach technology to enable the expansion and replication of resources on demand is presented in this paper. This approach is called CICA GRID and it provides service to research community in the Scientific Computer Centre of Andalusia (CICA). This approach is an alternative solution to the initial cost involved in building an own data center by public organizations for researches. This solution quickly provides resources with a minimal technical staff effort. Also, an architecture and user interface example called ReCarta was presented. This system supplies a private Cloud Computing system for non-technical end-users.

1 INTRODUCTION

In the last years, Cloud Computing has been launched as a concept that it has potential to transform the way in which computers are used and managed. This technology promises to realize the objective of transforming the computing resources into a single process. This process can use any quantity of resources during the needed time.

These features are especially interesting for the HPC/Grid Computing/Scientific area since they enable resources to be managed in a controlled environment. In this sense, researchers estimate their computing needs when a new project is considered. Therefore, researchers must spend their time to configure the available resources to satisfy their needs. Furthermore, the problem is bigger if researchers do not have technical computer knowledges.

In response to the needs of researchers and to improve the Andalusian Supercomputing Network (RASCI)[1], the Scientific Computer Center of Andalusia (CICA)[2] has implemented a technology solution called CICA GRID. It enables the expansion or replication of resources depending on research demands. CICA, in collaboration with the Spanish National Grid Initiative[3], features a high scalability cloud with a quick resources configuration: a GRID environment solution.

The developed approach incorporates three tools to carry out its function: Cobbler[4], Puppet[5] and OpenNebula[6]. It was called ReCarta[7] and it hides these tools to user by a web interface.

The paper is organized as follows: Cloud Computing technology is briefly presented in the next section. In Section 3, the project motivation and system architecture are analyzed and users' tools and examples are given in Section 4. Benchmarks and features are shown in Section 5. Section 6 provides a final discussion and concludes this paper.

2 CLOUD COMPUTING

Cloud Computing refers to hardware and software infrastructure which allows applications to be served across the web for end-users. Furthermore, it provides computational resources and virtual hosting to build their own applications for them and the hardware and software datacenter is called the Cloud.

There are two kinds of Cloud: Public Cloud (Armbrust et al., 2009) and Private Cloud. The first one is available for commercial purposes and pay-per-use (Stuer et al., 2007). The second one is found in an in-

[1]http://rasci.cica.es
[2]http://www.cica.es
[3]http://www.es-ngi.es
[4]http://cobbler.github.com
[5]http://reductivelabs.com/trac/puppet
[6]http://www.opennebula.org/doku.php
[7]http://trac.cica.es/recarta/

A. Alvarez M., Fernandez-Montes A., A. Ortega J. and Gonzalez-Abril L..
The CICA GRID - A Cloud Computing Infrastructure on Demand with Open Source Technologies.
DOI: 10.5220/0003992603010304
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (ICEIS-2012), pages 301-304
ISBN: 978-989-8565-11-2

dividual organization and the access is only allowed to authorized members. Also, Cloud Computing systems can be classified as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service).

3 PROJECT MOTIVATION AND SYSTEM ARCHITECTURE

A cluster for HPC is supported by CICA and the applied model is basically IaaS inside a private cloud which is accessed only by users from RASCI. It uses SGE as the Local Resource Manager (LRM) with Sun Grid Engine. Furthermore, this cluster has about 30 machines ant it is part of the Spanish National Grid Initiative.

The main motivation is to increase the requirements for access to excess computational resources of a working scheme in a cluster with an LRM. Thus, a project where authenticated and authorized users could design their own computational infrastructure and then use and manage it across a comfortable and simple interface was initiated. It was called “Recursos a la carta” (À-la-carte resources).

To achieve these goals some technical issues must be solved: i) machine supply; and ii) how to distribute the available physical resources among these virtual machines which they need them. The presented proposal, called CICA GRID, is developed as follows.

3.1 Provisioning and Management of Large-scale Virtual Systems

The CICA GRID is a private cloud with 35 virtual machines. It is composed by gLite’s working nodes and services (Andreetto et al., 2008). It is essential to have a tool that enables easy and flexible administration of these machines. The management of this cloud will be easier and more automated with it. Also, it must support the production control of the features and services of each machine. Hence, the problem of building the machines demanded by users has been resolved using Cobbler/Koan.

This tool facilitates the provisioning of virtual machines according to options given by users when they select how they want to build their machines and establishes an object hierarchy which defines the configuration characteristics at the highest levels. From the highest to the lowest level, they are Distro, Profile, Subprofile and System.

The relationship between objects that can be defined with Cobbler and the actual supplied machines are shown in Figure 1.

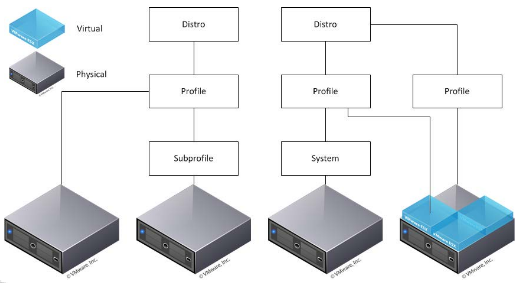


Figure 1: Cobbler object hierarchy.

At the time of computer installation, PXE boots the system while Cobbler shows a drop-down menu where the installation type can be chosen. If a virtual machine is going to be supplied, then it is possible to use the Koan command over the physical machine to specify what kind of machine is needed.

In the CICA GRID, users initiate a guided installation through a Cobbler profile. In answer to this request, the designated virtual machines are kept in a shared space (machine repository [see Figure 2]) where they are left available to OpenNebula for deployment.

Since provision and deployment of virtual machines does not resolve all infrastructure maintenance problems, a system for automating administration tasks is required and Puppet has been chosen. It provides a framework to simplify the work of system administrators, reusing the code as much as possible and allowing a modular system. Also, it is based on a client-server scheme and a declarative language that specifies administration tasks.

Puppet is used in the CICA GRID to configure and ensure that the NTP service of machines works correctly. Also, it must ensure that users are authenticated by LDAP and a basic backup configuration, security updates and certain file systems are set up. Through Cobbler profiles, each newly supplied virtual machine has a Puppet installed client.

Both Cobbler/Koan and Puppet have been proved to be capable of providing support for hundreds of machines.

3.2 Virtualized Systems

Open Nebula has been chosen to solve the problem of finding a system for an efficient deployment on virtual machine. It is an open-source virtual infrastructure engine and it enables dynamic deployment and re-placement of virtual machines using a pool of physical resources. It has achieved to decouple the server not only from the physical infrastructure, but also from the physical location.

Therefore, once Cobbler has provided machines requested by users and they have been saved in repository, the system will build needed files to enable OpenNebula to launch the deployment of machines as shown in Figure 2.

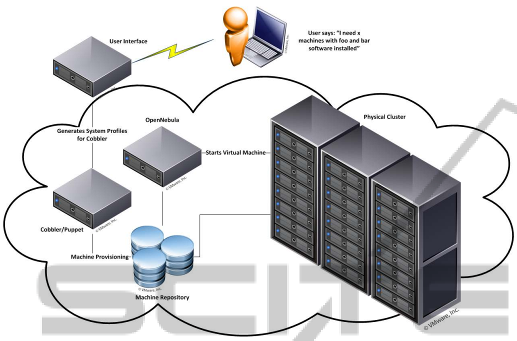


Figure 2: Architecture: connection scheme.

4 USER TOOL

The CICA GRID has a modular design. It facilitates its development and has been implemented in Python (Lutz and Van, 2001) language. Each module presents a well-defined interface so it can be easily used by other parts. They are Cobbler, DHCP and DNS management, Puppet, Open Nebula and User Interface modules.

4.1 ReCarta

A minimalist approach which attempts to show users the possible options is required. Therefore, the main focus is not on writing less code, but providing users with a useful system. This system is called ReCarta.

The created machine is composed for 2 steps. At first, users must define hardware and software features when they create a new group of machines. Later, users must indicate how many machines and the names of each have to be defined with these features.

At the end of this process, created system user data is shown along with the information needed for connection and start-up. Therefore, users have a project control panel at their disposal and they can see systems that they have been defined.

4.2 Code Example

A code example is given in order to illustrate the set of calls to defined API by different modules. They carry out tasks that a user has requested via web interface.

A new Cobbler profile (a new project in the user terminology) is created. It defines machines with 1 CPU, 512 MB of RAM, 4 GB of hard drive and Java language support.

```
import mod_cobbler
import mod_dhcpDns
import mod_puppet
import mod_nebula

miCobbler = mod_cobbler.Cobbler( 'john' )
miCobbler.setProfile( {
    'nombrePerfil':'project',
    'kickstart':'vm-kickstart.template',
    'diskSize':'4',
    'ram':'512',
    'cpus':'1',
    'comment':'project profile',
    'software':[ 'X-WINDOW', 'JAVA-SUPPORT' ]
} )
macs = miCobbler.setSystems( [
    { 'nombre':'project-vm',
        'comentario':'project VM',
        'perfil':'project' }
] )
mapIpNames = mod_dhcpDns.addSystemsDHCP(macs)
mod_dhcpDns.addEntryDns( mapIpNames )
miCobbler.provisionSystems( [ 'project-vm' ] )
```

It is important to note two ReCarta design features[8]. One is the high abstraction level offered by different methods. The kickstart template is modified to adapt it to different user requirements and it do not appear nowhere. Also, DHCP/DNS server configurations are modified to assign a place to new systems in network.

The other feature of ReCarta is the absence of system database to save information about defined projects and users, etc. ReCarta put the usernames as a prefix to data profiles and the project name to defined system names by users. This design decision has been taken to keep ReCarta as simple as possible.

5 BENCHMARKS AND FEATURES

The CICA GRID is a private cloud with 35 virtual machines with the following virtualized features: 1 core and 1 GB RAM. 6 physical servers are used to virtualize them with the following features: 2 cores and 4 GB RAM.

[8]http://trac.cica.es/recarta/wiki/RecartaDevel

Table 1: HPCC benchmarks.

	Intel 6400 Physical	Intel 6400 Virtualized
PTRANS(GB/s)	0.65	0.54
HPL(Gflops)	14.26	13.01
MPI Latency(ms)	0.00043	0.00053
MPI Bandwidth (ms)	1471.17	1477.64

Table 2: Bonnie++ write test - 2Gb blocks.

Server type	Sequential output (Kbs)	Sequential input (Kbs)	Random (seek/s)
Virtual	14014	34851	150.7
Physical	45678	49719	64.7

Table 3: Bonnie++ create test - 1Gb blocks.

Server type	Sequential Create (s)	Random Create (s)
Virtual	0.0000	0.0000
Physical	1621	891

Table 4: Consumption of 6 physical and 35 virtual servers.

Servers	Consumption (KWh/year)	Total (KWh/year)
35 virtual	222.6	7791
6 physical	516	3096

Nowadays, ReCarta creates systems compound of Xen (Barham et al., 2003) virtual machines. It has been used because it is proved that a paravirtualized virtual machine only loses 5-10% of CPU performances respect to equivalent physical machine.

Table 1 presents HPCC benchmark execution results. As expected, performances of the physical machine are better than virtualized machine. However, it is observed that the performance is about the same in both cases, so we can conclude that the proposal can be accepted as valid.

Table 2 and Table 3 show Bonnie++ execution results for virtual machine memory and equivalent physical machine. Also, a significant decrease in performance between virtual and physical machine can be seen for writing action to disc in these latter cases. In this case, the differences are slightly higher because the benchmark is performed on disk access. In this process, a virtual machine generates a very intense traffic on its virtual hard disk, especially reading.

The power consumption can see in Table 4. The use of virtualization allows the power consumption to be reduced to 39% can be seen in it.

6 CONCLUSIONS

Although the CICA GRID is still in its experimental phase, some case studies have been carried out. One of them is the creation of a small virtual cluster to be used with Apache Hadoop (Borthakur, 2007). Also, the project objectives have enabled that more job requests could be served without exceed the normal workload for a HPC cluster.

From the point of view of the energy-saving involved in virtualized environments, the CICA GRID renders research advances for the research groups with less cost by releasing electrical consumption.

We have learned during the launching of our pilot project of Cloud Computing that our users appreciate two advantages: i) the illusion of having a huge computing resource reserved exclusively for them; and ii) the possibility of increasing and decreasing the resources according to their needs.

ACKNOWLEDGEMENTS

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness ARTEMISA (TIN2009-14378-C02-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

REFERENCES

Andreetto, P., Andreozzi, S., Avellino, G., Beco, S., Cavallini, A., Cecchi, M., Ciaschini, V., Dorise, A., Giacomini, F., Gianelle, A., et al. (2008). The glite workload management system. In *Journal of Physics: Conference Series*. IOP Publishing.

Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., et al. (2009). Above the clouds: A berkeley view of cloud computing. Technical report, Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley.

Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A. (2003). Xen and the art of virtualization. *ACM SIGOPS Operating Systems Review*, 37(5):164–177.

Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*.

Lutz, M. and Van, G. (2001). *Programming Python: Object-Oriented Scripting*. O'Reilly & Associates, Inc., London, 2nd edition.

Stuer, G., Vanmechelen, K., and Broeckhove, J. (2007). A commodity market algorithm for pricing substitutable grid resources. *Future Generation Computer Systems*, 23(5):688–701.

CHAPTER 9

ACTIVITY RECOGNITION SYSTEM USING AMEVA METHOD

Overview

This article aims to develop a system of care and monitoring where the goal is to get a efficient way that allows to control the physical activity carried out by the user. It is based on the use of discrete variables which employ data from accelerometer sensors using an discretization and classification technique based on the chi^2 distribution.

The system is designed to be deployed and executed on smartphones, so the system energy consumption is taken into account. In this case, the results show that the developed algorithm reduces the computational cost of the system by about 50% compared on neural networks.

In other hand, the main problem of this system is the limit to the number of activities that can be recognized because is based on statistical learning. Also, working only with accelerometer sensors limits the number of variables that can be analyzed because they could lead to a strong correlation between them.

The approach could be tested in a real scenario during the Evaluating AAL Systems Through Competitive Benchmarking (EvAAL) competition in 2012 and it was a great chance to make a real stress test of AMEVA system and it allowed to know other techniques in activity recognition and new perspectives about this field.

Context

This research was the first collaboration between two different areas in the research group: discretization and activity recognition. It was based on the advantage of Ameva in terms of computation performance and pass it into the activity recognition approaches for smartphones getting a significant reduction of battery draining. Also, it was thought to use it in the EvAAL competition in 2012. This paper is the result of over 1 year's work and currently is one of the areas under investigation.

Research work information

The *Evaluating AAL Systems Through Competitive Benchmarking* conference aims at establishing benchmarks and evaluation metrics for comparing Ambient Assisted Living solutions. Since 2011 it has organised international competitions on indoor localization and indoor activity recognition.

It is indexed in SJR with a Impact Factor of 0.140 (Q4) and H Index of 19 in one category: Computer Science (miscellaneous).

DOI: https://doi.org/10.1007/978-3-642-37419-7_11

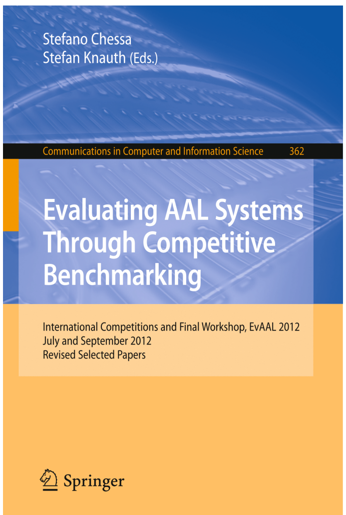


Figure 9.1: Evaluating AAL Systems Through Competitive Benchmarking cover.

Activity Recognition System Using AMEVA Method

Luis M. Soria Morillo[1], Luis González-Abril[2],
Miguel A. Álvarez de la Concepción[1], and Juan A. Ortega Ramírez[1]

[1] Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain
{lsoria,malvarez,jortega}@us.es
[2] Applied Economics I Dept., University of Seville, 41018 Seville, Spain
luisgon@us.es

Abstract. This article aims to develop a minimally intrusive system of care and monitoring. Furthermore, the goal is to get a cheap, comfortable and, especially, efficient system which controls the physical activity carried out by the user. For this purpose an innovative approach to physical activity recognition is presented, based on the use of discrete variables which employ data from accelerometer sensors. To this end, an innovative discretization and classification technique to make the recognition process in an efficient way and at low energy cost, is presented in this work based on the χ^2 distribution. Entire process is executed on the smartphone, by means of taking the system energy consumption into account, thereby increasing the battery lifetime and minimizing the device recharging frequency.

1 Introduction

Just 30 minutes of moderate activity five days a week, can improve your health according to the Centers for Disease Control and Prevention. By enabling activity monitoring on an individual scale, over an extended period of time in a ubiquitous way, physical and psychological health and fitness can be improved. Studies performed by certain health institutes initiative [7,3,10,6] have shown significant associations between physical activity and reduced risk of incident coronary heart disease and coronary events. Their results can be seen in Figure 1, where the inverse correlation between the risk of cardiovascular incidents and physical activity level is shown through the comparison of four separate studies.

In recent years, thanks largely to the increased interest in monitoring certain sectors of the population such those of as elderly people with dementia and of people in rehabilitation, activity recognition systems have increased in both number and quality. Furthermore, communication between relatives, friends and professionals can be improved by means of graphs of weekly activity (high relevant for sportsmen and for the relatives of elderly people) whereby the doctor can be automatically alerted if any strange activity is detected. By using data acquired from accelerometer, *NFC*, or even microphone sensors and applying

S. Chessa and S. Knauth (Eds.): EvAAL 2012, CCIS 362, pp. 137–147, 2013.

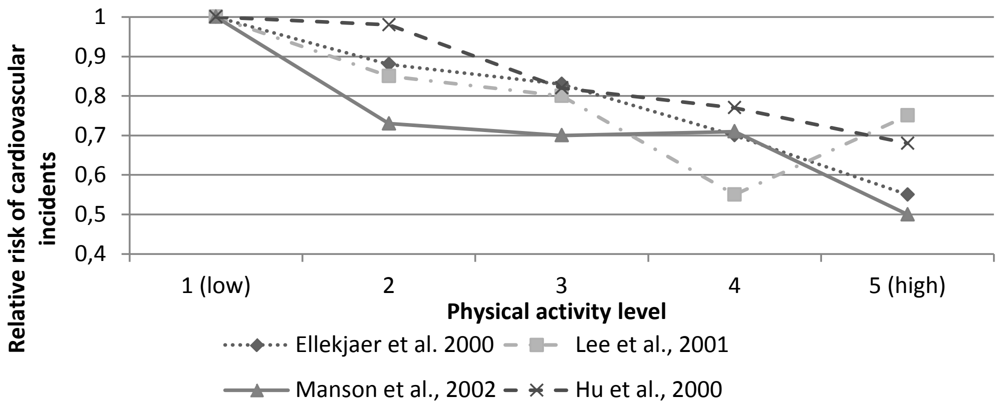


Fig. 1. Associations between physical activity and reduced risk of incident coronary heart disease and coronary events

some classification algorithm, it is possible to recognize human activities. Artificial neural networks (*ANN*) method will be analyzed and compared with our work. Results show the main differences between different studies, and certain drawbacks are determined which rules them out for development on users' smartphones To reduce the cost related to process accelerometer signals, this paper opts for an innovative technique, through which the work is performed in the field of discrete variables. Thanks to a discretization process, the classification cost is much lower than that obtained when working with continuous variables. Any dependence between variables during the recognition process is therefore eliminated and, on the other hand, energy consumption from the process itself is minimized.

2 Activity Recognition

2.1 Data Collection

Certain related studies attain results on activity recognition off-line. A comprehensive training set from the accelerometer output is first needed before data can be classified into any of the recognized activities. However, this paper has sought to minimize the waiting time for recognition, thereby providing valid information of the activity very frequently. To this end, both training and recognition sets are obtained using time windows [8] of fixed duration. After having conducted a performance and system accuracy analysis, it is determined that the optimum length for these windows is 5 seconds. Five seconds windows was chosen due to for our system it's extremely important to ensure that in each time window there is, at least, one activity cycle. Where activity cycle is define as an complete execution of some activity pattern. For instance, two steps are an activity cycle for walking and one pedal stroke is the activity cycle for cycling. If at least one activity cycle can not be ensure in each time window, it's not possible to determine, basing on accelerometer patterns, the activity performed. This statement could be seen in the next example. Suppose a two second cycle is having and the actor is jumping continuously, that is, we have a cadence of one jump for each two seconds. The system is configured with one second time window and thus, for each activity cycle will have two windows. In the first one, while the user is rising, vertical acceleration is negative. In the other one, because the user is falling, vertical acceleration will positive. If user increase the cadence by two, mean between acceleration set is close to , due to vertical positive and negative accelerations will be counteracted. For this reason, it's very important to ensure that one cycle of all activities, regardless of the speed performed, is contained in a time window. Segmentation process and activity cycle is shown in Figure 2.

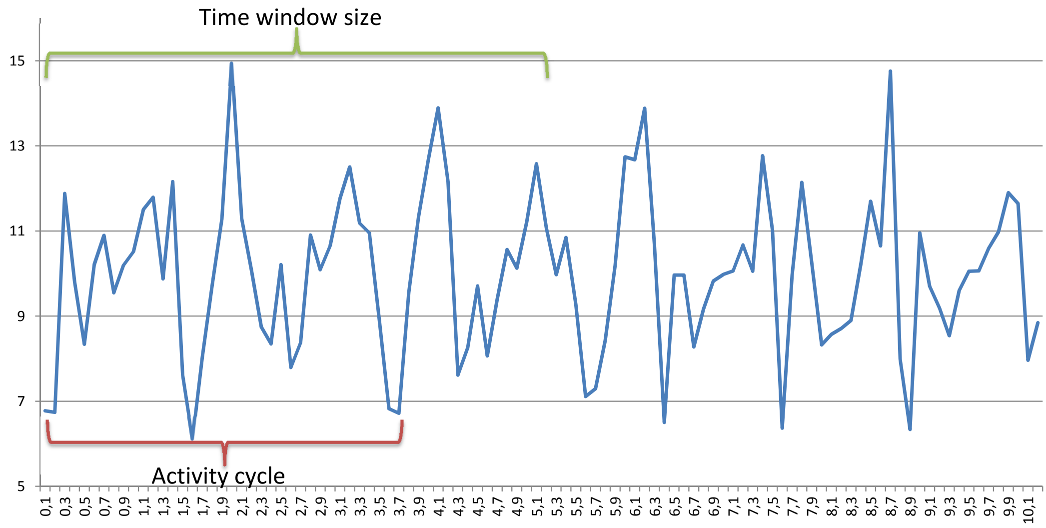


Fig. 2. Time windows split method over accelerometer signal

Based on these time windows, which contain data for each accelerometer axis, the signal module has been chosen in order to reduce the computational cost of the new solution. In addition to rendering the system more efficient, this choice of module eliminates the problem caused by device rotation [5,4]. Furthermore, user comfort with the system is decreased by removing the restriction that forces its orientation to be maintained during the process of learning and recognition. Using the accelerometer module, a data from each of the different readings taken within a time window $a_i = (a_{x,i}, a_{y,i}, a_{z,i})$ for the x, y, and z axes is defined as follows

$$|a_i| = \sqrt{(a_{x,i})^2, (a_{y,i})^2, (a_{z,i})^2} \tag{1}$$

For each temporal window is obtained Arithmetic Mean, Minimum, Maximum, Median, Std deviation, Geometric mean and other measures. In addition to the above variables, hereafter called temporal variables, a new set of statistics from the frequency domain of the problem is generated. This second set of variables will be called frecuencial variables. In order to obtain the frequency characteristics, the Fast Fourier Transform (FFT) for each time window is applied. In this way, and based on the frequency components obtained.

2.2 Set of Activities

Far from being a static system, the number and type of activities recognized by the system depends on the user [9]. However, to carry out a comparative analysis of the accuracy and performance of the discrete recognition method proposed below, 8 activities were taken into account. These activities are immobile, walking, running, jumping, cycling, drive, walking-upstairs and walking-downstairs. The learning system allows the user to decide what activities he/she wants the system to recognize. This is highly useful when the determination of certain very specific activities on monitored users is required. Examples of this situation include patients in rehabilitation who are monitored during their period of learning the various physical tasks prescribed by their doctors.

3 Qualitative Method

3.1 Ameva Algorithm

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a data set of a continuous attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of ℓ classes of the variable denoted by

$$\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2$$

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \to \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of the property that is being discretized.

Let us consider a discretization $\mathcal{D}$ which discretizes the continuous domain of $\mathcal{X}$ into k discrete intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{[d_0, d_1], (d_1, d_2], \cdots, (d_{k-1}, d_k]\}$$

In this discretization, d_0 is the minimum value and d_k is the maximum value of the attribute $\mathcal{X}$, and the d_i values are in ascendent order.

If L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2, 3, \ldots, k$, then

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \cdots, L_k\}$$

Therefore, the aim of the Ameva method [1] is to maximize the dependency relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k.

As a result from applying the above algorithm to each statistical value of the system, a series of intervals associated with a particular $\mathcal{C}$ tag is obtained. Thus, after processing all system statistics, a three-dimensional matrix is obtained. In the first two dimensions, the label of the activity $\mathcal{C}$ associated with the interval

$L_i = (L_i^l, L_i^s]$, as well as with the lower limit L_i^l and the upper limit L_i^s of that range is stored. In a third dimension, the matrix contains the above data for each statistic $\mathcal{S} = \{S_1, S_2, ..., S_{\mathbb{S}}\}, \mathbb{S} \geq 2$. This three-dimensional matrix containing the set of interval limits for each statistic is called the *Discretization Matrix* and is denoted by $Dm\{\mathcal{C}, L^{l,s}, \mathcal{S}\}$. The *Discretization Matrix* therefore determines the interval to which each item of data belongs with respect to each statistical value, by means of carrying out a simple and fast discretization process.

Class Integration. The next step of the algorithm determines the probability associated with the statistical data for each of the activities based on previously generated intervals. To this end, each element of the training set $x = \{\mathcal{X}; \mathcal{C}\}$ is processed, to which, in addition to the value of each statistic whose calculation is based on the time window, is also associated the label of the specific activity in the training set. In order to carry out this process, *Class-Matrix* is denoted by $Cm\{x, L_i, \mathcal{S}\}$ and is defined as a three-dimensional matrix that contains the number of data x from the training set associated with each L_i interval for each statistical $\mathcal{S}$ of the system. This matrix is defined as follows,

$$Cm_{x,i,s} = |x \in \mathcal{X} | x \geq L_i^l \wedge x < L_i^s \wedge x\{\mathcal{C}\} = C_s \tag{2}$$

Therefore, by this definition, each position in the *Class-Matrix* is uniquely associated with a position in the *Discretization-Matrix*, as determined by its range.

At this point not only is it possible to determine the discretization interval likelihood, but the *Class-Matrix* also helps to obtain the probability associated with the discretization process performed with the *Ameva* algorithm.

Activity-Interval Matrix. The next step in the learning process is to obtain the matrix of relative probabilities. This three-dimensional matrix, called the *Activity-Interval Matrix* and denoted by $AIm\{x, L_i, \mathcal{S}\}$, determines the likelihood that a given value x associated to an $\mathcal{S}$ statistic corresponds to a specific C_i activity. This ratio is based on the quality of the discretization performed by *Ameva*, and in order to determine the most probable activity from the generated data and the intervals of the training set. First the contents of the array AIm is defined as follows,

$$AIm_{c,i,s} = \frac{Cm_{c,i,s}}{total_{c,s}} \cdot \frac{1}{\ell - 1} \sum_{j=1, j \neq c}^{\ell} (1 - \frac{Cm_{j,i,s}}{total_{j,s}}) \tag{3}$$

where $total_{c,s}$ is the total number of time windows of the training process labeled with the c activity for the $\int$ statistic.

Figure 3 shows the overall process described on this section for carry on data analysis and interval determination.

3.2 Classification Process

Having obtained the discretization intervals and the probabilities of belonging to each interval, the process by which the classification is performed can be

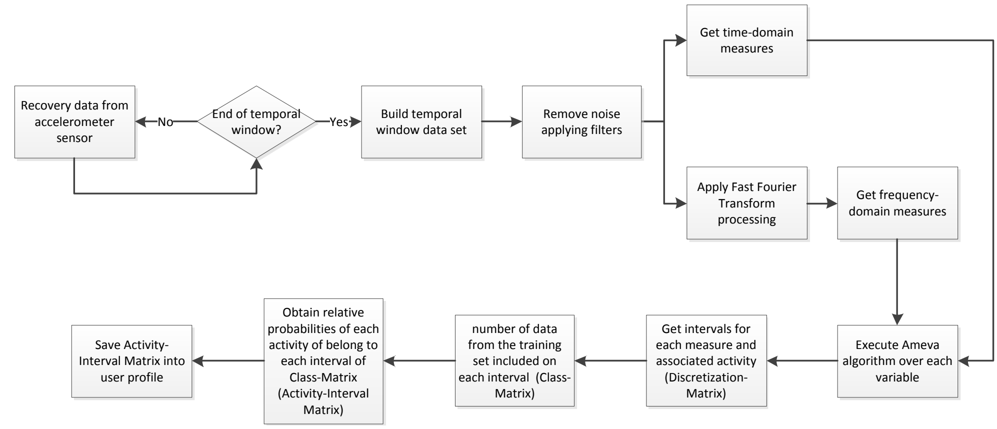


Fig. 3. Overall process of data analysis and interval determination

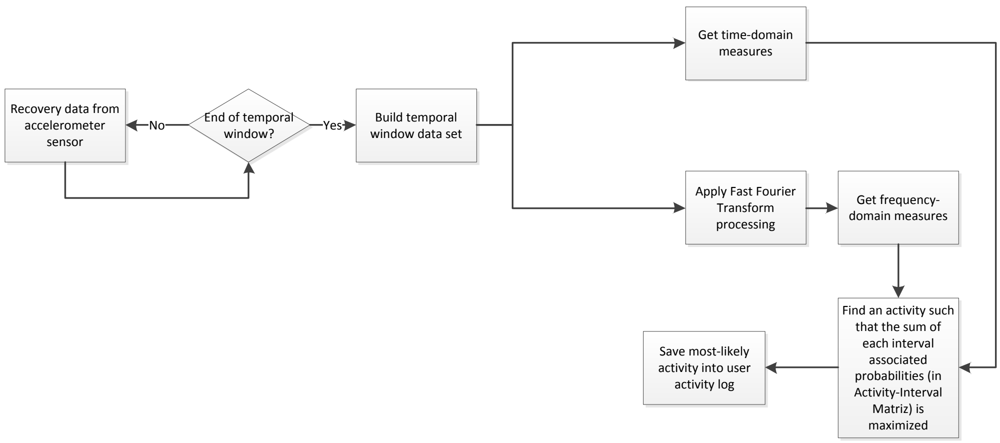


Fig. 4. Overall recognition process from data sensors

described. This classification is based on data from the analysis of time windows. The process is divided into two main steps: the way in which to perform the recognition of physical activity is first described; and the process to determine the frequency at which some particular activity is then presented.

Classifying Data. For the classification process, the most probable activity is decided by a majority voting system. This process starts from the *Activity-Interval Matrix* and uses a set of data $x \in \mathcal{X}$ for each of the statistics belonging to the $\mathcal{S}$ set. The process consists of finding an activity $mpa \in \mathcal{C}$ such that the likelihood is maximized. The above criterion is included in the following expression,

$$mpa(\mathcal{X}) = \max \sum_{s=1}^{s} AIm_{c,i,s} | x_s \in (L_i^l, L_i^s] \quad (4)$$

The expression shows that the weight contributed by each statistic to the calculation of the probability is identical. This can be carried out under the assumption that all statistics provide the same information to the system, and that there is no correlation between them. Thus, the most probable activity, or *mpa*, represents those activities whose data, obtained through the processing time window, is more suited to the AIm set values. In this way, the proposed algorithm not only determine the *mpa*, but also its associated probability. From this likelihood, certain activities that do not adapt well to sets of generic classification can be identified. This could be an indication that the user is carrying out new activities for which the system has not been previously trained.

Figure 4 shows the overall process described on this section for recognition process from Activity-Interval Matrix calculated in the previous stage.

4 Method Analysis

Now that the basis of the activity recognition algorithm has been laid out, an analysis of the new proposal can be performed. To this end, the new development is compared with a widely used recognition system based on neural networks [2]. In this case, both learning and recognition is performed by continuous methods. The test process is conducted on *Google Nexus S, Samsung Galaxy S2, and Google Nexus One* devices for a group of 40 users. Notably, the activity habits of these users are radically different, since 10 of them are under 25 years old, 20 users are between 25 and 40 years old, and the rest are over 40. An approximate distribution of the data for each subject regarding the eight activities in the study are: immobile (2800 min, 70 min per user), walking (2600 min, 65 min per user), running (2400 min, 60 min per user), jumping (2400 min, 60 min per user), cycling (2200 min, 55 min per user), driving (2200 min, 55 min per user), walking-upstairs (2400 min, 60 min per user), and walking-downstairs (2000 min, 50 min per user). Annotations are performed using a mobile application installed on the device itself with speech recognition software through which users dictate the name of the new activity when the physical activity being performance changes. Those unrecognizable activities conducted during the test process are dismissed to analyze the system accuracy. Data collection is obtained during four weeks.

Moreover, it is crucial to consider energy consumption and the processing cost of the system when it is working on a mobile device. In this case, after comparing the above methods, the conclusion reached is that the method based on *Ameva* reduces the computational cost of the system by about 50%, as can be seen in Figure 5. The time needed to process a time window by using the *Ameva*-based method is 0.6 seconds, while, for methods based on neural networks this figure is 1.2 seconds.

As can be seen in 6, Ameva battery consumption is lower than neural networks. For the first one, the battery lifetime is close to 25 hours while for the last one, it's only 16 hours. In the comparison can be observed the battery lifetime for decision tree but the main problem of this method, based on statistics chosen, is the low accuracy, not higher than 60%.

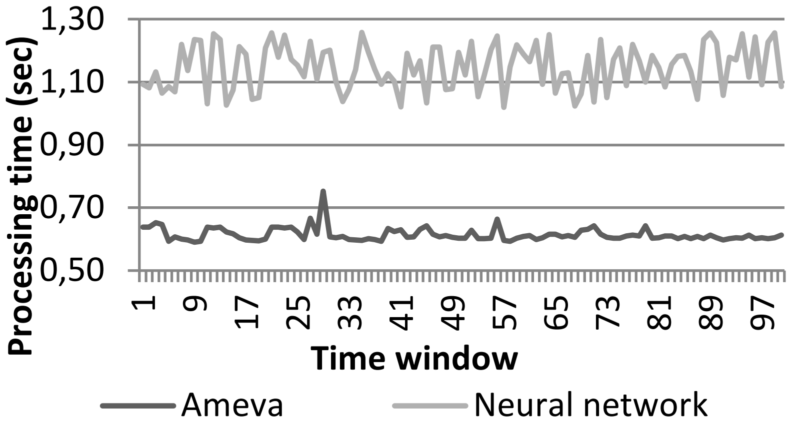


Fig. 5. Processing time of the Ameva and neural network methods on the device

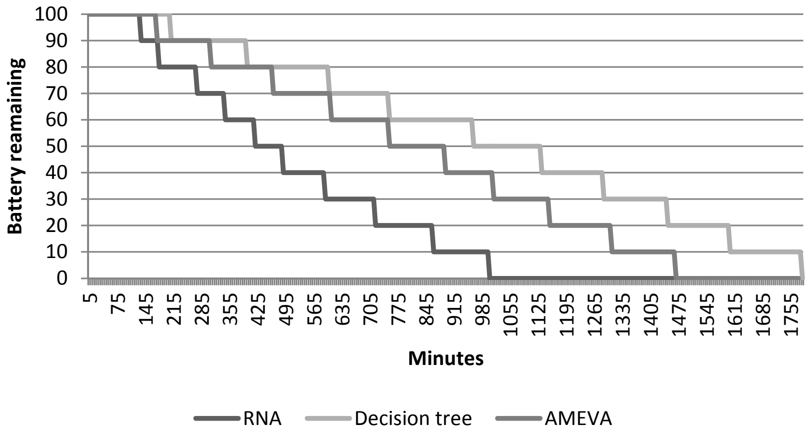


Fig. 6. Battery life for Ameva compared to neural network and decision tree methods

Based on Accuracy, Recall, Specificity, Precision, and F measure, Table 1 is presented. In this table, differences between the two methods, *RNA* and *Ameva* can be observed. Most values presented for each measure and activity show that the *Ameva* method performs better than *RNA*, especially as regards precision. That is to say, the number of false positive in the *Ameva* method is lower than that using the *RNA* method. *Immobile* and *Drive* are controversial activities due to their similar characteristics. Even under observation, it is difficult to differentiate between these two activities. For this reason and due to temporal nature of the *Immobile* activity, results from these two activities present a high level of disturbance in contrast to other activities.

Table 1. Performance comparison by using measures of evaluation

Activity	Accuracy		Recall		Specificity		Precision		F-measure (F_1)	
	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA
Walk	98.77%	97.93%	97.92%	93.95%	98.91%	98.57%	93.50%	91.36%	95.66%	92.64%
Jump	99.35%	98.87%	97.03%	96.44%	99.70%	99.25%	98.00%	95.12%	97.51%	95.77%
Immobile	98.69%	99.50%	94.57%	97.37%	99.42%	99.88%	96.60%	99.29%	95.58%	98.32%
Run	99.27%	98.35%	97.61%	92.62%	99.49%	99.14%	96.36%	93.64%	96.98%	93.13%
Up	98.93%	98.17%	95.40%	90.79%	99.43%	99.22%	96.00%	94.35%	95.70%	92.54%
Down	98.64%	98.25%	95.20%	92.68%	99.04%	98.89%	91.95%	90.62%	93.55%	91.64%
Cycle	99.32%	99.03%	96.13%	95.67%	99.73%	99.47%	97.91%	95.89%	97.01%	95.78%
Drive	98.14%	98.74%	90.02%	95.01%	99.20%	99.23%	93.63%	94.16%	91.79%	94.58%

5 Conclusions and Future Work

In this work, a highly successful recognition system based on discrete variables is presented, which uses the *Ameva* discretization algorithm and a new *Ameva*-based classification system. It has therefore been possible to achieve an average accuracy of 98% for the recognition of 8 types of activities. Furthermore, working with discrete variables has significantly reduced the computational cost associated to data processing during the recognition process. By using this process to increase recognition frequency, it has been possible to obtain a physical activity reading every 5 seconds and to enter these readings into the user activity log. However, the main problem of this system based on statistical learning is the limit to the number of activities that can be recognized. Working only with accelerometer sensors implies a limit to the number of system variables and therefore may lead to a strong correlation between these variables.

6 AMEVA Running in EvAAL Competition

During the competition, two test sessions were executed. In the first one, the training was performed prior to competition by an external actor not related to evaluation process. The training actor was 31 years old and the entire training process was performed with the smartphone in the hip, attached to the user's belt. In the competition, the actor was in a similar age range and thus, the way in that physical activity was executed was very similar. In other case, the system should be retrained for a better accuracy. Once finished the first evaluation session, intermediate data was analyzed. From this analysis, it was concluded that some activities was not well-recognized such as bending or cycling. This was a substantial impact in the accuracy due to cycling session was long. The accuracy for the other activities was promising but we detect that something was wrong for cycling detection. By using discrete techniques to perform the activity recognition, cycling is a easy activity to be detected because of the acceleration patterns presents an evident component in the advancing direction. Unfortunately, cycling activity was carried out on a stationary bike and thus, accelerations presented in movement direction was not detected. For the other controversial activity, bending, the system was not training to detect it because it was a important conflict with sitting activity. Both activities have a very similar acceleration profile and it can not be determine which is the right activity with a proper accurate. In the second test process, the system was retrained in order to achieve a most accurate recognition. Unfortunately, the Internet connection was not good enough to connect with training server placed at the University of Seville. For this reason, dataset from time windows was not properly sent to the server and therefore, the training parameters were wrong. After checking this problem, we decided to go on with the evaluation process to determine the impact of this problem in the accuracy. As it was thought, the second evaluation had a very low accuracy due to that problems. Furthermore, by studying intermediate data after the evaluation, temporal windows was misconfigured and it was set to 3 seconds and thus, some "fast activities" such as walking or cycling wasn't well recognized. Finally, EvAAL competition was a great chance to make a real stress test of AMEVA system since It's not usual in humans to make a long activities set in so quickly and so fast. In this regard, statistical-discrete classification for activity recognition based on AMEVA algorithm was designed to medium-long time activities. Transitions in discrete classification systems are really difficult to detect and, in AMEVA case, was not implemented any change activity detector. In conclusion, EvAAL offered a junction to test many systems and generate new ideas for competitors' systems. On the other hand, is very good to know other techniques in activity recognition and new perspectives about this field.

Acknowledgments. This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness ARTEMISA (TIN2009-14378-C02-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy.

References

1. González Abril, L., Cuberos, F.J., Velasco, F., Ortega, J.A.: Ameva: An autonomous discretization algorithm. Expert Syst. Appl. 36(3), 5327–5332 (2009)
2. Altun, K., Barshan, B., Tunçel, O.: Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recogn. 43(10), 3605–3620 (2010)
3. Ellekjaer, Holmen, Vatten: Physical activity and stroke mortality in women. Stroke 31(1), 14–18 (2000)
4. He, Z., Jin, L.: Activity recognition from acceleration data based on discrete consine transform and SVM. In: 2009 IEEE International Conference on Systems, Man and Cybernetics, pp. 5041–5044. IEEE (October 2009)
5. Khan, A.M., Lee, Y.-K., Lee, S., Kim, T.-S.: Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly. Medical & Biological Engineering & Computing 48(12), 1271–1279 (2010)
6. Lee, I.-M., Rexrode, K.M., Cook, N.R.: Physical activity and coronary heart disease in women. JAMA 285(11), 1447–1454 (2001)

7. Manson, J.E., Greenland, P., LaCroix, A.Z., Stefanick, M.L., Mouton, C.P., Oberman, A., Perri, M.G., Sheps, D.S., Pettinger, M.B., Siscovick, D.S.: Walking compared with vigorous exercise for the prevention of cardiovascular events in women. The New England Journal of Medicine 347(10), 716–725 (2002)
8. Paiyarom, S., Tungamchit, P., Keinprasit, R., Kayasith, P.: Activity monitoring system using Dynamic Time Warping for the elderly and disabled people, pp. 1–4. IEEE (2009)
9. Ravi, N., Nikhil, D., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence(IAAI), pp. 1541–1546 (2005)
10. Sattelmair, J.R., Kurth, T., Buring, J.E., Lee, I.-M.: Physical Activity and Risk of Stroke in Women. Stroke 41(6), 1243–1250 (2010)

CHAPTER 10

AN ADAPTIVE METHODOLOGY TO DISCRETIZE AND SELECT FEATURES

Overview

This paper designs an efficient and low computational way of finding independent features for getting the best accuracy on classification problem where an automatic method to select the best features is proposed. The methodology uses and extends the functionality of Ameva coefficient and allows to use it in other tasks of machine learning where it has not been defined.

One of the most important problems in classification processes is the selection of features. Usually, the obtained experimental data is not filtered about relevant features in systems and a lot of techniques for feature selection have been developed for this reason.

The Ameva discretization algorithm performs the discretization process effectively and quickly, so the set of values of a feature is greatly reduced. Taking this advantage as a premise, it is possible to use this algorithm to determine the relationship between features because Ameva uses the statistic to determine the relationship between features and classes. Thus, a new coefficient has been developed to determine the dependence between features that allows to reduce the number of values of features and the number of features from a qualitative reasoning.

This approach can be satisfactorily applied when the data set has a lot of instances and features, and one of these features determines the class which each instance belongs to.

Context

As a follow-up to the first paper on feature reduction, this research extended the same idea improving the algorithm in terms of personalization and performance. Feature reduction was one of the areas that this PhD candidate continued researching during the stay in Toulouse and this paper was one of the obtained results. This paper had over 5 months' work and is currently one of the areas under investigation.

Research work information

The journal *Artificial Intelligence Research* is a peer-reviewed, international scientific journal providing a forum for original research, reviews, experience exchange or conference reports related to the fields of Artificial Intelligence and Applications for researchers, programmers, software and hardware manufacturers.

DOI: `https://doi.org/10.5430/air.v2n2p77`

Figure 10.1: Artificial Intelligence Research cover.

ORIGINAL RESEARCH

An adaptive methodology to discretize and select features

Miguel Ángel Álvarez de la Concepción[1], Luis González Abril[2], Luis Miguel Soria Morillo[1], Juan Antonio Ortega Ramírez[1]

1. Computer Languages and Systems Department, University of Seville, Spain. 2. Department of Applied Economics I, University of Seville, Spain

Correspondence: Miguel Ángel Álvarez de la Concepción. Address: Computer Languages and Systems Department, University of Seville, Spain. Email: maalvarez@us.es.

Received: June 20, 2012 **Accepted:** December 15, 2012 **Online Published:** February 18, 2013
DOI: 10.5430/air.v2n2p77 **URL:** http://dx.doi.org/10.5430/air.v2n2p77

Abstract

A lot of significant data describing the behavior or/and actions of systems can be collected in several domains. These data define some aspects, called features, that can be clustered in several classes. A qualitative or quantitative value for each feature is stored from measurements or observations. In this paper, the problem of finding independent features for getting the best accuracy on classification problems is considered. Obtaining these features is the main objective of this work, where an automatic method to select features is proposed. The method extends the functionality of Ameva coefficient to use it in other tasks of machine learning where it has not been defined.

Key words

Ameva, Feature selection, Discretization, Entropy

1 Introduction

The problem to obtain the best accuracy, sensitivity, specificity, etc. in classification is one of the main problems in a lot of research areas like analysis and pattern recognition. It requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of features. One of them is in medical area when a doctor needs to know if a patient has cancer or not through thousands of gen values. In this sense, there are a lot of classifiers in the bibliography that process data sets to get the best results. Also, it is a central problem in machine learning. For example, there are classifiers based on SVM [1], Naive Bayesian [2], C4.5 [3], etc. that have been developed in the last years.

One of the most important preprocess in classification is the discretization because it allows algorithms to run very fast. This process establishes a relationship between continuous variables and their discrete transformation through developed functions. Therefore, it is possible to qualitatively model a series of continuous values if a label is assigned to them. Some studies [4] have shown that executing a prior process to discretize continuous features is more efficient than working directly with the continuous values. The discretization process reduces the computation memory usage and time in the application that develops classification algorithms. Also, it is used to manage the values of a feature more easily. As same as classifiers, there are a lot of discretization methods like GUDA-CCC [5], EDISC [6], CD [7] and others [8], [9]. Also, Ameva [10] is the discretization method that it is used in this paper.

It has been confirmed as one of the most promising correlation algorithms due to its reduced execution time and the small number of intervals provided. This behavior is outstanding when the data set has a large number of classes, although it has a slight reduction in the capacity of identification [11].

The other important problem in the classification process is the selection of features [12]. Usually, the obtained experimental data is not filtered about relevant features in systems. A lot of techniques for feature selection [13]-[15] have been developed. Some of these techniques are based on SVM [16] or Naïve Bayes [17].

The Ameva discretization algorithm [10] performs the discretization process effectively and quickly, so the set of values of a feature is greatly reduced. Because Ameva uses the statistic $\mathcal{X}^2$ to determine the relationship between features and classes, it is possible to use this algorithm to determine the relationship between features.

In this paper, a methodology based on Ameva is developed in order to select the main features of a data set. This method exploits the advantages of Ameva in runtime and brings a different approach which was developed on.

The rest of this paper is organized as follows: first, the definition of the problem is presented in Section 2. Also, the Ameva discretization algorithm and the entropy coefficient are presented. Section 3 presents the methodology to determine the best feature selection using the Ameva and the entropy coefficients. Section 4 reports the obtained results of applying the methodology in an example. The paper is finally concluded with a summary of the most important points.

2 Discretization

Let $X = \{x_1, x_2, \dots, x_N\}$ be a data set of a continuous attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of ℓ classes of the variable denoted by

$$\mathcal{C} = \{C_1, C_2, \dots, C_\ell\}, \ell \geq 2$$

A continuous attribute discretization is a function $\mathcal{D}: \mathcal{X} \longrightarrow \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of the property that is being discretized.

Let us consider a discretization $\mathcal{D}$ which discretizes $\mathcal{X}$ into k intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \dots L_k\}$$

where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2,3, \dots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1,2, \dots, N$ and $j = 1,2, \dots, k$. The discretization variable $\mathcal{L}(k)$ of attribute $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view. Having two discrete attributes, a two dimensional frequency table (called contingency table) as show in the Table 1 can be built.

In Table 1, n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j. $n_{i\cdot}$ is the total number of instances belonging to the class C_i, and $n_{\cdot j}$ is the total number of instances that belong to the interval L_j, for $i = 1,2, \dots, \ell$ and $j = 1,2, \dots, k$. So that:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, N = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij} \quad (1)$$

Table 1. Contingency table

$C_i \mid L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell k}$	$n_{\ell\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

2.1 The Ameva discretization

Given Table 1, a discretization criterion based on the Contingency Coefficient ($\mathcal{X}^2$) is defined which measures the independency between class variable $\mathcal{C}$ and discretization variable $\mathcal{L}(k)$.

It is well known in Statistics that two given discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are (statistically) independent if, for all $C_i \epsilon \mathcal{C}$ and $L_j \epsilon \mathcal{L}(k)$,

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, i = 1, \dots, \ell, j = 1, \dots, k$$

that is, no association exists between the two attributes.

Therefore, one way to measure the association (or independency) between class variable $\mathcal{C}$ and discretization variable $\mathcal{L}(k)$ is to analyse the value

$$\sum_{i=1}^{\ell} \sum_{j=1}^{k} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N} \right)^2$$

Nevertheless, it is better to consider a relative measure denoted by $\mathcal{X}^2(k) \stackrel{\text{def}}{=} \mathcal{X}^2(\mathcal{L}(k), \mathcal{C} | \mathcal{X})$:

$$\mathcal{X}^2(k) = \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N} \right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{N}}$$

By using (1), it is not difficult to prove that:

$$\mathcal{X}^2(k) = N \left(-1 + \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} \right) \tag{2}$$

and

$$\max_{X, \mathcal{L}(k), C} \mathcal{X}^2(k) = N \left(\min\{\ell, k\} - 1 \right) \tag{3}$$

In order to compare this coefficient against several discretization variables $\mathcal{L}(k)$ for $k \geq 2$, the Ameva coefficient, $Ameva(k) \stackrel{\text{def}}{=} Ameva(\mathcal{L}(k), \mathcal{C} | \mathcal{X})$, is defined as follows:

$$Ameva(k) = \frac{\mathcal{X}^2(k)}{k(\ell - 1)}$$

For $k, \ell \geq 2$. The Ameva criterion has the following properties:

- The minimum value of $Ameva(k)$ is 0 and when this value is achieved then both discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are statistically independent and viceversa.
- The maximum value of $Ameva(k)$ indicates the best correlation between class labels and discrete intervals. If $k \geq \ell$ then, for all $x \in C_i$ a unique $j0$ exists such that $x \in L_{j0}$ (remaining intervals $(k - \ell)$ have no elements); and if $k < \ell$ then, for all $x \in L_j$, a unique $i0$ exists such that $x \in C_{i0}$ (remaining classes have no elements) i.e. the highest value of the Ameva coefficient is achieved when all values within a particular interval belong to the same associated class for each interval.
- The aggregated value is divided by the number of intervals k, hence the criterion favors discretization schemes with the lowest number of intervals.
- From (3), it is followed that $Ameva_{max}(k) \stackrel{\text{def}}{=} \max_{\mathcal{X},\mathcal{L}(k),\mathcal{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ if $k < \ell$ and $\frac{N}{k}$ otherwise. Hence, $Ameva_{max}(k)$ is an increasing function of k if $k \leq \ell$, and a decreasing function of k if $k > \ell$. Therefore, $\max_{k\geq 2} Ameva_{max}(k) = Ameva_{max}(\ell)$ i.e. the maximum of the Ameva coefficient is achieved in the optimal situation, it is to say, when all values of C_i are in a unique interval L_j and viceversa.

Therefore, the aim of the Ameva method is to maximize the dependence relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k.

2.2 The entropy

If $\ell = 1$ or $k = 1$ then it is not possible to use the Ameva method (Note 1). Let us see these two cases (see Table 2 and Table 3).

Equation (2) can not be calculated using Table 2 because it is not possible to divide by 0. Nevertheless, all the instances belong to the same class, therefore can be concluded that the dependence is maximum. In this case, let us indicate that $A^*(1) = 1$.

Table 2. Contingency table at first case ($\ell = 1$).

$C_i \vert L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N
$n_{\cdot j}$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N

Regarding to Table 3, Ameva method can not be used because $\mathcal{X}^2(k) = 0$ and the Ameva coefficient does not give any information about the dependence. However, the dependence is not minimum and a new coefficient is necessary. By taking into account that if all instances are distributed equally in all classes, the dependence is minimum, and if exists i such that $n_{i1} = N$, the dependence is maximum. Hence the following coefficient, called Entropy, is considered:

$$A(1) = 1 + \frac{1}{N \ln \ell} \sum_{i=1}^{\ell} n_{i1} \ln\left(\frac{n_{i1}}{N}\right)$$

It holds that $0 \leq A(1) \leq 1$, and:

- If $A(1) = 0$, then $n_{i1} = \frac{N}{\ell}$ (minimum dependence).

- If $A(1) = 1$, then a unique n_{i1} exists that $n_{i1} = N$ (maximum dependence).

Table 3. Contingency table at second case ($k = 1$).

$C_i \mid L_j$	L_1	$n_{i\cdot}$
C_1	n_{11}	n_{11}
$\vdots$	$\vdots$	$\vdots$
C_i	n_{i1}	n_{i1}
$\vdots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$n_{\ell 1}$
$n_{\cdot j}$	N	N

3 The methodology

Given an attribute X_i where $i = 1,2,\dots,s$, the Ameva discretization algorithm is applied to this attribute so obtained intervals are considered as a new set of classes. This set of classes is denotes as follows:

$$\mathcal{C}^i = \{C_1^i, C_2^i, \dots, C_{\ell_i}^i\} \tag{4}$$

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathcal{C}^i$ where $p = 1,2,\dots,\ell$. From (4), for each attribute X_j with $j = 1,2,\dots,s$, a g_{ijp} value is obtained from $\mathcal{C}^i$ as follows:

- If the X^p data subset all belong to the same class C^i then $g_{ijp} = A^*(1) = 1$.
- If the subset of data belongs to different classes, then:
  - If values of the attribute X_j are always in the same interval, then $g_{ijp} = A(1)$.
  - If values of the attribute X_j are not always in the same interval, then $g_{ijp} = Ameva_N(\ell_i)$, where $Ameva_N(\ell_i)$ is defined as follows (Note 2):

$$Ameva_N(\ell_i) = \frac{\ell_i'}{N_p} Ameva(\ell_i)$$

provided that N_p is the number of instances of the class X^p and ℓ_i' is the number of intervals of the attribute X_i for which there is at least one value in the data subset.

Given $i, j = 1,2,\dots,s$, a g_{ijp} value can be obtained applying this methodology for all class $C_p \in \mathcal{C}$ ($p = 1,2,\dots,\ell$), and by considering different statistics as follows:

$$g_{ij}^{min} = \min_p g_{ijp}$$

$$g_{ij}^{geo} = \sqrt[\ell]{\prod_{p=1}^{\ell} g_{ijp}}$$

$$g_{ij}^{ari} = \frac{1}{\ell}\sum_{p=1}^{\ell} g_{ijp}$$

$$g_{ij}^{max} = \max_{p} g_{ijp}$$

It is well-known that the following relationship is holded:

$$g_{ij}^{min} \leq g_{ij}^{geo} \leq g_{ij}^{ari} \leq g_{ij}^{max}$$

The main properties of the matrix $G = (g_{ij})$, that is,

$$G = \begin{pmatrix} 1 & g_{12} & \cdots & g_{1s} \\ g_{21} & 1 & \cdots & g_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s1} & g_{s2} & \cdots & g_{ss} \end{pmatrix}$$

are the following: i) it is squared but non symmetric matrix; ii) the values of the main diagonal are 1; iii) $0 \leq g_{ij}, g_{ji} \leq 1$.

From the G matrix, a method of generating rules of dependence between attributes can be defined. For example, a possible rule is the next: given a threshold value, U, if max $\{g_{ij}, g_{ji}\} > U$ and $i < j$ where $i, j = 1,2, \dots, s$ and $i \neq j$, then the X_j variable is eliminated. Let us illustrate it with an example in the next section.

4 Two experiments

Let us consider the Iris Plant Database (Note 3) from UCI Repository which is perhaps the best known database to be found in the pattern recognition literature. This data set is considered due to its simplicity since this methodology is not completely defined yet.

The data set contains four attributes (sepal length, sepal width, petal length and petal width) and three classes (Setosa, Versicolor and Virginica) of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from each other.

The matrices generated by the presented methodology in this paper are:

$$G_{Iris}^{min} = \begin{pmatrix} 1 & 0.4898 & 0.6667 & 0.0998 \\ 0.3265 & 1 & 0.0350 & 0.0930 \\ 0.0280 & 0.0586 & 1 & 0.0303 \\ 0.0545 & 0.0998 & 0.0836 & 1 \end{pmatrix} \tag{5}$$

$$G_{Iris}^{geo} = \begin{pmatrix} 1 & 0.7883 & 0.8736 & 0.4638 \\ 0.6886 & 1 & 0.3271 & 0.4530 \\ 0.1727 & 0.2674 & 1 & 0.1293 \\ 0.1573 & 0.3222 & 0.2244 & 1 \end{pmatrix} \tag{6}$$

$$G_{Iris}^{ari} = \begin{pmatrix} 1 & 0.8299 & 0.8889 & 0.6999 \\ 0.7755 & 1 & 0.6783 & 0.6977 \\ 0.4039 & 0.4617 & 1 & 0.3672 \\ 0.3753 & 0.4783 & 0.4063 & 1 \end{pmatrix} \tag{7}$$

ISSN 1927-6974 E-ISSN 1927-6982

$$G_{Iris}^{max} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad (8)$$

This result shows that it is possible to determine the dependence of attributes of a data set from the Ameva discretization algorithm and the adjustments to resolve the inconsistencies outlined above with the entropy.

The coefficients in the minimum matrix (5) determine the lowest coefficients of dependence between two attributes. These coefficients provide information about there is a class for which the two attributes have less dependency. If these values are high, it is possible to conclude that the dependence between two attributes is high. Therefore, these coefficients are a minimum threshold for each pair of attributes.

A similar conclusion can be obtained from the maximum matrix (8). The coefficients provide information about there is a class for which the two attributes have a high dependence. In this case, these coefficients are the maximum threshold values for each pair of attributes.

Given a data set, the best result is achieved when the maximum and minimum matrix are the same. In this case, all the attributes are the same dependence with other regardless of the original class. Thus, there is only one matrix for generate the discrimination rules.

The geometric mean matrix (6) and the arithmetic mean (7) represent a global value of dependency. While the geometric mean matrix rewards the worst situations about a class, leading to a low value on the global coefficient, the arithmetic mean matrix balances the values of the coefficients.

A possible interpretation to determine which attributes are dependent of each other is to establish a threshold value. From this limit, two attributes are dependent if the average of the coefficients g_{ij} and g_{ji} of the arithmetic mean matrix is greater than or equal to this value.

In this case, the threshold value of 0.75 is established to check which attributes are dependents. The pair g_{ij}, g_{ji} that reaches this threshold is g_{12}, g_{21} because the arithmetic mean of g_{12} and g_{21} is greater than 0.75. It is necessary indicate that the sepal length and the sepal width features are the first and second attributes in the experiment.

Thus, in order to carry out a classification problem can be declared that the X_1 and X_2 features are similar. Let us see this affirmation by using as classification algorithm the Support Vector Machine (SVM) [18].

A performance for the 1-v-r SVM, in the form of accuracy rate, has been evaluated on models using the Gaussian kernel with $\sigma = 1$, and $C = 1$. The criteria employed to estimate the generalized accuracy is the 5-fold cross-validation on the whole set of training data. This procedure is repeated 120 times in order to ensure good statistical behavior. The obtained results are:

- With all features, the accuracy is 0.9320.
- Without the sepal length feature, the accuracy rate is 0.9341.
- Without the sepal width feature, the accuracy rate is 0.9667.

Furthermore to check that the accuracy rate is not less when a feature is eliminated, the methodology has discovered that these features introduce noise in the classification problem when both are used at the same time because the results are improved without the second feature.

Now, consider the Glass Identification (Note 4) Dataset, also from UCI Repository, to prove the methodology with another classification method.

The data set contains nine attributes (refractive index, Sodium, Magnesium, Aluminum, Silicom, Potassium, Calcium, Barium, Iron) (Note 5) and seven classes (building windows float processed, building windows non float processed, vehicle windows float processed, vehicle windows non float processed, containers, tableware and headlamps).

The generated matrices are:

$$G_{Glass}^{min} = \begin{pmatrix} 1 & 0.0864 & 0.1621 & 0.1526 & 0.1828 & 0.0937 & 0.2566 & 0.0200 & 0.0210 \\ 0.0607 & 1 & 0.1244 & 0.2184 & 0.0952 & 0.3224 & 0.0602 & 0.0064 & 0.0502 \\ 0.0968 & 0.1680 & 1 & 0.1577 & 0.1382 & 0.2324 & 0.1359 & 0.0676 & 0.0227 \\ 0.1344 & 0.0496 & 0.1062 & 1 & 0.0680 & 0.2534 & 0.0719 & 0.1239 & 0.0496 \\ 0.0236 & 0.2464 & 0.1361 & 0.0876 & 1 & 0.0972 & 0.0284 & 0.0045 & 0.0372 \\ 0.0416 & 0.3599 & 0.2023 & 0.1897 & 0.1244 & 1 & 0.0422 & 0.0060 & 0.0351 \\ 0.2505 & 0.1293 & 0.2047 & 0.1408 & 0.1755 & 0.3704 & 1 & 0.1247 & 0.0303 \\ 0.1080 & 0.1603 & 0.2230 & 0.3297 & 0.1016 & 0.2199 & 0.0936 & 1 & 0.0450 \\ 0.0606 & 0.0399 & 0.0428 & 0.0463 & 0.0643 & 0.0912 & 0.0493 & 0.0502 & 1 \end{pmatrix}$$

$$G_{Glass}^{geo} = \begin{pmatrix} 1 & 0.1105 & 0.1859 & 0.2392 & 0.2651 & 0.1680 & 0.2925 & 0.0762 & 0.0333 \\ 0.0809 & 1 & 0.2780 & 0.3359 & 0.1195 & 0.4461 & 0.0890 & 0.2031 & 0.0614 \\ 0.1433 & 0.2830 & 1 & 0.3181 & 0.1525 & 0.3355 & 0.2408 & 0.2626 & 0.0320 \\ 0.1755 & 0.2344 & 0.3042 & 1 & 0.0973 & 0.3202 & 0.1173 & 0.4599 & 0.0565 \\ 0.1952 & 0.2976 & 0.2422 & 0.1401 & 1 & 0.2040 & 0.1471 & 0.0593 & 0.0422 \\ 0.0682 & 0.4720 & 0.3455 & 0.2298 & 0.1515 & 1 & 0.0670 & 0.1598 & 0.0588 \\ 0.3566 & 0.1677 & 0.2744 & 0.2813 & 0.2178 & 0.4770 & 1 & 0.1541 & 0.0361 \\ 0.1506 & 0.3618 & 0.3457 & 0.5367 & 0.1609 & 0.2939 & 0.1249 & 1 & 0.0547 \\ 0.2040 & 0.0631 & 0.1552 & 0.1749 & 0.1652 & 0.2299 & 0.3136 & 0.3342 & 1 \end{pmatrix}$$

$$G_{Glass}^{ari} = \begin{pmatrix} 1 & 0.1131 & 0.1873 & 0.2449 & 0.2696 & 0.1757 & 0.2956 & 0.0945 & 0.0346 \\ 0.0832 & 1 & 0.2968 & 0.3421 & 0.1212 & 0.4515 & 0.0934 & 0.3401 & 0.0620 \\ 0.1486 & 0.2924 & 1 & 0.3426 & 0.1540 & 0.3449 & 0.2617 & 0.3011 & 0.0355 \\ 0.1828 & 0.2755 & 0.3369 & 1 & 0.1035 & 0.3263 & 0.1229 & 0.5214 & 0.0567 \\ 0.2630 & 0.3102 & 0.2499 & 0.1439 & 1 & 0.2318 & 0.1779 & 0.0839 & 0.0424 \\ 0.0745 & 0.4759 & 0.3567 & 0.2324 & 0.1534 & 1 & 0.0739 & 0.2609 & 0.0615 \\ 0.3634 & 0.1702 & 0.2865 & 0.3002 & 0.2215 & 0.4807 & 1 & 0.1554 & 0.0379 \\ 0.1621 & 0.3824 & 0.3549 & 0.5497 & 0.2124 & 0.2992 & 0.1269 & 1 & 0.0556 \\ 0.2686 & 0.0662 & 0.2454 & 0.2861 & 0.2110 & 0.2585 & 0.3897 & 0.4178 & 1 \end{pmatrix}$$

$$G_{Glass}^{max} = \begin{pmatrix} 1 & 0.1681 & 0.2279 & 0.3083 & 0.3234 & 0.2379 & 0.3757 & 0.2191 & 0.0494 \\ 0.1135 & 1 & 0.4177 & 0.3998 & 0.1497 & 0.5440 & 0.1529 & 0.4686 & 0.0730 \\ 0.2003 & 0.3868 & 1 & 0.5766 & 0.2017 & 0.4854 & 0.4861 & 0.4013 & 0.0751 \\ 0.3050 & 0.3512 & 0.5092 & 1 & 0.1585 & 0.4457 & 0.1927 & 0.7000 & 0.0642 \\ 0.4974 & 0.5004 & 0.3048 & 0.1836 & 1 & 0.5007 & 0.2904 & 0.1116 & 0.0494 \\ 0.1227 & 0.5554 & 0.4854 & 0.2824 & 0.1910 & 1 & 0.1486 & 0.4153 & 0.0917 \\ 0.4336 & 0.2137 & 0.4890 & 0.4973 & 0.2979 & 0.5689 & 1 & 0.1834 & 0.0692 \\ 0.2842 & 0.4686 & 0.4730 & 0.6650 & 0.6630 & 0.4153 & 0.1603 & 1 & 0.0695 \\ 0.5107 & 0.1041 & 0.5004 & 0.5153 & 0.5034 & 0.5013 & 0.5058 & 0.5287 & 1 \end{pmatrix}$$

As can be seen, the coefficients are lower than the matrices in the previous example. In this case, the threshold value of 0.4 is established to check which attributes are dependents in the arithmetic matrix. The pairs g_{ij}, g_{ji} that reaches this threshold is g_{26}, g_{62} and g_{48}, g_{84} because the arithmetic mean of g_{26} and g_{62}, and g_{48} and g_{84} are greater than 0.4. The Sodium, Aluminum, Potassium and Barium are the second, the fourth, the sixth and the eighth attributes in the experiment.

Thus, the X_2 and X_6 features and X_4 and X_8 features are similar. To prove this affirmation, a K-Nearest Neighbor classification algorithm is used with $k = 3$. The criteria employed to estimate the generalized accuracy is the 10-fold cross-validation on the whole set of training data in this case. The procedure is repeated 120 times. The obtained results are:

ISSN 1927-6974 E-ISSN 1927-6982

- With all features, the accuracy is 0.7152.
- Without the Sodium feature, the accuracy rate is 0.7284.
- Without the Potassium feature, the accuracy rate is 0.7058.
- Without the Aluminum feature, the accuracy rate is 0.6877.
- Without the Barium feature, the accuracy rate is 0.7149.
- Without the Sodium and the Aluminum features, the accuracy rate is 0.6762.
- Without the Sodium and the Barium features, the accuracy rate is 0.7286.

Without the Potassium and the Aluminum features, the accuracy rate is 0.6719.

Without the Potassium and the Barium features, the accuracy rate is 0.7156.

Once more, the methodology has discovered that these features introduce noise in the classification problem when both pairs are used at the same time.

5 Conclusions

We have studied a method of discretization, Ameva, whose objective is to maximize the dependence between the intervals that divide the values of an attribute and the classes to which they belong, providing at the same time the minimum number of intervals.

After that, we have developed a methodology to reduce the number of features of a data set based on the dependence between them. To the best of knowledge, there are not existing researches that directly address the problem to reduce the number of features using an approach similar to ours.

This development is based on taking advantage of Ameva discretization algorithm. Thus, a new coefficient has been developed to determine the dependence between features. Hence, we have reduced the number of values of features and the number of features from a qualitative reasoning.

To test the development of the methodology, it has been applied to two well-known data sets to obtain the dependent relationship between their features. Nevertheless, we think that this approach can be satisfactorily applied in this area when the data set has a lot of instances and features, and one of these features determines the class which each instance belongs to. Another data sets must fulfill these characteristics.

Finally, after applying the discrimination of features obtained in the methodology, the modified data sets have been carried out for the classification tests to verify the effectiveness of the methodology.

The next step to complement this development is the design of an automatic method of creation of feature discrimination rules. Subsequently, we must define some improvements in this methodology to automatically know the dependence between features without setting manually a threshold value.

Acknowledgement

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness ARTEMISA (TIN2009-14378-C02-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

References

[1] Ghoggali, N., Melgani, F. and Bazi, Y. "A multiobjective genetic SVM approach for classification problems with limited training samples". IEEE Transactions on Geoscience and Remote Sensing. 2009; 47 (6): 1707-1718. http://dx.doi.org/10.1109/TGRS.2008.2007128

[2] Stojadinovic, A., Potter, B.K., and et al. Development of a prognostic naive bayesian classifier for successful treatment of nonunions. The Journal of Bone and Joint Surgery (American). 2011; 93 (2): 187-194. http://dx.doi.org/10.2106/JBJS.I.01649

[3] Jiang, S.Y. and Yu, W. A combination classification algorithm based on outlier detection and C4.5. Advanced Data Mining and Applications. 2009; 5678: 504-511. http://dx.doi.org/10.1007/978-3-642-03348-3_50

[4] Entezari-Maleki, R., Iranmanesh, S.M. and Minaei-Bidgoli, B. An experimental investigation of the effect of discrete attributes on the precision of classification methods. World Applied Sciences Journal. 2009; 216-223. http://dx.doi.org/10.1109/ICICT.2009.5267189

[5] Zeng, A., Gao, Q. and Pan, D. A global unsupervised data discretization algorithm based on collective correlation coefficient. Modern Approaches in Applied Intelligence. 2011; 6703: 146-155. http://dx.doi.org/10.1007/978-3-642-21822-4_16

[6] Shehzad, K. EDISC: a class-tailored discretization technique for rule-based classification. IEEE Transactions on Knowledge and Data Engineering. 2012; 24 (8): 1435-1447. http://dx.doi.org/10.1109/TKDE.2011.101

[7] Wang, C., Wang, M., She, Z. and Cao, L. CD: a coupled discretization algorithm. Advances in Knowledge Discovery and Data Mining. 2012; 7302: 407-418. http://dx.doi.org/10.1007/978-3-642-30220-6_34

[8] Jiang, F., Zhao, Z. and Ge, Y. A supervised and multivariate discretization algorithm for rough sets. Rough Set and Knowledge Technology. 2010; 6401: 596-603. http://dx.doi.org/10.1007/978-3-642-16248-0_81

[9] Zhao, J., Han, C.Z., Wei, B. and Han, D.Q. A UMDA-based discretization method for continuous attributes. Advanced Materials Research. 2012; 403-408: 1834-1838. http://dx.doi.org/10.4028/www.scientific.net/AMR.403-408.1834

[10] González, L., Cuberos, F.J., Velasco, F. and Ortega, J.A. Ameva: an autonomous discretization algorithm. Expert Systems with Applications. 2009; 36 (3): 5327-5332. http://dx.doi.org/10.1016/j.eswa.2008.06.063

[11] González, L., Cuberos, F.J., Velasco, F. and Ortega, J.A. A new approach to qualitative learning in time series. Expert Systems with Applications. 2009; 36 (6): 9924-9927. http://dx.doi.org/10.1016/j.eswa.2009.01.066

[12] Saeys, Y., Inza, I. and Larrañaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23 (19): 2507-2517. http://dx.doi.org/10.1093/bioinformatics/btm344

[13] Hua, J., Tembe, W.D. and Dougherty, E.R. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition. 2009; 42 (3): 409-424. http://dx.doi.org/10.1016/j.patcog.2008.08.001

[14] Witten, D.M. and Tibshirani, R. A framework for feature selection in clustering. Journal of the American Statistical Association. 2010; 105 (490): 713-726. http://dx.doi.org/10.1198/jasa.2010.tm09415

[15] Ma, Y. and Zhan, L. Research on the evaluation of feature selection based on SVM. Informatics in Control, Automation and Robotics. 2012; 133: 407-414. http://dx.doi.org/10.1007/978-3-642-25992-0_57

[16] Maldonado, S. and Weber, R. A wrapper method for feature selection using Support Vector Machines. Information Sciences: an International Journal. 2009; 179 (13): 2208-2217. http://dx.doi.org/10.1016/j.ins.2009.02.014

[17] Chen, J., Huang, H., Tian, S. and Qu, Y. Feature selection for text classification with Naive Bayes. Expert Systems with Applications. 2009; 36 (3): 5432-5435. http://dx.doi.org/10.1016/j.eswa.2008.06.054

[18] González, L., Angulo, C., Velasco, F. and Català, A. Dual unification of bi-class support vector machine formulations. Pattern recognition. 2006; 39 (7): 1325-1332. http://dx.doi.org/10.1016/j.patcog.2006.01.007

CHAPTER 11

ACTIVITY RECOGNITION SYSTEM USING NON-INTRUSIVE DEVICES THROUGH A COMPLEMENTARY TECHNIQUE BASED ON DISCRETE METHODS

Overview

This paper aims to develop an innovative selection, discretization and classification technique to make the recognition process in an efficient way and at low energy cost. It controls the physical activity carried out by the user based on Ameva discretization. The entire process is executed on the smartphone and on a wireless health monitoring system is used when the smartphone is not used taking into account the system energy consumption.

Moreover, the conclusion reached in terms of energy consumption and the processing cost of the system when it is working on a mobile device is that the method based on Ameva reduces the computational cost of the system by about 50%.

The classification cost using discrete variables is much lower than working with continuous variables because it is possible to eliminate the correlation between variables during the recognition process and on the other hand, to minimize the energy consumption from the process.

In contrast, the number of activities recognized is limited because working only with the smartphone's sensors (accelerometer and barometer) limits the number of variables that can be used.

In order to improve the accuracy problems encountered during the celebration of the Evaluating AAL Systems Through Competitive Benchmarking (EvAAL) competition in 2012, some significant improvements in Ameva discretization algorithm are proposed. Also, in addition to detect specific activities, the barometric sensor which is being included in the latest generation of mobile devices is used. Finally, in order to answer the question about what would happen if you decide not to use your mobile device in an indoor environment, as happens in real life, a complementary wireless device is also optionally used.

The approach was tested in a real scenario during the EvAAL competition in 2013 and it was a great chance to make a real stress test of AMEVA system and checked if the improvements really got best accuracy that the previous year.

Context

This research was the continuation of the previous work in the discretization and activity recognition areas. It was based on the advantage of Ameva in terms of computation performance and pass it into the activity recognition approaches for smartphones getting a significant reduction of battery draining. Also, it improved the previous algorithm and it was tested during the EvAAL competition. This paper is the result of over 1 year's work and currently is one of the areas under investigation.

Research work information

The *Evaluating AAL Systems Through Competitive Benchmarking* conference aims at establishing benchmarks and evaluation metrics for comparing Ambient Assisted Living solutions. Since 2011 it has organised international competitions on indoor localization and indoor activity recognition.

It is indexed in SJR with a Impact Factor of 0.134 (Q4) and H Index of 19 in one category: Computer Science (miscellaneous).

DOI: https://doi.org/10.1007/978-3-642-41043-7_4

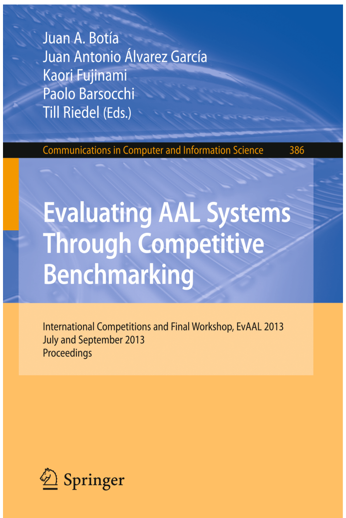


Figure 11.1: Evaluating AAL Systems Through Competitive Benchmarking cover.

Activity Recognition System Using Non-intrusive Devices through a Complementary Technique Based on Discrete Methods

Miguel Ángel Álvarez de la Concepción[1], Luis Miguel Soria Morillo[1], Luis González Abril[2], and Juan Antonio Ortega Ramírez[1]

[1] Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain
{maalvarez,lsoria,jortega}@us.es

[2] Applied Economics I Dept., University of Seville, 41018 Seville, Spain
luisgon@us.es

Abstract. This paper aims to develop a cheap, comfortable and, specially, efficient system which controls the physical activity carried out by the user. For this purpose an extended approach to physical activity recognition is presented, based on the use of discrete variables which employ data from accelerometer sensors. To this end, an innovative selection, discretization and classification technique to make the recognition process in an efficient way and at low energy cost, is presented in this work based on Ameva discretization. Entire process is executed on the smartphone and on a wireless health monitoring system is used when the smartphone is not used taking into account the system energy consumption.

Keywords: Contextual Information, Discretization Method, Mobile Environment, Qualitative Systems, Smart-Energy Computing.

1 Introduction

In recent years, thanks largely to the increased interest on monitoring certain sectors of population such as elderly people with dementia or people in rehabilitation, activity recognition systems have experienced an increase in both number and quality results. However, most of them are in a high computational cost and hence, it cannot be executed into a general purpose mobile device.

Calculation of the physical activity of a user based on data obtained from an accelerometer is a current research topic. Furthermore, many works is going to be analyzed showing some identified limitations that make these systems uncomfortable for users in general.

The first difference observed between the systems developed is the type of used sensor. There are systems using specific hardware [1], while others use general purpose hardware [2]. Obviously, the use of generic hardware is a benefit for users, since the cost of devices and versatility of them are points in their favor. Not to mention decreasing the loss and forgetting risk due to they have been integrated on an everyday object like users' smartphones.

J.A. Botía et al. (Eds.): EvAAL 2013, CCIS 386, pp. 36–47, 2013.

Another difference found between the surveyed proposals is the number and position of the sensors. In [3] can be seen that the accelerometer sensor is placed in a glove and a multitude of activities depending on the movement of the hand are recognized. In contrast, other studies use various sensors throughout the body [4], [5] or a wearable wireless sensor node with a static wireless non-intrusive sensory infrastructure [6] to recognize these activities. According to some comparative studies and previous works based on multiple sensors, they are more accurate.

Although, works like [2], where a sensor is at users' pocket or in the hipe, is more comfortable for them. By this way, place them in the monitored person is easier, not to mention that the infrastructure is much lower.

Thus, the presented work will is focus on the recognition of physical activities carried out by users throughout their mobile devices. So, it must be paid special attention to energy consumption and computational cost of used methods. Also, a wireless health monitoring system can be used to increment the user acceptance, i.e. the user does not carry the mobile devices all the time in an indoor environment.

One step further, some works do not only use data from accelerometers, but use other sources such as microphone, light sensor or voice recognition to determine the context of the user [7]. However, they present problems i.e. when the environment is noisy or the user is alone.

There are related works where data for activities recognition are obtained through mobile devices, but these data are sent to a server to process the information [8]. Thus, computational cost is not a handicap and because of this more complex methods are used. In contrast, the efficiency is a crucial issue when processing is carried out in the mobile device [9], [10].

To reduce the cost associated to accelerometer signal analysis, this paper opts for a novel approach based on a discretization method. Thanks to discretization process, classification cost is much lower than working with continuous variables. Because of this, it is possible to eliminate the correlation between variables during the recognition process and on the other hand, to minimize the energy consumption from the process.

Working in the domain of discrete variables to perform learning and recognition of activities is a new approach offered by this work. This decision was largely due to the high computational cost required for learning algorithms based on continuous variables used for this purpose over the years.

In [11], a labeling process, like a discretization process, is used to obtain a Qualitative Similarity Index (QSI), so it can be said that a transformation of the continuous domain to the discrete domain of values of the variables is beneficial in certain aspects.

But, before the self-recognition or learning, it is necessary to carry out a process of Ameva discretization from its algorithm [12]. It has a number of advantages over other well-known discretization algorithms like CAIM discretization algorithm [13], i.e. it is unsupervised and very fast. The most notable of these is the small number of intervals generated which facilitates and reduces the computational cost of the recognition process.

It should be noted that many of these studies could be seen in action during the competition EvAAL 2012 [14] in Activity recognition track. EvAAL is an annual international competition that addresses the challenge of evaluation and comparison of Ambient Assisted Living (AAL) systems and platforms, with the final goal to assess the autonomy, independent living and quality of life that AAL systems may grant to their end users.

In this track competition, four teams participated in the challenge: CUJ (from the University of Chiba, Japan) [15], CMU (from Carnegie Mellon and Utah Universities, USA) [16], DCU (from Dublin City University, Ireland) [17] and USS (from University of Seville, Spain) [12]. Finally, although CMU had the best accuracy in the results, USS won the competition because its simplicity and interoperability gave good marks in all the evaluated criteria.

In order to improve the accuracy problems encountered during the celebration of the EvAAL 2012 competition, some significant improvements in Ameva discretization algorithm are proposed. Also, in addition to detect specific activities, the barometric sensor which is being included in the latest generation of mobile devices is used.

Finally, in order to answer the question about what would happen if you decide not to use your mobile device in an indoor environment, as happens in real life, a complementary wireless device is also optionally used.

There are other similar EvAAL competitions such as HARL [18], OPPORTUNITY [19], HASC [20] or BSN contest [21].

The paper is organized as follows: first, the activity recognition step is presented in Section 2. Also, the data collection and the set of activities are presented. Section 3 presents the methodology to determine the activity using the Ameva discretization. Section 4 reports the obtained results of applying the methodology. Finally, the paper conclusions with a summary of the most important points are in Section 5.

2 Activity Recognition

The final real system consist only of a smartphone and, optionally, a wireless device, configured to detect the competition activities: lie, sit, stand, walk, bend, fall and cycle.

2.1 Data Collection

In contrast to the needs of some studies that require a training set to classify a recognized activity correctly, this paper reduces the waiting time for recognition, providing valid information for an activity frequently.

To this end, a training set and a recognition set are obtained using 5-second-time windows of fixed duration which has been determined empirically as optimum length from a performance and an accuracy analysis of the system.

The time length of five seconds of these windows has been chosen because for our system is very important to ensure that in each time window there is at least one cycle of activity, where activity cycle is defined as a complete execution of some activity patterns. For example, two steps are a walking activity cycle and one pedal stroke is the activity cycle for cycling. If at least one cycle of activity can not be guaranteed in each time window, it is not possible to determine the activity from accelerometer patterns.

This analysis is performed based on the values obtained from the accelerometer, which significantly improve the precision of the body-related activities, and a barometer to detect environment-related activities, such as going upstairs and downstairs. The latter sensor has most often been integrated in recent mobile devices, allow to increase the overall system accuracy detection of activities.

So, based on these time windows that contain data for each accelerometer axis and reducing the computational cost of the new solution, signal module has been chosen to work. This eliminates the problem caused by the device rotation [22]. Furthermore, it increases user comfort with the system by removing the restriction to keep the orientation during the learning and recognition process.

For each data in a time window size N, $a_i = (a_i^x, a_i^y, a_i^z)$, $i = 1, 2, \ldots, N$ where x, y and z represent the three accelerometer axis, the accelerometer module is defined as follow:

$$|a_i| = \sqrt{(a_i^x)^2 + (a_i^y)^2 + (a_i^z)^2}$$

Hence, the arithmetic mean, the minimum, the maximum, the median, the standard and the mean deviation, and the signal magnitude area statistics are obtained for each time window.

In addition to the above variables, hereafter called temporary variables, a new set of statistics called frequency-domain features from the frequency domain of the problem are generated. Thus, in order to obtain the frequency-domain features, Fast Fourier Transform (FFT) is applied for each time window.

For the barometer sensor, two measures are obtained for each time window: at the beginning and at the end, taking into account the difference between them.

$$b = b_N - b_1$$

It is important to note that in this case, the absolute value is not taken into account, contrary to what was done with the values obtained from the accelerometer.

2.2 Set of Activities

Far from being a static system, the number and type of activities recognized by the system depends on the user. Thanks to this proposal when users is carrying out activities that have not been learned before can be determined. This is achieved basing on the analysis of probability associated to each pattern while user is performing the activities. Obviously, the number of activities to be detected will impact on the accuracy of the system. Especially if acceleration patterns between activities are very similar.

For a large numbers of users could be interesting recognize a few activities, such as walking, sitting and falling. But for another users, activities like driving or biking would be important. However, to carry out a comparative analysis of the accuracy and performance of the discrete recognition method proposed below, 8 activities were taken into account. These activities are immobile, walking, running, jumping, cycling, drive, walking-upstairs and walking-downstairs.

Therefore, the learning system allows the user to decide what activities he/she wants the system to recognize. This is highly useful when the determination of certain very specific activities on monitored users is required.

3 Methodology

3.1 Ameva Algorithm

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a data set of an attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of the ℓ classes of class variable denoted by

$$\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}, \ell \geq 2$$

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \to \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in X$ in the domain of property that is being discretized. Let us consider a discretization $\mathcal{D}$ which discretizes $\mathcal{X}$ into k discrete intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \ldots, L_k\}$$

where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j]$, $j = 2, 3, \ldots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such $x_i \in L_j$ that for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$. The discretization variable $\mathcal{L}(k)$ of $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view.

The main aim of the Ameva method [12] is to maximize the dependency relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k. For this, the following statistic is used:

$$Ameva(k) = \frac{\chi^2(k)}{k(\ell - 1)} \; where \; \chi^2(k) = N \left(-1 + \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{n_{ij}^2}{n_{\cdot i} n_{j \cdot}} \right)$$

and n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j, $n_{i\cdot}$ is the total number of instances belonging to the class C_i and $n_{\cdot j}$ is the total number of instances that belong to the interval L_j, for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$, fulfilling the following:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}, \quad N = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij}$$

The original developed algorithm to obtain the best intervals with the Ameva discretization is based on finding the cutoff points that provide the best coefficient. To do this, the values of the variables are sorted to find the first cut (local maximum). Then, it returns the next cut, and so on, until the Ameva coefficient does not improve. This behavior causes the complexity of the algorithm is quadratic order, $O(n^2)$. A graphic with three local maximums can be seen in Figure 1.

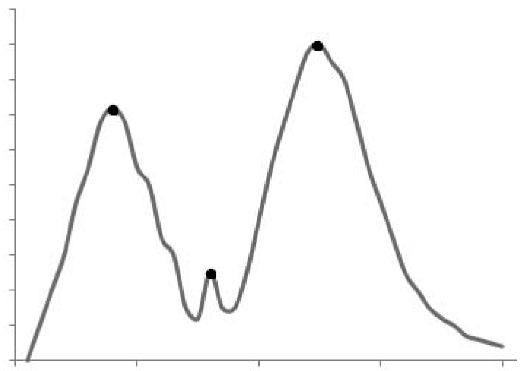


Fig. 1. An example of Ameva coefficient values with three local maximums

The presented improvement in this work allows to find all cuts, allowing the complexity of the algorithm would be of linear order, $O(n)$. Although there is a loss of precision, it is negligible for the field of study of this work, since it allows to obtain good results.

Finally, for each statistical $S_p \in \{S_1, S_2, \ldots, S_m\}$, the discretization process is performed, obtaining a matrix of order $k_p \times 2$, where k_p is the number of class intervals and 2 denotes the $inf(L_i^p)$ and $sup(L_i^p)$ interval limits i of p statistical. Hence, a three-dimensional matrix containing the statistics and the set of interval limits for each statistic is called Discretization Matrix and it is denoted by

$$\mathcal{W} = (w_{pij})$$

where $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2$.

Therefore, Discretization Matrix determines the interval at which each data belongs to the different statistical associated values, carrying out a simple and fast discretization process.

Class Integration. The aim in the next step of the algorithm is to provide a probability associated with the statistical data for each of the activities based on previously generated intervals. For this purpose, the elements of the training set $x \in X$ are processed to associate the label of the concrete activity in the training set. In addition, the value of each statistic is calculated based on the time window.

For carrying out the previous process, a Class Matrix, $\mathcal{V}$, is defined as a three-dimensional matrix that contains the number of data from the training set associated with a $\mathcal{L}$ interval in a $\mathcal{C}$ activity for each statistical $\mathcal{S}$ of the system. This matrix is defined as follows:

$$\mathcal{V} = (v_{pij})$$

where $v_{pij} = \#\{x \in X \mid inf(L_i^p) < x \leq sup(L_i^p)\}$, and $\mathcal{S} = S_p$, $\mathcal{C} = C_j$, $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$.

So, each position in the Class Matrix is uniquely associated with a position in the Discretization Matrix determined by its range.

At this point, there is not only possible to determine the discretization interval, but the Class Matrix helps to obtain the probability associated with the discretization process performed with the Ameva algorithm.

Activity-interval Matrix. The next step is determined a three-dimensional matrix, called Activity-Interval Matrix and denoted by $\mathcal{U}$, which determines the likelihood that a given value x associated to a S statistical corresponds to $\mathcal{C}$ activity in a $\mathcal{L}$ interval. This ratio is based on obtaining the goodness of the Ameva discretization and the aim is to determine the most probable activity from the data and the intervals generated for the training set.

Each value of $\mathcal{U}$ is defined as follows:

$$u_{pij} = \frac{v_{pij}}{v_{p\cdot j}} \frac{\sum_{q=1, q\neq j}^{\ell} \left(1 - \frac{v_{piq}}{v_{p\cdot q}}\right)}{\ell - 1}$$

where $v_{p\cdot j}$ is the total number of time windows of the training process labeled with the j activity for the p statistic, and $p = 1, 2, \ldots, m$, $i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, \ell$

Given these values, $\mathcal{U}$ for the p statistic is defined as

$$\mathcal{U}_p = \begin{pmatrix} u_{p00} & \ldots & u_{p0j} & \ldots & u_{p0\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pi0} & \ldots & u_{pij} & \ldots & u_{pi\ell} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{pk_p0} & \ldots & u_{pk_pj} & \ldots & u_{pk_p\ell} \end{pmatrix}$$

As can be seen in the definition of $\mathcal{U}$, the likelihood that a data x is associated with the interval L_i corresponding to the activity C_j, depends not only on data, but all the elements associated with the interval L_i for the other activities.

Thus, each u_{pij} matrix position can be seen as a grade of belonging that a given x is identified to C_j activity, that it is included in the L_i interval of the S_p statistic.

Similarly, the elements of $\mathcal{U}$ have the following properties:

– $u_{pij} = 0 \iff v_{pij} = 0 \vee v_{piq} = v_{p\cdot q}, q \neq j$
– $u_{pij} = 1 \iff v_{pij} = v_{p\cdot j} = v_{pi\cdot}$

Figure 2 shows the overall process described on this section for carry on data analysis and interval determination.

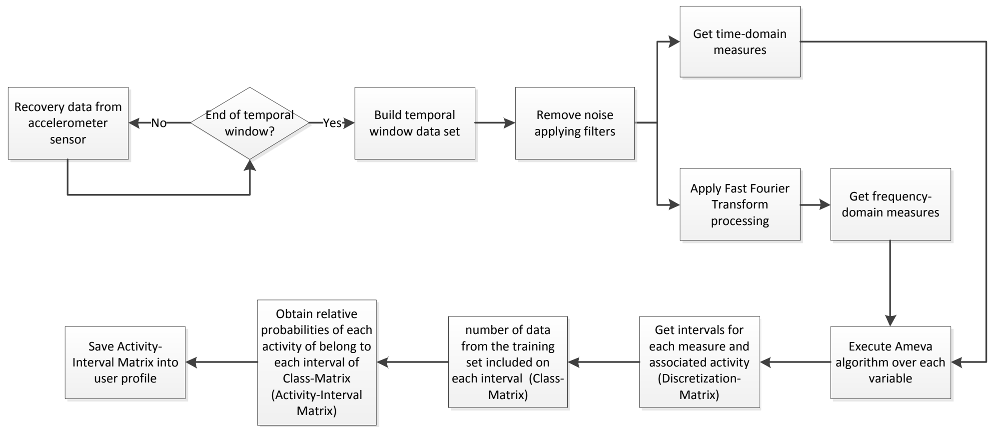


Fig. 2. Overall process of data analysis and interval determination

3.2 Classification Process

Having obtained the discretization intervals and the probabilities of belonging to each interval, the process by which the classification is performed can be described. This classification is based on data from the analysis of time windows. The process is divided into two main steps: the way in which to perform the recognition of physical activity is first described; and the process to determine the frequency at which some particular activity is then presented.

Classifying Data. For the classification process, the more likely activity is decided by a majority voting system. As said above, this process parts from the Activity-Interval Matrix and a set of data $x \in X$ for the $\mathcal{S}$ set.

Therefore, it consists in finding an activity $C_i \in \mathcal{C}$ that maximizes the likelihood. The above criterion is collected in the following expression, denoted by mpa (most likely activity):

$$mpa(x) = C_k$$

where $k = arg(\max_j \sum_{p=1}^{m} u_{pij} \mid x \in (inf(L_i^p), sup(L_i^p)])$. The expression shows that the weight contributed by each statistical to the likely calculation function is the same. This can be done under the assumption that all statistical provide the same information to the system and there is not correlation between them.

Thus, the *mpa* represents the activity whose data, obtained through the processing time window, is more suited to the value set from $\mathcal{U}$. In this way, the proposed algorithm not only determine the *mpa*, but its associated probability.

From this likelihood, certain activities that do not adapt well to sets of generic classification can be identified. It is an indication that user is carrying out new activities for which the system has not been trained previously.

Figure 3 shows the overall process described on this section for recognition process from Activity-Interval Matrix calculated in the previous stage.

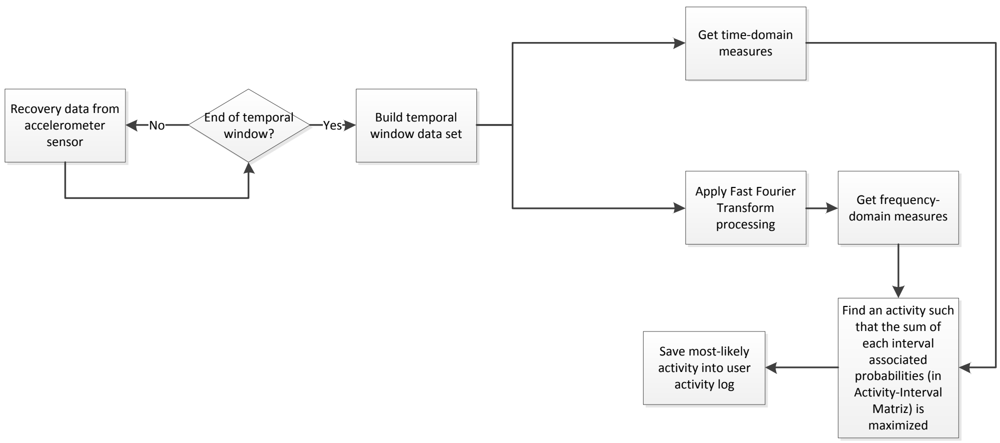


Fig. 3. Overall recognition process from data sensors

4 Method Analysis

Once exposed the bases of the developed activities recognition algorithm, an analysis of the new proposal was performed. To do this, the new development was compared with a recognition system widely used based on neural network. In this case, both learning and recognition was performed by continuous methods.

The test process was conducted in a Google Nexus One for a group of 10 users. Notably, the activity habits of these users were radically different, since 5 of them were under 30 years while the rest were older than this age. For this purpose, a document was delivered to each user for describing the activity performed, start time and end time.

Finally, the learning process consisted on the performing of each activity recognized by the system for a time of 6 minutes. As for the recognition process, users were followed over a period of 72 hours.

Moreover, the energy consumption and the processing cost of the system when it is working on a mobile device are considered. In this case, the conclusion reached is that the method based on Ameva reduces the computational cost of the system by about 50% (see Figure 4. The time needed to process a time window by using nueral networks methods is 1.2 seconds, while, for the Ameva-based method is 0.6 seconds.

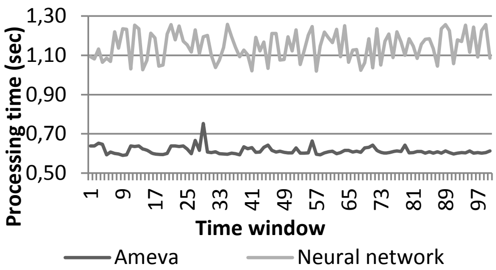


Fig. 4. Processing time of the Ameva and neural network methods on the device

As can be seen in 5, Ameva battery consumption is lower than neural networks. For the first one, the battery lifetime is close to 25 hours while for the last one, it's only 16 hours. In the comparison can be observed the battery lifetime for decision tree but the main problem of this method, based on statistics chosen, is the low accuracy, not higher than 60%.

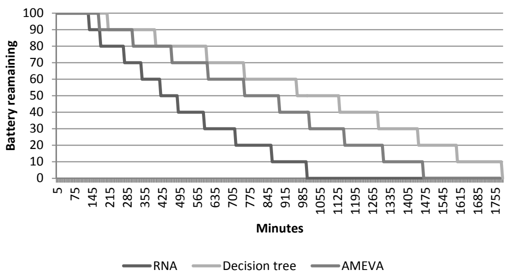


Fig. 5. Battery life for Ameva compared to neural network and decision tree methods

Based on Accuracy, Recall, Specificity, Precision, and F measure, Table 1 is presented. In this table, differences between the two methods, RNA and Ameva can be observed. Most values presented for each measure and activity show that the Ameva method performs better than RNA, especially as regards precision.

Table 1. Performance comparison by using measures of evaluation

Measure / Activity	Accuracy		Recall		Specificity		Precision		F-measure (F_1)	
	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA	Ameva	RNA
Walk	98.77%	97.93%	97.92%	93.95%	98.91%	98.57%	93.50%	91.36%	95.66%	92.64%
Upstairs	98.93%	98.17%	95.40%	90.79%	99.43%	99.22%	96.00%	94.35%	95.70%	92.54%
Downstairs	98.64%	98.25%	95.20%	92.68%	99.04%	98.89%	91.95%	90.62%	93.55%	91.64%
Cycle	99.32%	99.03%	96.13%	95.67%	99.73%	99.47%	97.91%	95.89%	97.01%	95.78%
Immobile	98.69%	99.50%	94.57%	97.37%	99.42%	99.88%	96.60%	99.29%	95.58%	98.32%

5 Conclusions

In this work, a recognition system based only on a smartphone and, optionally, a wireless device is presented obtaining very good results. It should be noted that the system does not have communication with a server, thus it does not affect too much to de battery duration life.

Also, the Ameva discretization algorithm has been modified in order to improved the accuracy to obtain best results as the last implemented system. It has therefore been possible to achieve an average accuracy of 98% for the recognition of 7 types of activities.

In contrast, the number of activities that the system can recognize is limited, because working only with accelerometer and barometer limits the number of system variables that can be used, that it can cause that the correlation between these variables tends to be high.

Acknowledgments. This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness ARTEMISA (TIN2009-14378-C02-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy.

References

1. Ravi, N., Dandekar, N., Mysore, P., Littman, M.: Activity recognition from accelerometer data. In: Proceedings of the National Conference on Artificial Intelligence, vol. 20, p. 1541. AAAI Press, MIT Press, Menlo Park, Cambridge (2005)
2. Hong, Y., Kim, I., Ahn, S., Kim, H.: Activity recognition using wearable sensors for elder care. In: Second International Conference on Future Generation Communication and Networking, vol. 2, pp. 302–305. IEEE (2008)
3. Brezmes, T., Gorricho, J.-L., Cotrina, J.: Activity recognition from accelerometer data on a mobile phone. In: Omatu, S., Rocha, M.P., Bravo, J., Fernández, F., Corchado, E., Bustillo, A., Corchado, J.M. (eds.) IWANN 2009, Part II. LNCS, vol. 5518, pp. 796–799. Springer, Heidelberg (2009)
4. Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., Zancanaro, M.: What is happening now? Detection of activities of daily living from simple visual features. Personal and Ubiquitous Computing 14(8), 749–766 (2010)
5. Bicocchi, N., Mamei, M., Zambonelli, F.: Detecting activities from body-worn accelerometers via instance-based algorithms. Pervasive and Mobile Computing 6(4), 482–495 (2010)
6. Paoli, R., Fernández-Luque, F., Zapata, J.: A system for ubiquitous fall monitoring at home via a wireless sensor network and a wearable mote. Expert Systems with Applications (2011)
7. Kwapisz, J., Weiss, G., Moore, S.: Activity recognition using cell phone accelerometers. ACM SIGKDD Explorations Newsletter 12(2), 74–82 (2011)
8. Altun, K., Barshan, B., Tunçel, O.: Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognition 43(10), 3605–3620 (2010)
9. Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks 6(2), 13 (2010)
10. Fuentes, D., Gonzalez-Abril, L., Angulo, C., Ortega, J.: Online motion recognition using an accelerometer in a mobile device. Expert Systems with Applications 39(3), 2461–2465 (2012)

11. Cuberos, F., Ortega, J., Velasco, F., González, L.: Qsi-alternative labelling and noise sensitivity. In: 17th International Workshop on Qualitative Reasoning (2003)
12. Gonzalez-Abril, L., Cuberos, F., Velasco, F., Ortega, J.: Ameva: An autonomous discretization algorithm. Expert Systems with Applications 36(3), 5327–5332 (2009)
13. Kurgan, L., Cios, K.: Caim discretization algorithm. IEEE Transactions on Knowledge and Data Engineering 16(2), 145–153 (2004)
14. Álvarez García, J.A., Barsocchi, P., Chessa, S., Salvi, D.: Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 evaal competition. Journal of Ambient Intelligence and Smart Environments 5(1), 119–132 (2013)
15. Nergui, M., Yoshida, Y., Gonzalez, J., Koike, Y., Sekine, M., Yu, W.: Human motion tracking and measurement by a mobile robot. In: 7th International Conference on Intelligent Unmanned Systems (2011)
16. Hong, J.H., Ramos, J., Dey, A.K.: Understanding physiological responses to stressors during physical activity. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 270–279. ACM (2012)
17. Li, N., Crane, M., Ruskin, H.J.: Automatically detecting "significant events" on sensecam. Ercim News 2011(87) (2011)
18. Wolf, C., Mille, J., Lombardi, E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandréa, E., Bichot, C.E., Garcia, C., Sankur, B.: The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. Technical Report LIRIS RR-2012-004, Laboratoire d'Informatique en Images et Systmes d'Information, INSA de Lyon, France (2012)
19. Sagha, H., Digumarti, S.T., del Millan, J., Chavarriaga, R., Calatroni, A., Roggen, D., Troster, G.: Benchmarking classification techniques using the opportunity human activity dataset. In: 2011 IEEE International Conference on Systems, Man and Cybernetics, pp. 36–40. IEEE (2011)
20. Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada, T., Murao, K., Inoue, S., Kawahara, Y., Sumi, Y., Nishio, N.: Hasc challenge: gathering large scale human activity corpus for the real-world activity understandings. In: Proceedings of the 2nd Augmented Human International Conference, p. 27. ACM (2011)
21. Loseu, V., Jafari, R.: Power aware wireless data collection for bsn data repositories. In: 2011 International Conference on Body Sensor Networks, pp. 19–21. IEEE (2011)
22. He, Z., Jin, L.: Activity recognition from acceleration data based on discrete consine transform and svm. In: 2009 IEEE International Conference on Systems, Man and Cybernetics, pp. 5041–5044. IEEE (2009)

CHAPTER 12

EVALUATING WEARABLE ACTIVITY RECOGNITION AND FALL DETECTION SYSTEMS

Overview

This paper proposes a protocol that fuses activity recognition (AR) and fall detection (FD) research areas to achieve a large, open and growing dataset that could, potentially, provide an enhanced understanding of the activities and fall process and the information needed to design and evaluate high-performance systems.

Some repositories contain datasets combining AR and FD but reported in and AR manner, missing some important information from FD point of view like demographics or context. We defined a protocol focused on create an open and growing AR-FD dataset with simple and complex activities and, more important, simulated and real falls.

Due to the fact that current mobile phones include triaxial accelerometers, gyroscopes and magnetometers and are wearable and ubiquitous devices, one or more are used to acquire the data. Also, to avoid only one configuration, all the positions are detailed in the dataset, so some features will be present and others no. In addition, a video recording, time-stamped and synchronized to the acquisition data mobile phones must be included to the dataset because some information are hard to report from non-medical people.

From the previous information, the missing information can be completed analysing the video and even discard some not suitable events. Also, more instances can be added to the dataset by uploading the video and the generated files to a server.

Fusing FD and AR must support active independent living in aged people achieving two main challenges: having public datasets in realistic environments with a rich configuration of sensors and a good description of users, sensors, activities and falls and environment; and evaluating multiple algorithms/systems through either live competitions or new and independent datasets. Competitions are good examples of how to evaluate different hardware and software approaches with the same goal, but it requires an investment of both time and money.

This high replicable protocol creates an open, flexible, growing and community maintained dataset of multiple activities and simulated or real falls. Also, it allows a software competition based on multiple hardware configurations using diverse positions of the acquisition tools to use it.

Context

This research was the result of a decision about covering an empty area in AR-FD databases. It was based on the need of supporting common and standarized information for researchers and non-medical people in order to compare the result between different activity recognition and fall detections systems. This paper is the result of over more than 3 year's work and currently is used as an standard protocol for the activity recognition and fall detections areas of investigation.

Research work information

The *International Federation for Medical and Biological Engineering* is primarily a federation of national and transnational societies. These professional organizations represent interests in medical and biological engineering. Its objectives are scientific and technological as well as educational and literary and it encourages research and application of knowledge, disseminates information and promotes collaboration.

It is indexed in SJR with a Impact Factor of 0.118 and H Index of 12 in two categories: Bioengineering; Biomedical Engineering.

DOI: https://doi.org/10.1007/978-3-319-11128-5_163

Figure 12.1: International Federation for Medical and Biological Engineering cover.

Evaluating Wearable Activity Recognition and Fall Detection Systems

Juan Antonio Álvarez-García, Luis Miguel Soria Morillo, Miguel Ángel Álvarez de La Concepción, Alejandro Fernández-Montes, and Juan Antonio Ortega Ramírez

Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain

Abstract— **Activity recognition (AR) and fall detection (FD) research areas are very related in assistance scenarios but evolve independently. Evaluate them is not trivial and the lack of FD real-world datasets implies a big issue. A protocol that fuses AR and FD is proposed to achieve a large, open and growing dataset that could, potentially, provide an enhanced understanding of the activities and fall process and the information needed to design and evaluate high-performance systems.**

Keywords— **AAL, Activity recognition, Fall detection, inertial sensor.**

I. INTRODUCTION

Demographic tendencies in today's societies lead to gentrification of the population both in developed and developing countries as well as in third world countries. According to UNFPA[1] although currently only Japan has an older population of more than 30 per cent, by 2050, 64 countries are expected to join it. These countries are supporting Ambient Assisted Living (AAL) research programs that address ICT technologies for the independent living of elders and disabled [2].

To achieve this goal, two main pillars are needed: Activity Recognition (AR) and Fall Detection (FD).

AR is a research area where the objective is to recognize human activities. The automatic and unobtrusive identification of users activities is one of the challenging goals of context-aware computing [3] and is expected to be a practical solution to monitor aged people. AR can be focused on basic activities (lying, sitting, standing up, etc.) or in complex ones (watching TV, cooking, having a shower, etc.). AR can be a good feedback tool to advise the user, relatives or doctors about the accomplishment of rehabilitation, preventive exercises or specific activity goals such as get some number of steps a day.

FD can be defined as an assistive technology whose main objective is to alert when a fall event occurs. In a real-life scenario, it has the potential to mitigate some of the adverse consequences of a fall. Specifically, FD can have a direct impact on the reduction in the fear of falling and the rapid provision of assistance after a fall. In fact, falls and fear of falling depend on each other: an individual who falls may subsequently develop fear of falling and, viceversa, the fear of falling may increase the risk of suffering from a fall [4]. Fear of falling has been shown to be associated with negative consequences such as avoidance of activities, less physical activity, falling, depression, decreased social contact and lower quality of life.

According to the World Health Organization [5] more than 28% of people aged 65 and over fall each year increasing to more than 32% for those over 70 years of age. If preventive measures are not taken in the immediate future, the number of injuries caused by falls is projected to be a 100% higher in 2030. In this context, assistive devices that could help to alleviate this major health problem are a social necessity. Indeed, fall detectors are being actively investigated.

The rest of the paper is organized as follows: Section 2 gives an overview of how AR systems are being evaluated and compared each other, Section 3 reviews the same for FD. The proposal of a high replicable protocol to create an open, flexible, growing and community maintained AR and FD dataset is presented in sections 4. Section 5 draws the conclusions.

II. EVALUATING AR

AR using wearable sensors [6] allows monitor user exercises and activities or detect abnormal behavior. AR also supports independent living, the main focus of some European Projects [7].

AR is mature enough from datasets point of view: Multiple datasets can be found in UCI Machine Learning Repository composed by different Activities of Daily Living (ADL). However, since each research group use different hardware or place their sensors in different positions, not all the datasets are feasible to every system and comparisons between different AR systems is not possible. To solve this problem two solutions have been proposed:

- Software-based competitions with a high number of sensors where the systems can choose the preferred sensors close to his hardware configuration. AR Challenge OPPORTUNITY [8], is the best example: an extremely sensor-rich and activity-rich common dataset against which all participants benchmark their proposed activity recognition software. The dataset includes 72 body-worn, ambient, and object sensors, a very high number of activity instances (more than 2500 instances of gestures) labeled at various levels of abstractions, executed by multiple persons.

I. Lacković and D. Vasić (eds.), *6th European Conference of the International Federation for Medical and Biological Engineering*,
IFMBE Proceedings 45, DOI: 10.1007/978-3-319-11128-5_163

- Live competitions or (hardware + software)-based competitions where the competitors are requested to install and run their systems during a set of benchmarks. This approach is more challenging because it is often the data-acquisition part (the sensors) that limits the systems' reliability and acceptability, and thus their real-life usability. EvAAL (standing for Evaluating AAL Systems through Competitive Benchmarking) AR track competition [9] is the only competition hold with these characteristics.

A. EvAAL Experience

The main objective of AR track is to evaluate AR systems intended to be used by the elderly in real life.

During the '12 and '13 competitions, the following seven activities were recognized: lying, sitting, standing, walking, bending, cycling on a stationary bike, and falling. Most of them were selected because they are common in daily life and thus recognizing them is the starting point for AR. Cycling was included because it is a recommended exercise for older people, while falling is a major health hazard for the elderly. All the activities were included in a scenario that lasted approximately 5 minutes. The scenario included actions of daily living (watching TV, working in the kitchen, bathroom activities, sleeping) and was repeated twice, with the better run by each competitor counting towards the final score. In order to get approximately the same ground truth for all the competitors, audio cues were played from a file to signal the actor as to which activity should be performed three seconds in advance, giving the actor the time to prepare for it. This ground truth was refined by an evaluator who followed the actor and used an Android phone with a custom application to mark the precise time-stamps of the activities. In the competition, there is no limitation to the number and type of devices comprising the competing ARS. The only constraint that ARS should satisfy is the compatibility with the physical limitations of the hosting living lab.

While the evaluation scenario was short and relatively simple, the impression of people involved was that it is a decent indicator or real-life performance. An elderly simulation kit helped emulate the movements of a >65 years old person. A longer and more complex evaluation (multiple days of real life) would be preferable, but too difficult and expensive to organize.

The competition was a good opportunity for discussion, resulting in valuable feedback to improve both the AR and FD systems and the competition.

B. AR Datasets information

- Type of activities reported: Simple activities (lying, sitting, standing, etc.) or complex activities (watching TV, working in the kitchen, bathroom activities, sleeping, etc.).
- Demographics information: number of participants and in some cases age, weight and height.
- Sensor information: sampling frequency and range, fixation site, number and type of sensors.
- Attribute information: Description of all the recorded attributes.

III. EVALUATING FD

Many different approaches have been explored to automatically detect a fall using inertial sensors [10]. The biggest problem of this research area is datasets. To the best knowledge of authors, there are only a few in the AmI repository [11,12] and in the EvAAL website [13] with simulated falls. Furthermore there are no public datasets with real falls a very important issue according Bagalà et al. work [14]: published algorithms report high sensitivity (SE) and high specificity (SP) being tested on simulated falls performed by healthy volunteers, but applying the same algorithms to a real fall database SP and SE average is considerably lower. For instance, the best one [15] provides 83% SE and 97% SP but the results are still different from those obtained by the authors on their simulated-falls database (100% SE and SP). Moreover, Kangas et al. [16] also found differences between simulated and real-world falls on beds in terms of low impact magnitude.

Without public real fall datasets it is difficult to evaluate and compare FD systems such as AR systems. Only projects working with "fallers" can compare different algorithms with his private databases.

According Schwickert et al. [17] only 6 of 96 studies from 1998 to 2012 were performed including real-world fall data. From these papers only one [18] reported more than 10 falls (n=20). So the private datasets reports less than 100 falls. Privacy issues and the analysis of FD from a biomedical point of view promote this lack of public real-world datasets.

A. FD Datasets information

- Type of falls reported: Forward falls, backward falls, mixed direction falls
- Context information: Location (indoor/outdoor), activity before the fall (standing, sitting, walking forward, walking backwards, sit-to stand, stand-to-sit, etc.), reported direction of fall (Forward, backward, sideward), Impact spot

(floor, against wall/locker before hitting the floor, bed/sofa, desk, etc.), mats thickness (if simulated) and soil type (if real or quasi-real).

- Demographics information: number of participants, age, weight, height, BMI, gait speed and in some cases balance [18].
- Sensor information: sampling frequency and range, fixation site, number and type of sensors.
- Attribute information: Description of all the recorded attributes.

IV. Proposal

The difficulties to access to a group of high risk of falling patients, to report in a medical way the results (patient agreement, patient privacy, trial protocol approved by Ethics Committee, written informed consent before the trial, etc.) and to obtain a high number of falls is a chimeric task but we can support an open dataset of AR and FD in order to obtain a high number of simulated kind of falls and multiple kind of simple and complex activities accepted from both communities.

AmI repository and EvAAL achieved the first step, with datasets combining AR and FD but reported in an AR manner, missing some important points from FD point of view (Demographics, context).

To avoid the appearance of multiple datasets with different sensor, activities and falls configurations we defined a protocol focused on create an open and growing AR-FD dataset with multiple activities (simple and complex) and simulated and real falls. The main idea is to allow to other people replicate the data acquisition system. Due to the fact that current mobile phones include tri-axial accelerometers, gyroscopes and magnetometers and are 'wearable and ubiquitous devices' one or more will be used to acquire the data. How to attach to the human body the mobile phone(s) is capital. To avoid only one configuration all the positions will be detailed in the dataset so some features will be present and other no (assuming multiple contributors this will not be an issue). Our first trial includes a mobile on the hip and another on the chest (sternum) using a elastic-band. To avoid different inertial sensors' frequency rate depending on the mobile device, an Android application (we are developing the iOS one) captures data annotating the position of the sensors at a fixed rate (discarding the readings out of this rate) and generates the files. Currently we work with 20 Hz to allow the longer range of Android mobile devices.

According to FD and AR Datasets, some information such as gait speed, activities, reported direction of fall or mats thickness are hard to report from people out of medical or research environments. To simplify the reporting task, a video recording, time-stamped and synchronized to the acquisition data mobile phones must be included to the dataset. To simplify the time-stamped and synchronization process, another mobile phone is used to record the video through another application prepared for this task.

Community can complete the missing information analysing the video and even discard some not suitable events (e.g. not realistic simulated falls).

To add instances to the dataset, contributors will upload the video and the generated files to a server. A visual description of the infrastructure is shown in Figure 1.

We hope to publish the initial instances in the dataset June 2014.

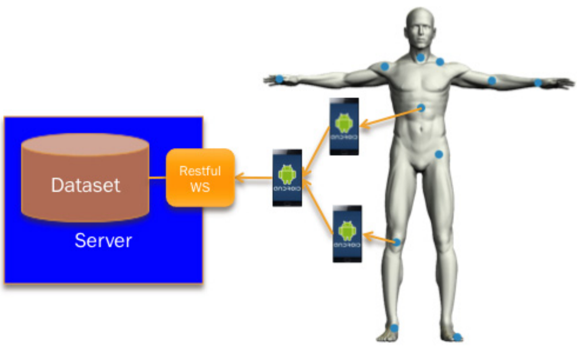


Fig. 1 Infrastructure used to collect the dataset

V. Conclusions

Fusing FD and AR, including rehabilitation exercises within the recognized activities, must support active independent living in aged people. To achieve this fusion some challenges must be complete:

1. Having public datasets in realistic environments with a rich configuration of sensors and a good description of users, sensors, activities and falls and environment
2. Evaluate multiple algorithms or systems through live competitions or new and independent datasets.

Although the EvAAL-AR competition is a good example of how to evaluate different hardware and software approaches with the same goal, participating in the competition requires an investment in time and money and only teams that were confident in the quality of their system participated, only four competitors each year.

Bagalà et al. approach is cheaper for "competitors" but it is time consuming for organizer: implementing all the algorithms (AR algorithms are usually more complex FD ones) and applying them to an independent and real fall database. Furthermore data collection must be prepared to consider diverse AR and FD sensor location configuration.

Our proposal, a high replicable protocol to create an open, flexible, growing and community maintained dataset of multiple activities and simulated or real falls could allow a software competition based on multiple hardware configurations using diverse positions of the acquisition tools, the mobile phones.

ACKNOWLEDGMENT

This research is partially supported by the Andalusian Regional Ministry of Economy projects P11-TIC-8052 (Simon), P11-TIC-7124 and by Spanish Ministry of Economy and Competitiveness through the project HERMES (TIN2013-46801-C4).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. United Nations Population Fund and Help Age International. (2012). Ageing in the twenty-first century: a celebration and a challenge.
2. AAL Joint Programme. Catalogue for Projects 2013. Retrieved May 4, 2014, from AAL-Europe: http://www.aal-europe.eu/wp-content/uploads/2013/09/AALCatalogue2013_Final.pdf
3. Brush, A.B., Krumm, J., Scott, J., Saponas, S.: Recognizing activities from mobile sensor data: Challenges and opportunities. In: Ubicomp' 11. (2011)
4. Legters, K. (2002). Fear of falling. Physical therapy, 82(3), 264-272.
5. World Health Organization. Ageing, & Life Course Unit. (2008). WHO global report on falls prevention in older age. World Health Organization.
6. Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. Communications Surveys & Tutorials, IEEE, 15(3), 1192-1209.
7. EU-funded research and innovation in fall prevention. Retrieved May 4, 2014 from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4080
8. Roggen, D.; Calatroni, A.; Rossi, M.; Holleczek, T.; Forster, K.; Troster, G.; Lukowicz, P.; Bannach, D.; Pirkl, G.; Ferscha, A.; Doppler, J.; Holzmann, C.; Kurz, M.; Holl, G.; Chavarriaga, R.; Sagha, H.; Bayati, H.; Creatura, M.; del R Millan, J., "Collecting complex activity datasets in highly rich networked sensor environments," Networked Sensing Systems (INSS), 2010 Seventh International Conference on , vol., no., pp.233,240, 15-18 June 2010
9. Álvarez-García, J. A., Barsocchi, P., Chessa, S., & Salvi, D. (2013). Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 EvAAL competition. Journal of Ambient Intelligence and Smart Environments, 5(1), 119-132.
10. Sannino, G., De Falco, I., & De Pietro, G. (2014, January). Effective supervised knowledge extraction for an mHealth system for fall detection. In XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013 (pp. 1378-1381). Springer International Publishing.
11. Kaluza, B., Kozina, S., & Lustrek, M. (2012, July). The Activity Recognition Repository: Towards Competitive Benchmarking in Ambient Intelligence. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.
12. Kozina, S., Gjoreski, H., Gams, M., & Lustrek, M. (2013). Three-layer activity recognition combining domain knowledge and meta-classification. Journal of Medical and Biological Engineering, 33(4), 406-414.
13. EvAAL website. Retrieved May 4, 2014 from: http://evaal.aaloa.org/
14. Bagalà F, Becker C, Cappello A, Chiari L, Aminian K, et al. (2012) Evaluation of Accelerometer-Based Fall Detection Algorithms on Real-World Falls. PLoS ONE 7(5): e37062. doi:10.1371/journal.pone.0037062
15. Bourke, A. K., Van de Ven, P., Gamble, M., O'Connor, R., Murphy, K., Bogan, E., ... & Nelson, J. (2010). Evaluation of waist-mounted tri-axial accelerometer based fall-detection algorithms during scripted and continuous unscripted activities. Journal of biomechanics, 43(15), 3051-3057.
16. Kangas, M., Vikman, I., Nyberg, L., Korpelainen, R., Lindblom, J., & Jämsä, T. (2012). Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects. Gait & posture, 35(3), 500-505.
17. Schwickert, L., Becker, C., Lindemann, U., Maréchal, C., Bourke, A., Chiari, L., ... & Klenk, J. (2013). Fall detection with body-worn sensors. Zeitschrift für Gerontologie und Geriatrie, 46(8), 706-719.
18. Klenk, J., Becker, C., Lieken, F., Nicolai, S., Maetzler, W., Alt, W., ... & Lindemann, U. (2011). Comparison of acceleration signals of simulated and real-world backward falls. Medical engineering & physics, 33(3), 368-373.

Author: Juan Antonio Alvarez-Garcia
Institute: Computer Languages and Systems Department, University of Seville
Street: Avda. Reina Mercedes s/n ETSI Informática
City: Seville
Country: Spain
Email: jaalvarez@us.es

PART IV

Final remarks

CHAPTER 13

CONCLUSIONS AND FUTURE WORK

I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right - Albert Einstein

13.1 Conclusions

This thesis focuses on the problem of recognizing activities and fall detection in mobile systems, which presents a major difficulties in terms of battery draining and accuracy. This obstacle has two bases: on one hand, amount of consumed energy, which can decrease the battery level; on the other hand, it is known that a fast response is desired, and the dependency on high performance features in mobile systems is not sustainable. Moreover, this work takes into account both energy consumption and performance in mobile systems, and provides a solution to recognize activities based on discretization processes.

An activity recognition algorithm and, on top of that, a mobile application has been developed in order to identify activities in real time in mobile systems whereby a discretization process have been used as the core of it. This application has been widely tested both in the labs and in some real scenarios like the EvAAL competition in order to obtain reliable results, and the mobile application has been published online through the Google Store, thereby making it available worldwide for people interested on activity recognition.

The innovation of the thesis is also presented in the use of discretization algorithms, specifically the Ameva discretization algorithm, which is applied in order to achieve a trade-off between accuracy and energy consumption. This innovation has proved itself to be a useful improvement for mobile systems in the evaluation of efficiency and performance of activity recognition, who in turn detects activities with an high accuracy.

Finally, as result of the applied researching, the developed system got two awards in international competitions as can be seen in A.3.

13.2 Future work

In terms of activity recognition, future work will focus on the next topics:

- Increasing the number of activities that can be recognized using solely the accelerometer sensors. The current activities that the system can recognized have a high dependency on the statistics defined in the process, so defining new features or removing existing ones could extend the set of activities. Also, this problem could be solved by including new sensors which provide more information to the system.

- Reducing the correlation between variables using the approach started in 7. To date, the current system does not have any preprocess where the statistics are removed due to the correlation existing between them. The feature reduction methodology published was only one basic approach and the next step is to design an automatic method of creation of feature discrimination rules and to define some improvements in this methodology to automatically know the dependence between features without setting manually a threshold value.

- Finding better locations where the mobile device should be placed. The system gets good results when the device is placed on the hip but it was not widely tested in some other natural places like the arm (using an sport band) or the hand when walking.

- Bringing the system to trending mobile devices like smartwatches. Currently, the application has been developed for smartphones because it is the most common mobile device but we consider other ones that are connected to it by Bluetooth or WiFi.

- Extending the mobile application developed as result of the applied researching. As an output of the system for fall detection, it can transmit an alarm signal complementing existing telecare services allowing a faster response in case of an medical emergency.

APPENDIX A

CURRICULUM

A.1 Published papers

During last years, a set of research papers has been published related to the Ameva discretization algorithm and how it has been used into the activity recognition area to improving it.

[**2009 - 2010**]**:** I participated as a researcher both in a national researching project called InCare and in a research project of excellence called CUBICO that improved services for dependant people. Some papers were published in conferences.

[**2011**]**:** During 2011 the research was focused on extending the Ameva discretization algorithm. The new extension allowed not only to discretize features, but also to filter them as well. As a result, the paper [6] was a result of this researching.

[**2012**]**:** In 2012, we focused our efforts in the use of the Ameva methodology in Activity Recognition Systems. We developed an Android application that allows to train a recognize activities using an smartphone. We published a research paper [87] in this area. Also, as a result of the position as an employee in the CICA, the paper [5] was published.

[**2013**]**:** This year we decided to continue with the two previous topics: extending Ameva and applying it in the Activity Recognition area. The papers [26] and [27] presented the results of researching on them. This new approach comes as a result of my stage at Laboratoire d'Analyse et d'Architecture des Systèmes of the CNRS with Dra. Louise Travé-Massuyès during 2012.

[**2014**]**:** Finally, during this year we published a well defined methodology that used the benefits of the Ameva methodology in a well defined Activity Recognition system. The paper [25] was the result of this researching.

[**2015**]**:** Once we have a successful activity recognition methodology, we started to improve it in terms of battery consumption and we were able to publish the paper [88]. Also, we decided to focus on wearable devices and as result on it, the paper [9] was published.

[**2016**]**:** This year, we focused our work in fall detection that allowed elderly people to get an accurate response from medical systems. The result was the paper [28].

A.1.1 JCR Indexed Journals

1. Title: **Mobile activity recognition and fall detection system for elderly people using Ameva algorithm**. Authors: **M.Á. Álvarez de la Concepción, L.M. Soria Morillo, J.A. Álvarez García, L. González Abril**.

   Published in: **Pervasive and Mobile Computing**, Elsevier, ISSN: 1574-1192, Date of Publication: January 2017, Volume: 34, On Pages: 3-13, DOI: `https://doi.org/10.1016/j.pmcj.2016.05.002`, **Q2 in two categories. JCR-2016 IF: 2.349**.

2. Title: **Low Energy Physical Activity Recognition System on Smartphones**. Authors: **L.M. Soria Morillo, L. González Abril, J.A. Ortega Ramírez, M.Á. Álvarez de la Concepción**.

Published in: **Journal of Sensors**, MDPI AG, ISSN: 1424-8220, Date of Publication: March 2015, Volume: 15, Issue: 3, On Pages: 5163-5196, DOI: `https://doi.org/10.3390/s150305163`, **Q1, Q2 and Q3 in one category each. JCR-2015 IF: 2.033**.

3. Title: **Discrete techniques applied to low-energy mobile human activity recognition. A new approach**. Authors: **M.Á. Álvarez de la Concepción, L.M. Soria Morillo, L. González Abril, J.A. Ortega Ramírez**.

   Published in: **Journal of Expert Systems with Applications**, Elsevier, ISSN: 0957-4174, Date of Publication: October 2014, Volume: 41, Issue: 14, On Pages: 6138-6146, DOI: `https://doi.org/10.1016/j.eswa.2014.04.018`, **Q1 in three categories. JCR-2014 IF: 2.240**.

A.1.2 Other Journals

4. Title: **An adaptive methodology to discretize and select features**. Authors: **M.Á. Álvarez de la Concepción, L. González Abril, L.M. Soria Morillo and J.A. Ortega Ramírez**.

   Published in: **Artificial Intelligence Research**, ISSN: 1927-6974, Date of Publication: February 2013, On Pages: 77-86, DOI: `https://doi.org/10.5430/air.v2n2p77`.

5. Title: **Extensiones para el ciclo de mejora continua en la enseñanza e investigación de Ingeniería Informática**. Authors: **M.Á. Álvarez de la Concepción, A. Jiménez Ramírez, M.M. Martínez Ballesteros, R. Martínez Gasca, L. Parody Núñez and L.M. Soria Morillo**.

   Published in: **Revista de Enseñanza Universitaria**, ISSN: 1131-5245, Date of Publication: December 2011, On Pages: 4-26.

A.1.3 International Conferences

6. Title: **Evaluating Wearable Activity Recognition and Fall Detection Systems**. Authors: **J.A. Álvarez García, L.M. Soria Morillo, M.Á. Álvarez de la Concepción, A. Fernández-Montes González and J.A. Ortega Ramírez**.

   Published in: **6th European Conference of the International Federation for Medical and Biological Engineering**, ISBN: 978-3-319-11127-8, Date of Publication: January 2015, On Pages: 653-656, DOI: `https://doi.org/10.1007/978-3-319-11128-5_163`.

7. Title: **Hi-Res activity recognition system based on EEG and WoT**. Authors: **L.M. Soria Morillo, M.Á. Álvarez de la Concepción, J.A. Álvarez García, J.A. Ortega Ramírez and R. Vergara**.

   Published in: **Ambient Intelligence for Telemedicine and Automotive**, ISBN: 978-84-697-0147-8, Date of publication: November 2013, On Pages: 7-11.

8. Title: **A new back-propagation algorithm with momentum coefficient for medical datasets**. Authors: **V. Montes Valencia, J.M. Chincho Cardosa, M.Á. Álvarez de la Concepción, L.M. Soria Morillo and J.A. Ortega Ramírez and R. Vergara**.

   Published in: **Ambient Intelligence for Telemedicine and Automotive**, ISBN: 978-84-697-0147-8, Date of publication: November 2013, On Pages: 19-21.

9. Title: **Activity Recognition System Using Non-intrusive Devices through a Complementary Technique Based on Discrete Methods**. Authors: **M.Á. Álvarez de la Concepción, L.M. Soria Morillo, L. González Abril and J.A. Ortega Ramírez**.

   Published in: **Evaluating AAL Systems Through Competitive Benchmarking**, ISSN: 1865-0929, Date of Publication: June 2013, On Pages: 36-47, DOI: `https://doi.org/10.1007/978-3-642-41043-7`.

10. Title: **Activity Recognition System Using AMEVA Method**. Authors: **L.M. Soria Morillo, L. González Abril, M.Á. Álvarez de la Concepción and J.A. Ortega Ramírez**.

    Published in: **Evaluating AAL Systems Through Competitive Benchmarking**, ISSN: 1865-0929, Date of Publication: January 2013, On Pages: 137-147, DOI: `https://doi.org/10.1007/978-3-642-37419-7`.

11. Title: **Aprendizaje dirigido por etapas para la consecución de aptitudes profesionales en asignaturas de Ingeniería del Software**. Authors: **M.Á. Álvarez de la Concepción, H. Sarmiento, L. González Abril and J.A. Ortega Ramírez**.

    Published in: **IV Congreso Internacional de Ambientes Virtuales de Aprendiza Adaptativos y Accesibles**, ISSN: 2323-0010, Date of Publication: December 2012, On Pages: 107-110, DOI: `https://doi.org/10.5220/0003992603010304`.

12. Title: **The CICA grid: a cloud computing infrastructure on demand with open source technologies**. Authors: **M.Á. Álvarez de la Concepción, A. Fernández-Montes González, J.A. Ortega Ramírez and L. González Abril**.

    Published in: **14th International Conference on Enterprise Information Systems**, ISBN: 978-989-8565-11-2, Date of Publication: January 2012, On Pages: 301-304, DOI: `https://doi.org/10.5220/0003992603010304`.

13. Title: **A qualitative methodology to reduce features in classification problems**. Authors: **M.Á. Álvarez de la Concepción, L. González Abril, J.A. Ortega Ramírez and L.M. Soria Morillo**.

    Published in: **25th International Workshop on Qualitative Reasoning**, Date of Publication: August 2011, On Pages: 1-4.

A.1.4 National Conferences

14. Title: **Generación de dataset flexible para el reconocimiento de actividades y caídas**. Authors: **J.A. Álvarez García, L.M. Soria Morillo and M.Á. Álvarez de la Concepción**.

    Published in: **XVI Jornadas de Arca. Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robótica e Inteligencia Ambiental**, Date of publication: June 2014, On Pages: 1-4.

15. Title: **A comparative about chronicles**. Authors: **M.Á. Álvarez de la Concepción, L.M. Soria Morillo, Luis González Abril and J.A. Ortega Ramírez**.

    Published in: **XV Jornadas de Arca. Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robótica e Inteligencia Ambiental**, ISBN: 978-84-616-7622-4, Date of publication: June 2013, On Pages: 89-94.

16. Title: **An extended chronicle discovery approach to find temporal patterns between sequences**. Authors: **M.Á. Álvarez de la Concepción, A. Subias, L. Travé-Massuyès, Luis González Abril and J.A. Ortega Ramírez**.

    Published in: **XIV Jornadas de Arca. Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robótica e Inteligencia Ambiental**, Date of publication: June 2012, On Pages: 51-54.

17. Title: **A quantitative methodology to identify related features in data sets**. Authors: **M.Á. Álvarez de la Concepción, Luis González Abril, J.A. Ortega Ramírez, L.M. Soria Morillo and F.J. Cuberos García-Baquero**.

    Published in: **XIII Jornadas de Arca. Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robótica e Inteligencia Ambiental**, ISBN: 978-84-615-5513-0. Date of Publication: June 2011, On Pages: 39-43.

18. Title: **Benchmarking on Ameva: a performance test on discretization Algorithm**. Authors: **M.Á. Álvarez de la Concepción, F.J. Cuberos García-Baquero, Luis González Abril and J.A. Ortega Ramírez**.

    Published in: **XII Jornadas de Arca. Eficiencia Energética y Sostenibilidad en Inteligencia Ambiental**, ISBN: 978-84-614-6457-9. Date of Publication: June 2010, On Pages: 101-106.

19. Title: **An Approach to the Implementation of Web TV Architecture with Interactive Services**. Authors: **A.M. Bellido Romero, M.Á. Álvarez de la Concepción, L.M. Soria Morillo, J.A. Ortega Ramírez and J. Torres Valderrama**.

    Published in: **XII Jornadas de Arca. Eficiencia Energética y Sostenibilidad en Inteligencia Ambiental**, ISBN: 978-84-614-6457-9. Date of Publication: June 2010, On Pages: 83-88.

20. Title: **Controlled monitoring in intelligent environments**. Authors: **M.Á. Álvarez de la Concepción, A.M. Bellido Romero, J.A. Álvarez García, J.A. Ortega Ramírez and F. Velasco Morente**.

    Published in: **XI Jornadas de Arca. Sistemas Cualitativos, Diagnosis, Robótica, Sistemas Domóticos y Computación Ubicua**, ISBN: 978-84-613-71-587. Date of Publication: June 2009, On Pages: 49-52.

21. Title: **Social networks applications: detecting school bullying**. Authors: **A.M. Bellido Romero, M.Á. Álvarez de la Concepción, J.A. Álvarez García, J.A. Ortega Ramírez, L. González Abril and J. Torres Valderrama**.

    Published in: **XI Jornadas de Arca. Sistemas Cualitativos, Diagnosis, Robótica, Sistemas Domóticos y Computación Ubicua**, ISBN: 978-84-613-71-587. Date of Publication: June 2009, On Pages: 45-48.

22. Title: **A location of robots proposal in collaborative environments**. Authors: **M.Á. Álvarez de la Concepción, A. Fernández-Montes González,**

J.A. Ortega Ramírez and F. Velasco Morente.

Published in: **X Jornadas de Arca. Sistemas Cualitativos y Diagnosis, Robótica, Sistemas Domóticos y Computación Ubicua**, ISBN: 978-84-89315-54-9. Date of Publication: June 2008, On Pages: 40-43.

A.2 Patents

23. Title: **Grid'5000 Toolbox. A simulating tool to apply energy saving policies in Grid Computing and Internet Datacenters**. Authors: **A. Fernández-Montes González, J.I. Sánchez Venzalá, J.A. Ortega Ramírez, L. González Abril, F. Velasco Morente, J.A. Álvarez García, M.Á. Álvarez de la Concepción, and L.M. Soria Morillo**. Reference: **Request: 2012-12-20, Number: SE-1244-12**.

24. Title: **Arquitectura para la gestión de tareas de usuarios en un cluster a través de la web. SCI (SIMPLE CLUSTER INTERFACE)**. Authors: **J.A. Ortega Ramírez, L. González Abril, J.A. Álvarez García, M.Á. Álvarez de la Concepción, D. Fuentes Brenes, A. Fernández-Montes González, J. Cantón Ferrero, A. Silva, D. Bosque, F. Velasco Morente, J. Torres Valderrama, M.J. Escalona Cuaresma and L.M. Soria Morillo**. Reference: **Request: 2010-10-22, Number: P1001371**.

25. Title: **Dilos: Dispositivo de Localización y Seguimiento Energéticamente Eficiente**. Authors: **L.M. Soria Morillo, M.A. Álvarez de la Concepción, J.A. Álvarez García, A.M. Bellido Romero, D. Fuentes Brenes, A. Fernández-Montes González, L. González Abril, J.A. Ortega Ramírez, F. Velasco Morente, J. Torres Valderrama and J.I. Sánchez Venzalá**. Reference: **Request: 2010-03-22, Number: P201000969**, International Application Number: PCT/ES2011/000237, Publication Number: WO/2012/010727.

A.3 Awards

- **Third prize in Activity Recognition track** in Evaluating AAL Systems through Competitive Benchmarking in July 2013.
- **First prize in Activity Recognition track** in Evaluating AAL Systems through Competitive Benchmarking in July 2012.

A.4 R&D projects

This thesis dissertation has been developed within the framework of the following research projects:

- Title: **La Formación a Través de Dispositivos Móviles. Diseño y Evaluación de Contenidos y Actividades Formativas a Través de M-Learning**.

  Main researchers: **Carlos Marcelo García**. Granting Entity: **Junta de Andalucía. Consejería de Innovación, Ciencia y Empresa**. Period: **2013-2016**. Reference: **P11-TIC-7124**.

- Title: **Healthy and Efficient Routes in Massive Open-Data Based Smart Cities-Citizen**.

  Main researchers: **Juan Antonio Ortega Ramírez and Juan Antonio Álvarez García**. Granting Entity: **Gobierno de España. Ministerio de Economía y Competitividad**. Period: **2014-2017**. Reference: **TIN2013-46801-C4-1-R**.

- Title: **Simon. Saving Energy by Intelligent Monitoring**.

  Main researcher: **Juan Antonio Ortega Ramírez**. Granting Entity: **Junta de Andalucía. Consejería de Economía, Innovación y Ciencia**. Period: **2013-2017**. Reference: **P11-TIC-8052**.

- Title: **Arquitectura para la eficiencia energética y sostenibilidad en entornos residenciales**.

  Main researcher: **Juan Antonio Ortega Ramírez**. Granting Entity: **Gobierno de España. Ministerio de Ciencia e Innovación**. Period: **2010-2013**. Reference: **TIN2009-14378-C02-01**.

A.5 Scientific Outreach Events

- Title: **Speaker at La Noche de los Investigadores**. Year: **2014**.

- Title: **Member of the Editorial Board of the Journal of Computer Engineering (COES&RJ-JCE)**. Year: **2013**.

- Title: **Member of the Track Chair of the Cognitive Informatics and Computing track of 1st International Conference on Cognitive and Sensor Networks**. Year: **2013**.

- Title: **Member of the Technical Program Committee of the Machine Learning for Signal Processing track of 2nd International Conference on Digital Signal Processing**. Year: **2013**.

- Title: **Member of the Technical Program Committee of the Cognitive Informatics and Computing track of 1st International Conference on Cognitive and Sensor Networks**. Year: **2013**.

- Title: **Member of the Technical Program Committee of the Software Engineering and Applications track of 2nd International Conference on Computer Science and Engineering**. Year: **2013**.

- Title: **Member of the Publicity Chair of the 2nd International Conference on Digital Signal Processing**. Year: **2013**.

- Title: **Member of the Track Chair of the Signal, Image and Speech Processing track of 5th International Conference on Communications, Signals and Coding**. Year: **2012**.
- Title: **Member of the Technical Program Committee of the Signal, Image and Speech Processing track of Mosharaka International Conference on Communications, Signals and Coding**. Year: **2012**.
- Title: **Member of the Technical Program Committee of the Data Mining and Knowledge Discovery track of 2nd International Conference on Computing and Artificial Intelligence**. Year: **2012**.
- Title: **Member of the Academic Coordinator Team of the Congreso Internacional de Ambientes Virtuales de Aprendizaje Adaptativos y Accesibles (CAVA)**. Year: **2012**.
- Title: **Member of the Editorial Review Board of the journal Artificial Intelligence Research (AIR)**. Year: **2012**.
- Title: **Member of the Working Group of ADMS.F/OSS**. Year: **2012**.
- Title: **Member of the Organizing Committee of Café con Ciencia**. Year: **2011**.
- Title: **Speaker at seminar in Laboratoire d'Analyse et d'Architecture des Systèmes**. Year: **2011**.

A.6 Supervision of Degree Projects

- Title: **Tienda Virtual**.

  Author: **Hatim Khrichfa**. Year: **2012-2013**.
- Title: **REDSOVI**.

  Author: **Felipe García Ojeda**. Year: **2012-2013**.

- Title: **Brain Application**.

  Author: **Esteban Álvarez Catalán**. Year: **2013-2014**.

- Title: **Sistema multiplataforma de visionado de contenido audiovisual**.

  Author: **Jesús Manuel Vargas Sosa**. Year: **2013-2014**.

BIBLIOGRAPHY

[1] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy. Adaptive mobile activity recognition system with evolving data streams. *Neurocomputing*, 150:304–317, 2015.

[2] N. Agarwal and M. Sebastian. Wireless infrastructure setup strategies for healthcare. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, page 66. ACM, 2014.

[3] W. H. O. Ageing and L. C. Unit. *WHO global report on falls prevention in older age*. World Health Organization, 2008.

[4] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

[5] M. Alvarez, A. Fernández-Montes, J. Ortega, and L. G. Abril. The cica grid-a cloud computing infrastructure on demand with open source technologies. In *ICEIS (2)*, pages 301–304, 2012.

[6] M. Alvarez, L. Gonzalez-Abril, J. Ortega, and L. Soria. A qualitative methodology to reduce features in classification problems.

[7] J. Alvarez-Garcia, J. Ortega, L. Gonzalez-Abril, and F. Velasco. Trip destination prediction based on past gps log using a hidden markov model. *Expert Systems with Applications*, 37(12):8166–8171, 2010.

[8] J. A. Álvarez García, P. Barsocchi, S. Chessa, and D. Salvi. Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 evaal competition. *Journal of Ambient Intelligence and Smart Environments*, 5(1):119–132, 2013.

[9] J. A. Álvarez García, L. M. S. Morillo, M. A. A. de La Concepción, A. Fernández-Montes, and J. A. O. Ramírez. Evaluating wearable activity recognition and fall detection systems. In *6th European Conference of the International Federation for Medical and Biological Engineering*, pages 653–656. Springer, 2015.

[10] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Ambient Assisted Living and Home Care*, pages 216–223, 2012.

[11] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

[12] S. A. Antos, M. V. Albert, and K. P. Kording. Hand, belt, pocket or bag: Practical activity tracking with mobile phones. *Journal of neuroscience methods*, 231:22–30, 2014.

[13] S. Azzi, C. Dallaire, A. Bouzouane, B. Bouchard, and S. Giroux. Human activity recognition in big data smart home context. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 1–8. IEEE, 2014.

[14] F. Bagalà, C. Becker, A. Cappello, L. Chiari, K. Aminian, J. M. Hausdorff, W. Zijlstra, and J. Klenk. Evaluation of accelerometer-based fall detection algorithms on real-world falls. *PloS one*, 7(5):e37062, 2012.

[15] L. Bao. *Physical activity recognition from acceleration data under semi-naturalistic conditions*. PhD thesis, Massachusetts Institute of Technology, 2003.

[16] A. Bhattacharya, E. McCutcheon, E. Shvartz, and J. Greenleaf. Body acceleration distribution and O2 uptake in humans during running and jumping. *Journal of Applied Physiology*, 49(5):881–887, 1980.

[17] N. Bicocchi, M. Mamei, and F. Zambonelli. Detecting activities from body-worn accelerometers via instance-based algorithms. *Pervasive and Mobile Computing*, 6(4):482–495, 2010.

[18] T. Brezmes, J. Gorricho, and J. Cotrina. Activity recognition from accelerometer data on a mobile phone. *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, pages 796–799, 2009.

[19] M. Carman, N. Tam, and M. Woodward. Why fitbit has succeeded when other pedometers have failed. 2013.

[20] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[21] T. Choudhury, S. Consolvo, B. Harrison, J. Hightower, A. LaMarca, L. LeGrand, A. Rahimi, A. Rea, G. Bordello, B. Hemingway, et al. The mobile sensing platform: an embedded activity recognition system. *Pervasive Computing*, 7(2):32–41, 2008.

[22] D. Chu, N. D. Lane, T. T.-T. Lai, C. Pang, X. Meng, Q. Guo, F. Li, and F. Zhao. Balancing energy, latency and accuracy for mobile sensor data classification. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pages 54–67. ACM, 2011.

[23] N. Costa, P. Domingues, F. Fdez-Riverola, and A. Pereira. A mobile virtual butler to bridge the gap between users and ambient assisted living: A smart home case study. *Sensors*, 14(8):14302–14329, 2014.

[24] S. Daskalaki, I. Kopanas, and N. Avouris. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence*, 20(5):381–417, 2006.

[25] M. A. De La Concepción, L. S. Morillo, L. Gonzalez-Abril, and J. O. Ramírez. Discrete techniques applied to low-energy mobile human activity recognition. a new approach. *Expert Systems with Applications*, 41(14):6138–6146, 2014.

[26] M. A. A. de la Concepción, L. G. Abril, L. M. S. Morillo, and J. A. O. Ramírez. An adaptive methodology to discretize and select features. *Artificial Intelligence Research*, 2(2):77, 2013.

[27] M. A. A. de la Concepción, L. M. S. Morillo, L. G. Abril, and J. A. O. Ramírez. Activity recognition system using non-intrusive devices through a complementary technique based on discrete methods. In *International Competition on Evaluating AAL Systems through Competitive Benchmarking*, pages 36–47. Springer, 2013.

[28] M. A. A. de la Concepción, L. M. S. Morillo, J. A. A. García, and L. González-Abril. Mobile activity recognition and fall detection system for elderly people using ameva algorithm. *Pervasive and Mobile Computing*, 34:3–13, 2017.

[29] P. Dell'Acqua, L. V. Klompstra, T. Jaarsma, and A. Samini. An assistive tool for monitoring physical activities in older adults. In *Serious Games and Applications for Health (SeGAH), 2013 IEEE 2nd International Conference on*, pages 1–6. IEEE, 2013.

[30] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook. Simple and complex activity recognition through smart phones. In *Intelligent Environments (IE), 2012 8th International Conference on*, pages 214–221. IEEE, 2012.

[31] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 838–845, 2005.

[32] H. Ellekjær, J. Holmen, E. Ellekjær, and L. Vatten. Physical activity and stroke mortality in women: ten-year follow-up of the nord-trøndelag health survey, 1984-1986. *Stroke*, 31(1):14–18, 2000.

[33] K. Ellis, J. Kerr, S. Godbole, and G. Lanckriet. Multi-sensor physical activity recognition in free-living. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 431–440. ACM, 2014.

[34] R. Entezari-Maleki, S. M. Iranmanesh, and B. Minaei-Bidgoli. An experimental investigation of the effect of discrete attributes on the precision of classification methods. In *Information and Communication Technologies, 2009. ICICT'09. International Conference on*, pages 215–220. IEEE, 2009.

[35] M. Ermes, J. PÄrkkÄ, J. MÄntyjÄrvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):20–26, Jan 2008.

[36] Z. Falomir, L. Gonzalez-Abril, L. Museros, and J. A. Ortega. Measures of similarity between objects based on qualitative shape descriptions. *Spatial Cognition & Computation*, 13(3):181–218, 2013.

[37] Z. Falomir, L. Museros, L. Gonzalez-Abril, M. T. Escrig, and J. A. Ortega. A model for the qualitative description of images based on visual and spatial features. *Computer Vision and Image Understanding*, 116(6):698–714, 2012.

[38] D. Ferreira, A. K. Dey, and V. Kostakos. Understanding human-smartphone concerns: a study of battery life. In *International Conference on Pervasive Computing*, pages 19–33. Springer, 2011.

[39] J. Fogarty, C. Au, and S. E. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 91–100. ACM, 2006.

[40] Forrester. Customer technology survey. Technical report, North American Technologies, 2011.

[41] Forrester. North american technographics consumer technology online survey, q1 2011 (us). Technical report, Cambridge, 2011.

[42] D. Fuentes, L. Gonzalez-Abril, C. Angulo, and J. Ortega. Online motion recognition using an accelerometer in a mobile device. *Expert Systems with Applications*, 39(3):2461–2465, 2012.

[43] U. N. P. Fund and H. A. International. Ageing in the twenty-first century: a celebration and a challenge, 2012.

[44] H. Gjoreski. *Context-based Reasoning in Ambient Intelligence.* PhD thesis, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia, 2015.

[45] H. Gjoreski, S. Kozina, M. Gams, M. Lustrek, J. A. Álvarez García, J.-H. Hong, J. Ramos, A. K. Dey, M. Bocca, and N. Patwari. Competitive live evaluations of activity-recognition systems. *Pervasive Computing, IEEE*, 14(1):70–77, 2015.

[46] L. González, C. Angulo, F. Velasco, and A. Catala. Dual unification of bi-class support vector machine formulations. *Pattern recognition*, 39(7):1325–1332, 2006.

[47] L. Gonzalez-Abril, F. J. Cuberos, F. Velasco, and J. A. Ortega. Ameva: An autonomous discretization algorithm. *Expert Systems with Applications*, 36(3):5327–5332, 2009.

[48] L. Gonzalez-Abril, F. Velasco, J. Ortega, and F. Cuberos. A new approach to qualitative learning in time series. *Expert Systems with Applications*, 36(6):9924–9927, 2009.

[49] M. Govindarajan. Text mining technique for data mining application. In *Proceedings of world academy of science, engineering and technology*, volume 26, pages 544–549, 2007.

[50] F. Guo, Y. Li, M. S. Kankanhalli, and M. S. Brown. An evaluation of wearable activity monitoring devices. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, pages 31–34. ACM, 2013.

[51] M. A. Habib, M. S. Mohktar, S. B. Kamaruzzaman, K. S. Lim, T. M. Pin, and F. Ibrahim. Smartphone-based solutions for fall detection and prevention: challenges and open issues. *Sensors*, 14(4):7181–7208, 2014.

[52] C. W. Han, S. J. Kang, and N. S. Kim. Implementation of hmm-based human activity recognition using single triaxial accelerometer. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(7):1379–1383, 2010.

[53] Z. He and L. Jin. Activity recognition from acceleration data based on discrete cosine transform and SVM. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 5041–5044, 2009.

[54] Y. Hong, I. Kim, S. Ahn, and H. Kim. Activity recognition using wearable sensors for elder care. In *Second International Conference on Future Generation Communication and Networking*, volume 2, pages 302–305, 2008.

[55] Y.-J. Hong, I.-J. Kim, S. C. Ahn, and H.-G. Kim. Mobile health monitoring system based on activity recognition using accelerometer. *Simulation Modelling Practice and Theory*, 18(4):446–455, 2010.

[56] F. B. Hu, M. J. Stampfer, G. A. Colditz, A. Ascherio, K. M. Rexrode, W. C. Willett, and J. E. Manson. Physical activity and risk of stroke in women. *Jama*, 283(22):2961–2967, 2000.

[57] L. Hu, Y. Chen, S. Wang, and Z. Chen. b-coelm: A fast, lightweight and accurate activity recognition model for mini-wearable devices. *Pervasive and Mobile Computing*, 15:200–214, 2014.

[58] J. Huang and C. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[59] G. A. Index. Insight report, 2014.

[60] P.-S. Jeong and Y.-H. Cho. Fall detection system using smartphone for mobile healthcare. *Journal of the Korea society of IT services*, 12(4):435–447, 2013.

[61] R. Jia and B. Liu. Human daily activity recognition by fusing accelerometer and multi-lead ecg data. In *Signal processing, communication and computing (ICSPCC), 2013 IEEE international conference on*, pages 1–4. IEEE, 2013.

[62] M. Kanis, S. Robben, J. Hagen, A. Bimmerman, N. Wagelaar, and B. Krose. Sensor monitoring in the home: giving voice to elderly people. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pages 97–100. IEEE, 2013.

[63] T. P. Kao, C. W. Lin, and J. S. Wang. Development of a portable activity detector for daily activity recognition. In *2009 IEEE International Symposium on Industrial Electronics*, pages 115–120, July 2009.

[64] A. M. Khan, Y.-K. Lee, S. Lee, and T.-S. Kim. Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly. *Medical & biological engineering & computing*, 48(12):1271–1279, 2010.

[65] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE transactions on information technology in biomedicine*, 14(5):1166–1172, 2010.

[66] C. R. Kothari. *Research methodology: Methods and techniques.* New Age International, 2004.

[67] S. Kozina, H. Gjoreski, M. Gams, and M. Luštrek. Efficient activity recognition and fall detection using accelerometers. In *Evaluating AAL Systems Through Competitive Benchmarking*, pages 13–23. Springer, 2013.

[68] A. Krause, M. Ihmig, E. Rankin, D. Leong, S. Gupta, D. Siewiorek, A. Smailagic, M. Deisher, and U. Sengupta. Trading off prediction accuracy and power consumption for context-aware wearable computing. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 20–26. IEEE, 2005.

[69] L. Kurgan and K. Cios. CAIM discretization algorithm. *IEEE transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.

[70] J. Kwapisz, G. Weiss, and S. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011.

[71] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, 15(3):1192–1209, 2013.

[72] I.-M. Lee, K. M. Rexrode, N. R. Cook, J. E. Manson, and J. E. Buring. Physical activity and coronary heart disease in women: Is no pain, no gain passé? *Jama*, 285(11):1447–1454, 2001.

[73] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. What is happening now? Detection of activities of daily living from simple visual features. *Personal and Ubiquitous Computing*, 14(8):749–766, 2010.

[74] M. Li, V. Rozgic, G. Thatte, S. Lee, B. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan. Multimodal physical activity recognition by fusing temporal and cepstral information. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 18(4):369–380, 2010.

[75] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou. Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on*, pages 138–143. IEEE, 2009.

[76] G. Liang, J. Cao, and W. Zhu. Circlesense: A pervasive computing system for recognizing social activities. In *Pervasive computing and communications (percom), 2013 ieee international conference on*, pages 201–206. IEEE, 2013.

[77] Y. Liang, X. Zhou, Z. Yu, and B. Guo. Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare. *Mobile Networks and Applications*, 19(3):303–317, 2014.

[78] B. Longstaff, S. Reddy, and D. Estrin. Improving activity classification for health applications on mobile devices using active and semi-supervised learning. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on-NO PERMISSIONS*, pages 1–7. IEEE, 2010.

[79] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM conference on embedded networked sensor systems*, pages 71–84. ACM, 2010.

[80] S. Maloney and I. Boci. Survey: techniques for efficient energy consumption in mobile architectures. *Power (mW)*, 16(9.56):7–35, 2012.

[81] J. E. Manson, P. Greenland, A. Z. LaCroix, M. L. Stefanick, C. P. Mouton, A. Oberman, M. G. Perri, D. S. Sheps, M. B. Pettinger, and D. S. Siscovick.

Walking compared with vigorous exercise for the prevention of cardiovascular events in women. *New England Journal of Medicine*, 347(10):716–725, 2002.

[82] A. Martínez, M. Llorente, and J. P. Lázaro. Ciami living lab: an economically sustainable technological tool for open innovation. *I ENoLL Living Lab Summer School*, 2010.

[83] M. Mathie, A. Coster, N. Lovell, B. Celler, S. Lord, and A. Tiedemann. A pilot study of long-term monitoring of human movements in the home using accelerometry. *Journal of Telemedicine and Telecare*, 10(3):144–151, 2004.

[84] U. Maurer, A. Rowe, A. Smailagic, and D. Siewiorek. *Location and Activity Recognition Using eWatch: A Wearable Sensor Platform*, pages 86–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[85] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350. ACM, 2008.

[86] L. M. S. Morillo. *Middleware para el desarrollo de aplicaciones ubicuas en dispositivos móviles.* PhD thesis, Universidad de Sevilla, 2011.

[87] L. M. S. Morillo, L. González-Abril, M. A. A. de la Concepción, and J. A. O. Ramírez. Activity recognition system using ameva method. In *International Competition on Evaluating AAL Systems through Competitive Benchmarking*, pages 137–147. Springer, 2012.

[88] L. M. S. Morillo, L. Gonzalez-Abril, J. A. O. Ramirez, and M. A. A. de la Concepcion. Low energy physical activity recognition system on smartphones. *Sensors*, 15(3):5163–5196, 2015.

[89] L. S. Morillo, J. O. Ramirez, J. A. Garcia, and L. Gonzalez-Abril. Outdoor exit detection using combined techniques to increase gps efficiency. *Expert Systems with Applications*, 39(15):12260–12267, 2012.

[90] M. Musolesi, M. Piraccini, K. Fodor, A. Corradi, and A. T. Campbell. Supporting energy-efficient uploading strategies for continuous sensing applications on mobile phones. In *International Conference on Pervasive Computing*, pages 355–372. Springer, 2010.

[91] F. J. Ordónez, P. de Toledo, and A. Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, 13(5):5460–5477, 2013.

[92] S. Paiyarom, P. Tungamchit, R. Keinprasit, and P. Kayasith. Activity monitoring system using dynamic time warping for the elderly and disabled people. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–4. IEEE, 2009.

[93] R. Paoli, F. Fernández-Luque, and J. Zapata. A system for ubiquitous fall monitoring at home via a wireless sensor network and a wearable mote. *Expert Systems with Applications*, 39(5):5566–5575, 2011.

[94] T. Pawar, S. Chaudhuri, and S. P. Duttagupta. Body movement activity recognition for ambulatory cardiac monitoring. *IEEE transactions on biomedical engineering*, 54(5):874–882, 2007.

[95] G. Plasqui, A. Bonomi, and K. Westerterp. Daily physical activity assessment with accelerometers: new insights and validation studies. *Obesity Reviews*, 14(6):451–462, 2013.

[96] S. J. Preece, J. Y. Goulermas, L. P. Kenney, D. Howard, K. Meijer, and R. Crompton. Activity identification using body-mounted sensors - a review of classification techniques. *Physiological measurement*, 30(4):R1, 2009.

[97] R. Rana, M. Hume, J. Reilly, R. Jurdak, and J. Soar. Transforming knowledge capture in healthcare: Opportunistic and context-aware affect sensing on smartphones. *arXiv preprint arXiv:1502.02796*, 2015.

[98] T. Rault, A. Bouabdallah, Y. Challal, and F. Marin. Context-aware energy-efficient wireless sensor architecture for body activity recognition. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 203–206. IEEE, 2014.

[99] N. Ravi, N. Dandekar, P. Mysore, and M. Littman. Activity recognition from accelerometer data. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 1541, 2005.

[100] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, 6(2):13, 2010.

[101] A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 108–109. IEEE, 2012.

[102] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. 3d head tracking for fall detection using a single calibrated camera. *Image and Vision Computing*, 31(3):246–254, 2013.

[103] J. Sharkey. Coding for life-battery life, that is. In *Google IO Developer Conference*, volume 2009, 2009.

[104] C. Shen, S. Chakraborty, K. R. Raghavan, H. Choi, and M. B. Srivastava. Exploiting processor heterogeneity for energy efficient context inference on mobile phones. In *Proceedings of the Workshop on Power-Aware Computing and Systems*, page 9. ACM, 2013.

[105] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176, 2014.

[106] M. Shoaib, H. Scholten, and P. J. Havinga. Towards physical activity recognition using smartphone sensors. In *Ubiquitous Intelligence and Computing,*

2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), pages 80–87. IEEE, 2013.

[107] A. Sixsmith and N. Johnson. A smart sensor to detect the falls of the elderly. *Pervasive Computing, IEEE*, 3(2):42–47, 2004.

[108] L. Soria Morillo, J. Ortega Ramirez, and L. Gonzalez-Abril. *Aplicaciones contextuales en dispositivos móviles: Arquitectura para la mejora de la eficiencia energética*. EAE. Spanish Academic Editorial, 2012.

[109] V. Srinivasan and T. Phan. An accurate two-tier classifier for efficient duty-cycling of smartphone activity recognition systems. In *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*, page 11. ACM, 2012.

[110] G. Srivastava, J. Park, A. C. Kak, B. Tamersoy, and J. Aggarwal. Multi-camera human action recognition. In *Computer Vision*, pages 501–511. Springer, 2014.

[111] M. Stager, P. Lukowicz, and G. Troster. Implementation and evaluation of a low-power sound-based user activity recognition system. In *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, volume 1, pages 138–141. IEEE, 2004.

[112] V. Taipale. Global trends, policies and gerontechnology. *Gerontechnology*, 12(4):187–193, 2014.

[113] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *Wearable Computers, 2007 11th IEEE International Symposium on*, pages 37–40. IEEE, 2007.

[114] J. E. van Bronswijk, H. Bouma, J. L. Fozard, W. D. Kearns, G. C. Davison, and P.-C. Tuan. Defining gerontechnology for r&d purposes. *Gerontechnology*, 8(1):3–10, 2009.

[115] K. Van Laerhoven. Iswc 2010: The latest in wearable computing research. *IEEE Pervasive Computing*, 10(1):8–10, 2011.

[116] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[117] Y. Wang, B. Krishnamachari, and M. Annavaram. Semi-markov state estimation and policy optimization for energy efficient mobile sensing. In *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012 9th Annual IEEE Communications Society Conference on*, pages 533–541. IEEE, 2012.

[118] Y. Wang, B. Krishnamachari, Q. Zhao, and M. Annavaram. The tradeoff between energy efficiency and user state estimation accuracy in mobile sensing. In *International Conference on Mobile Computing, Applications, and Services*, pages 42–58. Springer, 2009.

[119] Y. Wang, J. Lin, M. Annavaram, Q. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh. A framework of energy-efficient mobile sensing for automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 179–192, 2009.

[120] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1553–1567, 2006.

[121] M. Weiser. The computer for the 21st century: specialized elements of hardware and software, connected by wires, radio waves and infrared, will be so ubiquitous that no one will notice their presence. In *Readings in Human-Computer Interaction*, pages 933–940. Elsevier, 1995.

[122] G. M. Weiss and J. W. Lockhart. The impact of personalization on smartphone-based activity recognition. In *AAAI Workshop on Activity Context Representation: Techniques and Languages*, 2012.

[123] S. Weng, L. Xiang, W. Tang, H. Yang, L. Zheng, H. Lu, and H. Zheng. A low power and high accuracy mems sensor based activity recognition algorithm. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 33–38. IEEE, 2014.

[124] C. Wojek, K. Nickel, and R. Stiefelhagen. Activity recognition and room-level tracking in an office environment. In *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 25–30. IEEE, 2006.

[125] Y.-h. Wu, J. Wrobel, M. Cornuet, H. Kerhervé, S. Damnée, and A.-S. Rigaud. Acceptance of an assistive robot in older adults: a mixed-method study of human-robot interaction over a 1-month period in the living lab setting. *Clinical interventions in aging*, 9, 2014.

[126] Y.-H. Wu, J. Wrobel, V. Cristancho-Lacroix, L. Kamali, M. Chetouani, D. Duhaut, B. Le Pévédic, C. Jost, V. Dupourque, M. Ghrissi, et al. Designing an assistive robot for older adults: The robadom project. *IRBM*, 34(2):119–123, 2013.

[127] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 17–24. Ieee, 2012.

[128] O. Yurur, C. Liu, and W. Moreno. A survey of context-aware middleware designs for human activity recognition. *Communications Magazine, IEEE*, 52(6):24–31, 2014.

[129] O. Yurur, C. H. Liu, X. Liu, and W. Moreno. Adaptive sampling and duty cycling for smartphone accelerometer. In *Mobile Ad-Hoc and Sensor Systems (MASS), 2013 IEEE 10th International Conference on*, pages 511–518. IEEE, 2013.

[130] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster. Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In *Wireless Sensor Networks*, volume 4913, pages 17–33. 2008.

[131] M. Zhang and A. A. Sawchuk. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *ACM International Conference on Ubiquitous Computing (Ubicomp) Workshop on Situation, Activity and Goal Awareness (SAGAware)*, Pittsburgh, Pennsylvania, USA, September 2012.

[132] Q. Zhang, Y. Su, and P. Yu. Assisting an elderly with early dementia using wireless sensors data in smarter safer home. In *Service Science and Knowledge Innovation*, pages 398–404. Springer, 2014.

[133] Z. Zhao, Z. Chen, Y. Chen, S. Wang, and H. Wang. A class incremental extreme learning machine for activity recognition. *Cognitive Computation*, 6(3):423–431, 2014.

[134] A. Zinnen, U. Blanke, and B. Schiele. An analysis of sensor-oriented vs. model-based activity recognition. In *Wearable Computers, 2009. ISWC'09. International Symposium on*, pages 93–100. IEEE, 2009.