Expert Systems With Applications 115 (2019) 1–15

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Subjective data arrangement using clustering techniques for training expert systems

Isaac Martín de Diego[a,*], Oscar S. Siordia[b], Alberto Fernández-Isabel[a], Cristina Conde[a], Enrique Cabello[a]

[a] Face Recognition and Artificial Vision Group, Data Science Laboratory, Rey Juan Carlos University, c/ Tulipán, s/n, 28933, Móstoles, Spain

[b] Centro de Investigación en Ciencias de Información Geoespacial (CentroGeo), Laboratorio Nacional de Inteligencia (GeoInt), Parque Científico Tecnológico Yucatán (PCTY), México

ARTICLE INFO

Article history:
Received 27 April 2018
Revised 2 July 2018
Accepted 29 July 2018
Available online 31 July 2018

Keywords:
Subjective sequential data
Subjective data arrangement
Combination of similarities
Driving risk assessment
Driving risk prediction

ABSTRACT

The evaluation of subjective data is a very demanding task. The classification of the information gathered from human evaluators and the possible high noise levels introduced are ones of the most difficult issues to deal with. This situation leads to adopt individuals who can be considered as experts in the specific application domain. Thus, the development of Expert Systems (ES) that consider the opinion of these individuals have been appeared to mitigate the problem. In this work an original methodology for the selection of subjective sequential data for the training of ES is presented. The system is based on the arrangement of knowledge acquired from a group of human experts. An original similarity measure between the subjective evaluations is proposed. Homogeneous groups of experts are produced using this similarity through a clustering algorithm. The methodology was applied to a practical case of the Intelligent Transportation Systems (ITS) domain for the training of ES for driving risk prediction. The results confirm the relevance of selecting homogeneous information (grouping similar opinions) when generating a ground truth (a reliable signal) for the training of ES. Further, the results show the need of considering subjective sequential data when working with phenomena where a set of rules could not be easily learned from human experts, such as risk assessment.

1. Introduction

The practice of Knowledge Engineering (Van Do, Le Thi, & Nguyen, 2018) has become a very useful approach to solve complex problems that require a high level of human expertise. This discipline involves integrating knowledge into computer systems which emulates the decision-making ability of a human expert in a specific domain. The systems in charge of achieving these tasks are the Expert Systems (ES) (Agarwal & Goel, 2014).

The building, maintaining and development of ES (Djamal et al., 2017) are mainly based on the interaction between the knowledge engineer and the domain expert (Yau & Sattar, 1994). The development of a reliable ES requires a deep understanding and a good representation of the knowledge of the domain expert.

In most of the cases, the knowledge representation is based on a set of rules (a production system) that ease the explanation of the decision-making made by the inference engine (Wick & Slagle, 1989). These rules are build from the knowledge acquired from human experts with the application of Machine Learning techniques (such as Neural Networks (Lin & Zhang, 2012), Deep Learning (Wei, He, Chen, Zhou, & Tang, 2017), Decision Trees (Sriram & Yuan, 2012), Fuzzy Logic (Wang, Lee, & Ho, 2007), Bayesian methods (WenBin, XiaoLing, YiJun, & Yu, 2010), Genetic Algorithms (Daza et al., 2011), among others).

Knowledge acquisition is a process which aims to extract knowledge, experience and problem-solving procedures from one or more domain experts. Several techniques have been proposed for a correct knowledge acquisition (see Hua, 2008 for a complete review). Nevertheless, there are several problems that must be considered when acquiring knowledge from human experts (Gaines, 1987):

- Experts may not be able to express their knowledge in a structured way.

[*] Corresponding author.

E-mail addresses: isaac.martin@urjc.es (I. Martín de Diego), oscar.sanchez@centrogeo.edu.mx (O.S. Siordia), alberto.fernandez.isabel@urjc.es (A. Fernández-Isabel), cristina.conde@urjc.es (C. Conde), enrique.cabello@urjc.es (E. Cabello).

URL: http://www.frav.es, http://www.datasciencelab.es (I. Martín de Diego), http://www.frav.es, http://www.datasciencelab.es (A. Fernández-Isabel), http://www.frav.es, http://www.datasciencelab.es (C. Conde), http://www.frav.es, http://www.datasciencelab.es (E. Cabello)

https://doi.org/10.1016/j.eswa.2018.07.058
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

- Experts may not be aware of the significance of the knowledge they have used.
- The expressed knowledge may be irrelevant, incomplete or not understandable.

In some cases, depending on the field of application, it may be easier to extract the knowledge from human experts through a continuous scale. This is the case of the risk assessment, where the knowledge could be acquired in a predefined scale (e.g. from 0, no risk, to 100, maximum risk). Here, the knowledge of the experts is gathered in form of subjective sequential data (Prelec, 2004) and could be treated as time series for its study and integration (see, for instance, de Diego, Crespo, Siordia, Conde, & Cabello, 2011; de Diego, Siordia, Conde, & Cabello, 2011; Siordia, de Diego, Conde, & Cabello, 2011a).

However, the integration of several opinions into a unique ground truth (i.e. a reliable signal) is a hard-to-achieve task (Liou & Nunamaker, 1990). Two different scenarios appear. The consideration of knowledge from too few experts could provide a ground truth with insufficient information. In contrast, the consideration of knowledge from too many experts could generate a noisy ground truth due to the appearance of possible contradictions between their evaluations (Turban, 1991). Different statistical approaches have been proposed in the past (see, for instance, meta-analysis methods in Brockwell & Gordon (2001)).

In this paper, it is presented a novel methodology for the selection of subjective sequential data for the training of ES. This methodology upgrades the previous approaches in the domain (Siordia, de Diego, Conde, & Cabello, 2014) focusing on the inclusion of more experts. This increment of sources of information leads to produce heterogeneous and noisy evaluations that have to be arranged. A novel definition of similarity between experts' evaluations will be firstly presented here. In addition, in the previous method, the agreement between two or more evaluations was enough to define a unique ground truth. However, in the present paper, all the homogeneous evaluations will be used.

Delving into the main idea behind, the methodology consists of the arrangement of a set of evaluations acquired from human experts through a hierarchical clustering technique. In this way, similarities between the evaluations of experts could be identified and grouped together, filtering the contradictions. The resulting groups (clusters) could be analyzed in order to select the most appropriate ground truth labels (Healey, 2011) for the training of the ES.

The proposed methodology is a general purpose approach. Thus, it can be used in several domains where different human opinions should be managed. In this paper, the methodology is applied to a practical case on the Intelligent Transportation Systems (ITS) domain (Alam, Ferreira, & Fonseca, 2016). It is focused on the characterization of risky or safe situations for the driving task.

Regarding the experiments, three different have been considered to illustrate the performance of the approach. First, an experiment has been developed using synthetic data for demonstrative purposes. The other experiments are based on the practical case presented above. Thus, they have been achieved using real driving risk evaluations made by experts from urban and interurban scenarios respectively.

The paper is organized as follows: Section 2 situates the approach in the domain. Section 3 introduces the proposed methodology, explaining in detail the similarity measures to evaluate subjective sequential data. Section 4 describes the practical case where the approach has been applied. Section 5 presents the achieved experiments and their most relevant results. Finally, Section 6 concludes and provides future lines of work.

2. Related work

The ES have been widely used for multiple purposes (Wagner, 2017). They are systems that are able to exhibit features associated with human intelligence (e.g. problem solving or reasoning) (Hodson, 2018). They have a common architecture based on two main modules: a domain dependent knowledge database and the inference mechanism. Examples of them are Attwell, Leask, Meyer, Rokkas, and Ward (2017) or Meza-Palacios et al. (2017).

The architecture of the ES presented here comprehends both modules. The knowledge base is acquired from traffic experts that evaluate the behavior of drivers, while the inference mechanism is built applying similarity measures and unsupervised learning techniques.

Delving into these unsupervised learning techniques, clustering (see, for instance, Aggarwal, 2015) is an initial and fundamental step in data analysis. It has as a main goal to reveal a natural partition of data into a number of meaningful subclasses or clusters. Clustering of sequential data differs from clustering of static feature data mainly in how to compute the similarity between two data objects.

In the presented approach, *Agnes* clustering algorithm has been selected. It is an agglomerative hierarchical clustering technique that provides real-time updating (see Kaufman & Rousseeuw, 2009 for a complete description).

Regarding the characteristics of subjective sequential data (where sudden changes occur and where the key information is given by its trend), it is appropriated a piecewise representation of the data. Thus, a variety of algorithms to obtain a proper linear representation of sequential data have been proposed in the literature (see, for instance, Keogh, Chu, Hart, & Pazzani, 2004; Lachaud, Vialard, & De Vieilleville, 2005; Zhu, Wu, & Li, 2007)

Focusing on driving risk situations, there are multiple examples of their characterization through the analysis of data collected on driving sessions. These approaches are usually focused on the study of the drivers behavior and how their acts affect to the driving task. For instance, Cheng, Park, and Trivedi (2007) introduces an approach based on multi-perspective (several cameras recording the driver) in order to analyze the different body movements (mainly head and hands). In the case of Malta, Miyajima, Kitaoka, and Takeda (2011), it is oriented to identify the frustration and the different emotions of the driver and how these emotions affect to the driving task. These systems are related to the approach presented in this paper. Both examples use cameras to identify the movements of the driver, though in our case the face expressions are not considered.

Other studies have their key topic in learning from specific risk situations identifying patterns. For example, Wang, Zhu, and Gong (2010) has as a main purpose to infer the safe or dangerous actions achieved by drivers using time series and unsupervised learning. In this case, the presented approach could be considered as one of this type of systems.

There are similar approaches that evaluate specific tasks of the driver and not only the hands or the facial expressions. The pressures exerted on the break and throttle pedals are also interesting parameters to evaluate. Examples of these are Sathyanarayana, Boyraz, and Hansen (2008), that is oriented to route paths recognition and Rakha, El-Shawarby, and Setti (2007), which addresses the behavior of driver in intersections.

Delving into the behavior of drivers, multiple theoretical models have been developed. They can be classified into: taxonomic models and functional models. The firsts usually produce descriptive classifications of certain elements of traffic based on a context. They can be decomposed into features-based models (Bone & Mowen, 2006) and task-analysis models (Fastenmeier & Gstalter, 2007). The second ones can be organized into mechanical models, adaptive control models and cognitive models. Mechanical models (see, for instance (Greenberg, 1959)) are based on the representation of the mutual influences among individuals. Cognitive models (see, for example MacAdam, 1981) collect information from the external environment (e.g. traffic signals or the presence of a vehicle), and produce the outputs that determine the behavior of the element according to it. Cognitive models (see, for instance Dia, 2002) use cognitive entities (usually intelligent agents) in order to emulate the complex behaviors of drivers.

These theoretical models have provided support to the practical case presented (see Section 4). Nevertheless, they are not included in the development of the approach due to the subjectivity of the traffic domain. Instead, traffic experts were selected to evaluate the drivers' behavior.

3. Methodology

The rules of the production system of an ES are usually achieved with an experimental ground truth generated at the experiment design step (see Fig. 1(a)). This experimental ground truth (i.e. a reliable signal) is usually build through interviews, process or concept mapping, commentating, card sorting, tables, or transcriptions (Hua, 2008).

However, depending on the requirements of the field of application, it may be mandatory the inclusion of subjective sequential data acquired from human experts' opinions. And as it was already mentioned, when working with subjective sequential data, it is possible to obtain noisy information due to contradictions among the acquired evaluations.

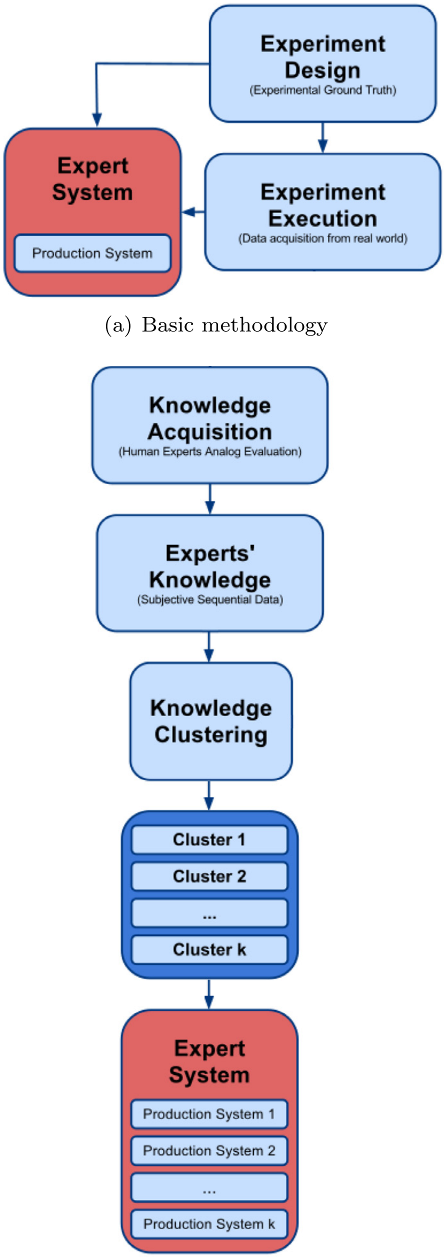
This leads to develop a methodology for the arrangement and selection of subjective sequential data acquired from a group of human experts (see Fig. 1(b)). First, the knowledge from human experts is acquired through an analog evaluation. Next, the experts' knowledge is represented in a proper set of features as subjective sequential data. Once the feature from each expert are defined, a clustering technique is applied in order to group the experts into homogeneous groups. Thus, a set of k clusters is obtained, each of them representing different experts evaluations. Finally, different prediction models are trained and tested in each cluster.

Therefore, the main idea behind the proposed methodology is the arrangement of subjective sequential data into groups in accordance to the opinions provided by the experts through a hierarchical clustering technique. For that purpose, a linear representation of the main trend of the sequential data is used.

However, special considerations must be taken into account when selecting the cut points where a linear model will be fitted. In this case, it has been used a linear segmentation algorithm based on the search of feature points where extreme changes on the data trend are produced. This method has been called Trend Segmentation Algorithm (TSA).

Summed up briefly, the TSA algorithm is as follows. The input of the algorithm is the evaluations made by the experts. A number of feature points are selected and linear regressions for each pair of consecutive points (a segment) are fitted. For each segment, if the regression error is high, the segment is divided. Otherwise, the points are stored as final points. The output of TSA is the optimal linear representation of the input evaluation achieved as a trade-off between the global error and the complexity of the representation (number of generated segments). A complete description of TSA can be seen in Siordia et al. (2011a) and Siordia, de Diego, Conde, and Cabello (2011b).

Regarding the hierarchical clustering algorithm applied, *Agnes*, an agglomerative nesting technique is selected (Guerraz et al., 2010). It uses a bottom-up approach. Thus, it is useful for the approach presented, as it provides real-time updating. Each observation starts in its own cluster. Then, clusters are merged until only one large cluster remains which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster.



(a) Basic methodology

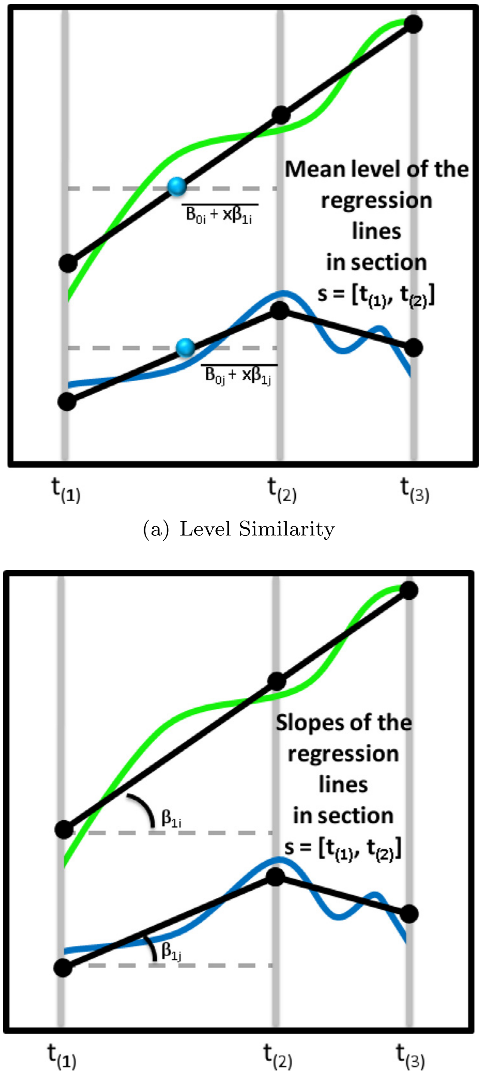
(b) Proposed methodology

Fig. 1. Comparison among schemes of the basic methodology and the methodology proposed in this work.

A similarity measure between the linear representation of the data is calculated. In order to achieve it, two different and complementary similarity measures are used (Siordia et al., 2011b) and an original method to combine then based on clustering is proposed. Next sections explain them in detail.

3.1. Similarity definitions

Given a pair of aligned linearized evaluations (f_i, f_j), it is possible to define a set of similarity measures taking advantage of the characteristics of the linear representation proposed by the TSA. In this work, two similarity measures have been proposed. They are


Mean level of the regression lines in section s = [t(1), t(2)]
B0i + xβ1i
B0j + xβ1j
t(1)
t(2)
t(3)
Slopes of the regression lines in section s = [t(1), t(2)]
β1i
β1j
t(1)
t(2)
t(3)

(a) Level Similarity

(b) Angle Similarity

Fig. 2. Similarities between two sections of two segmented sequential series, where $t_{(1)} < t_{(2)} < t_{(3)}$ are three consecutive points (instants of time) selected by the TSA algorithm.

based on the difference of levels (i.e. *Level Similarity*) and angles between the linear regression lines obtained from the linearized evaluations (i.e. *Angle Similarity*). In addition, a new similarity measure from the combination of them is also proposed.

Section 3.1.1 introduces the *Level Similarity*. Section 3.1.2 presents the *Angle Similarity*. Finally, Section 3.1.3 describes the new proposed similarity measure based on the combination of the two previous similarities and the clustering results.

3.1.1. Level similarity

Let $k = [t_{(initial)}, t_{(final)}]$ be a common section defined for the linearized sequential data f_i and f_j. The width of the section is defined by $w(k) = t_{(final)} - t_{(initial)}$. Let $\hat{Y}_i = \beta_{0i} + x\beta_{1i}$ and $\hat{Y}_j = \beta_{0j} + x\beta_{1j}$ be the regression lines fitted in the section k of f_i and f_j, respectively. The *Level Similarity* is based on the mean levels of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$ over the section k (see Fig. 2(a)).

Let d the distance of the mean levels of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$ calculated as:

$$d = \left|(\beta_{0i} + \frac{w(k)}{2}\beta_{1i}) - (\beta_{0j} + \frac{w(k)}{2}\beta_{1j})\right|$$

$$= \left|(\beta_{0i} - \beta_{0j}) + \frac{w(k)}{2}(\beta_{1i} - \beta_{1j})\right|. \tag{1}$$

The *Level Similarity* calculated in section k is obtained as follows:

$$s_L(k) = 1 - \frac{d}{\max(d)}, \tag{2}$$

where $\max(d)$ is the maximum possible distance between the mean levels. Notice that for a set of evaluations ranging in [0, 100], the maximum possible distance is 100. Thus, $s_L(k)$ is in [0,1].

The overall *Level Similarity* for f_i and f_j is calculated as the weighted sum of all the sectional similarities as follows:

$$S_L(f_i, f_j) = \frac{\sum_{k=1}^{K} w(k)\, s_L(k)}{\sum_{k=1}^{K} w(k)}. \tag{3}$$

3.1.2. Angle similarity

The *Angle Similarity* considers the angle between the regression lines defined in sections $k = 1, \ldots, K$. Let β_{1i} and β_{1j} be the slopes of the regression lines $\hat{Y}_i$ and $\hat{Y}_j$, respectively (see Fig. 2(b)). The angle between the regression lines is calculated as:

$$\theta = atan(|\beta_{1i} - \beta_{1j}|). \tag{4}$$

The *Angle Similarity* in section k, denoted by $s_A(k)$, is obtained as:

$$s_A(k) = 1 - \frac{\theta}{\breve{\theta}_k}, \tag{5}$$

where the worst angle $\breve{\theta}_k$ is calculated as the maximum possible change in section k:

$$\breve{\theta}_k = atan\left|\frac{2\max(d)}{w(k)}\right|. \tag{6}$$

Notice that $s_A(k)$ is in [0,1].

The overall *Angle Similarity* for the linearized sequential data f_i and f_j is calculated as the weighted sum of all the sectional similarities as follows:

$$S_A(f_i, f_j) = \frac{\sum_{k=1}^{K} w(k)\, s_A(k)}{\sum_{k=1}^{K} w(k)}. \tag{7}$$

3.1.3. Clustering similarity

In de Diego, Muñoz, and Moguerza (2010), the Pick-out method is used to fuse information from several feature representations employing label information for classification task. The input of the method are two similarity measures between two different points in a training data set. When the two points belong to the same class (they have the same label), the final similarity is defined as the maximum of the two original ones. When the two points belong to different classes (they have different labels), the final similarity is defined as the minimum of the two original ones. The same idea is here proposed to combine the *Level Similarity* and the *Angle Similarity* for clustering.

First, the *Level Similarity* and the *Angle Similarity* are used in two different and independent cluster analysis. Let $C_L(f_i, f_j) = 1$ if the linearized sequential data f_i and f_j are grouped in the same cluster when the *Level Similarity* S_L is used, and $C_L(f_i, f_j) = 0$ otherwise. In the same way, let $C_A(f_i, f_j) = 1$ if the linearized sequential data f_i and f_j are grouped in the same cluster when the *Angle Similarity* S_A is used, and $C_A(f_i, f_j) = 0$ otherwise.

As the *Level Similarity* and the *Angle Similarity* have been developed to collect different characteristics of the data, we propose to

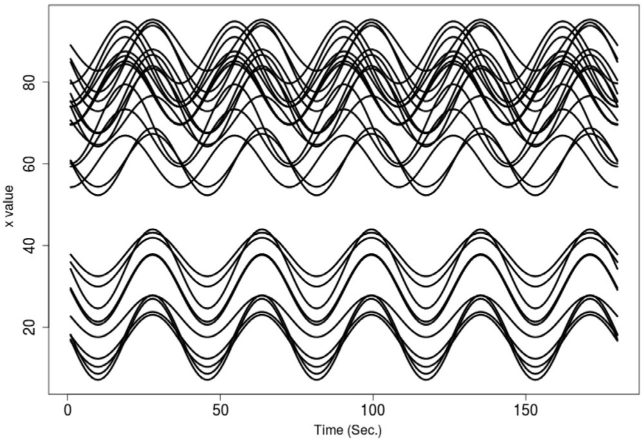


Fig. 3. Synthetic data: sines and cosines with random bias and amplitude.

build the *Clustering Similarity* as follows: $S_C(f_i, f_j) =$

$$\begin{cases} \max(S_{\mathrm{L}}(f_i, f_j), S_{\mathrm{A}}(f_i, f_j)), & \text{if } C_L(f_i, f_j) = C_A(f_i, f_j) = 1\,, \\ \dfrac{S_{\mathrm{L}}(f_i, f_j) + S_{\mathrm{A}}(f_i, f_j)}{2}, & \text{if } C_L(f_i, f_j) \neq C_A(f_i, f_j)\,, \\ \min(S_{\mathrm{L}}(f_i, f_j), S_{\mathrm{A}}(f_i, f_j)), & \text{if } C_L(f_i, f_j) = C_A(f_i, f_j) = 0\,. \end{cases} \tag{8}$$

Thus, if f_i and f_j are grouped in the same class when both similarities are used, it is guaranteed that $S_C(f_i, f_j)$ will be the largest possible according to the available information.

In addition, if f_i and f_j are grouped in different classes when both similarities are used, it is guaranteed that $S_C(f_i, f_j)$ will be the lowest possible according to the available information.

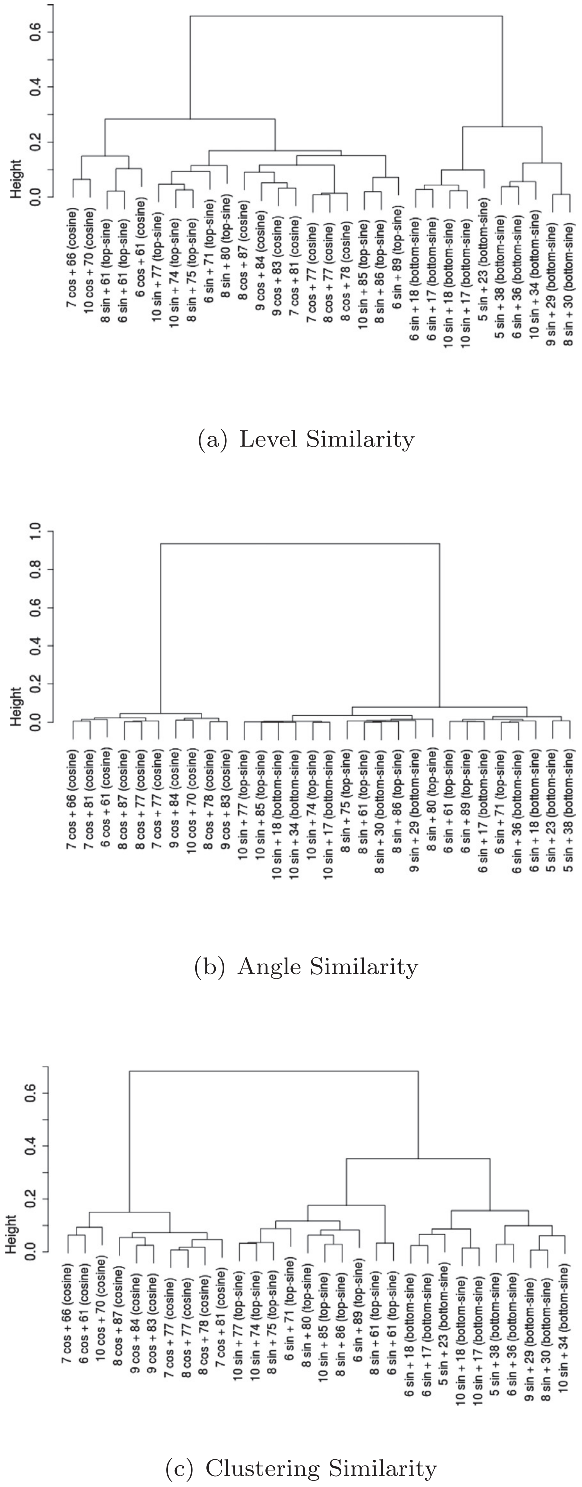
Further, if f_i and f_j are grouped in the same or in different classes according to the similarity used, the average of the two similarities is considered. Hence, as the original Pick-out method (for classification tasks), the proposed similarity tends to move closer those sequential data belonging to the same group, and tends to separate points belonging to different groups.

3.2. Synthetic example

In order to illustrate the proposed methodology, an experiment with synthetic data is presented. A set of 30 random series of a numerical variable x were generated from 0 to 180 seconds:

- 10 sines with bias from 50 to 90 and amplitude from 5 to 10 (top-sines).
- 10 cosines with bias from 50 to 90 and amplitude from 5 to 10 (cosines).
- 10 sines with bias from 10 to 50 and amplitude from 5 to 10 (bottom-sines).

Thus, the most difficult cases when working with subjective data have been considered: conflictive opinions (sines and cosines) and conflictive level (top-sines and bottom-sines). Fig. 3 shows the 30 series generated for this example. Following the proposed methodology, the TSA algorithm has been applied to all the data in order to transform the original information into proper linear representations. Then, both the *Level Similarity* and the *Angle Similarity* are calculated. Fig. 4 shows a dendrogram of the arrangement produced by the *Agnes* clustering technique using the *Level Similarity*, the *Angle Similarity*, and the *Clustering Similarity*.



(a) Level Similarity

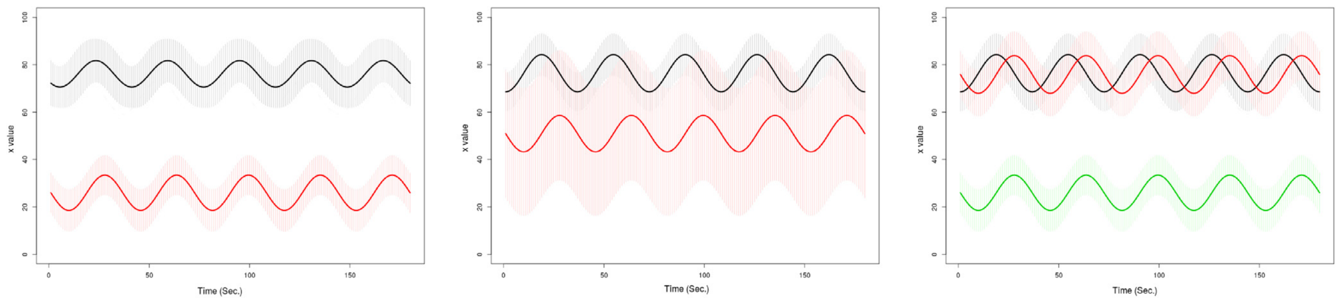
(b) Angle Similarity

(c) Clustering Similarity

Fig. 4. Clusters of the synthetic data (Dendrograms).

For the *Level Similarity*, two main groups have been clearly identified by its level. In the first one, the series generated with a bias from 50 to 90 (top-sines and cosines) are grouped together. The mean of the series belonging to this group is shown in a black line in Fig. 5(a). The standard deviation of this group is shown as a black shadow behind the mean line. In the second group, the series generated with a bias from 10 to 50 formed a second group (bottom-sines). The mean and standard deviation of the series belonging to this second group are shown in red in the same figure (see Fig. 5(a)).

In the same way, two main groups have been clearly identified by the *Angle Similarity*. The first group, formed by all the generated cosines, is shown in black in Fig. 5(b). The second group, formed



(a) Level Similarity (b) Angle Similarity (c) Clustering Similarity

Fig. 5. Clusters of the synthetic data (Series). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by all the generated sines (top-sines and bottom-sines) is shown in red in the same figure. In this case the series have been grouped by its behavior (sines and cosines).

Finally, for the *Clustering Similarity* three main groups were identified. The mean and standard deviation of each group are shown on Fig. 5(c). In this case, the top-sines (shown in red), bottom-sines (shown in green), and cosines (shown in black) were pooled in individual groups.

In all the cases, the arrangements produced by the cluster analysis shows the properties of each of the proposed similarities. The *Level Similarity* was able to separate the data by its level. Also, the *Angle Similarity* was able to separate the data by its behavior. Further, the *Clustering Similarity* (as a proper combination of both similarities) was able to separate the synthetic data in each of the three simulated groups, separating data with conflictive angles (sines and cosines) and conflictive levels (top-sines and bottom-sines).

4. Practical case

The study of driver's behavior has become a topic of interest in the last years for the ITS domain (Kircher & Ahlström, 2009). It has been estimated that 25 − 50% of all vehicle crashes are caused by human factors (Ranney, Mazzae, Garrott, & Goodman, 2000; Wang, Knipling, Goodman et al., 1996).

CABINTEC (Intelligent cabin truck for road transport) is a project focused on the study of human factors for the improvement of traffic safety (Brazalez, Ares, & Matey, 2006). The key idea of this project is the development of an ES to provide assistance to drivers (see de Diego, Crespo et al., 2011 for a complete description). Thus, the main task of the system consists of the timely notification of imminent risky situations. It uses a buffered risk model to penalize inappropriate driver's behavior (activators) and to praise correct driver's behavior (inhibitors) with a set of rules.

However, the characterization of risky or safe situations for the driving task is hard to achieve due to its subjective entity and the great number of factors involved (Schneider & Kiesler, 2005; Siordia, de Diego, Conde, Reyes, & Cabello, 2010; Zhang, Schreiner, Zhang, & Torkkola, 2007). One of the most common problems related to the development of driving risk detection systems is the absence of a reliable risk signal (i.e. a driving risk ground truth) against which they could be compared and evaluated (Kircher & Ahlström, 2009).

The proposed methodology fits properly in this context. It allows to analyze and select subjective sequential data acquired from human experts in order to generate a reliable driving risk ground truth for the training of automatic risk detection systems.

Section 4.1 describes the data gathering step focusing on the tools and techniques used. Section 4.2 introduces the knowledge acquisition step where the human experts evaluate the data previously collected. Finally, Section 4.3 presents the cluster analysis.

4.1. Database acquisition

The database used in this practical case has been collected from a set of driving sessions executed by a professional driver using a truck simulator. This simulator is located at the Centre of Studies and Technical Research of Gipuzkoa (Centro de Estudios e Investigaciones Técnicas de Gipuzkoa (CEIT)). It presents a real truck cockpit mounted over a Gough–Stewart platform (Zakaria, Abdel-Moneim, Abdin, Hafez, & Darwish, 2017) to provide a natural driving sensation (see Fig. 6(a)). Furthermore, the driver's visual field is covered by a detailed simulated 3D scene using rear projection onto three independent screens (left, center and right).

The data acquisition process was performed during four driving sessions using two different scenarios: urban and interurban. Each driving session had a length between 3 and 5 minutes. The first driving session recorded in each scenario will be used for training purposes in the experiments. The second driving session of each scenario will be used for testing purposes in the experiments.

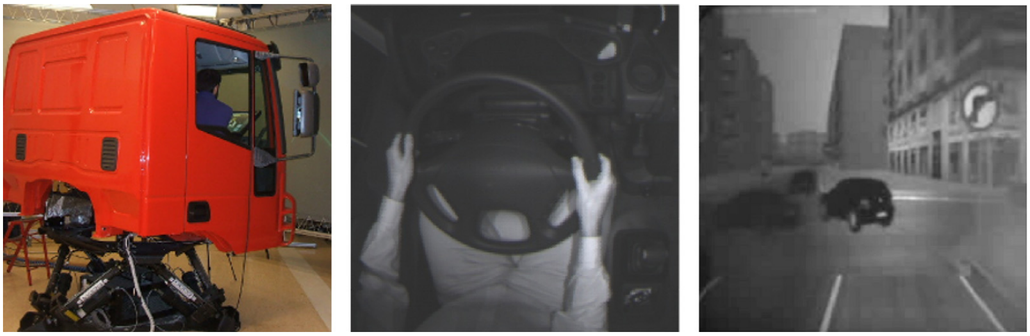
The three basic elements of traffic safety: driver, road and vehicle have been covered on the recording of the driving sessions. Data registers of the vehicle dynamics and road characteristics have been obtained from the simulator. Table 1 shows a brief description of the vehicle and road variables.

Visual information has been obtained from two video sources: sequences of the driver's top view (see Fig. 6(b)) and sequences of the simulator central screen, which presents the main view of the road to the driver (see Fig. 6(c)). Both scenarios involved real traffic and interactions with other vehicles.

Further, in order to induce risky situations, a set of tasks have been designed on the laboratory to be executed by the driver at specific time periods (i.e. the driver follows a set of predefined guidelines). Fig. 7 shows the time periods where the risky situations were induced along the driving sessions.

In this kind of experiments, the driver behavior is usually labeled through a binary signal in accordance to these planned time periods. Thus, a binary label is defined: risk behavior vs. safe behavior. Instances of the first one are impaired driving, non use of seat belts, speeding, following too closely or traffic violations (Simons-Morton, Lerner, & Singer, 2005). Instances of the second could be the opposite ones (e.g. drivers respect the traffic normative and their aggressiveness levels are low).

However, as shown in Siordia et al. (2011a) and de Diego, Siordia, Conde, and Cabello (2012), this binary experimental risk signal does not provide enough information about the quantitative real risk that the driver is taking at each moment. Risk perception is subjective (Sjöberg, Moen, & Rundmo, 2004) and it is also related to the skills of the driver. Thus, these skills could lead into a good

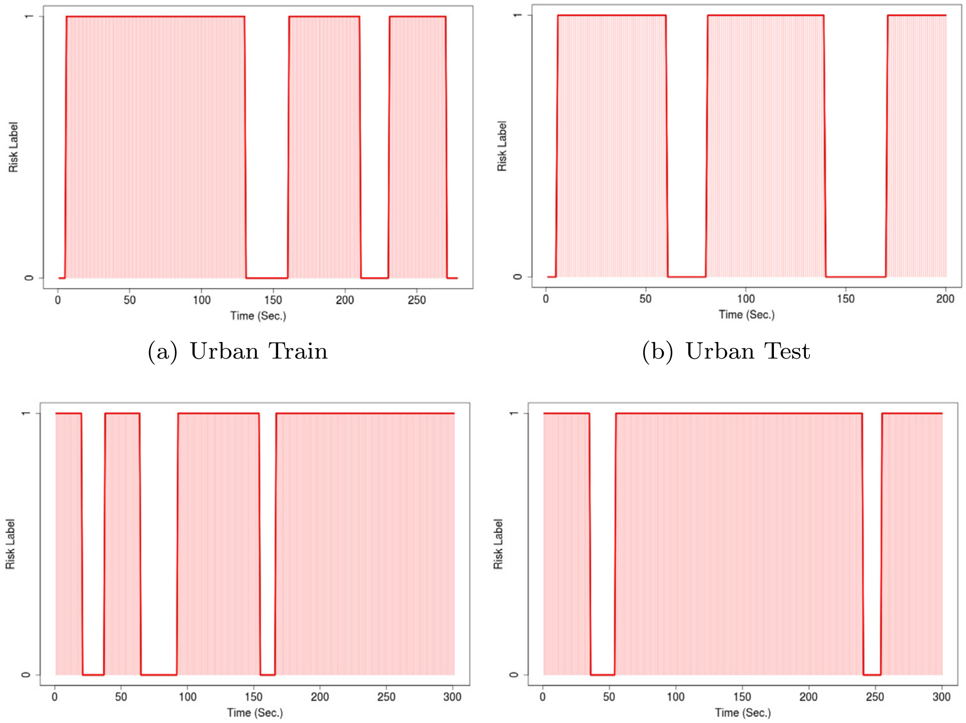


(a) Truck simulator (b) Driver's top view (c) Road main view

Fig. 6. Simulator used on the acquisition process and samples of visual information acquired.

Table 1
Variables acquired during the driving sessions recorded on the truck simulator.

Variable	Description
HANDS POSITION	Number of hands on the steering wheel (a) and on the gear stick (b) coded as "a-b"
ELAPSED TIME	Time elapsed from the start of the simulation and the current instant.
LANE INVASION	Establishes whether the vehicle is invading the opposite lane.
SPEED LIMIT	Establishes whether the vehicle speed is higher than the speed limit of the road.
BRAKE PEDAL	Percentage of pressure applied to the brake pedal.
BRAKING	First derivative of the pressure in the brake pedal.
ACCELERATOR PEDAL	Percentage of pressure applied to the accelerator pedal.
ACCELERATION	First derivative of the pressure in the accelerator pedal.
SECURITY DISTANCE	Distance between the desired vehicle and the next vehicle in the simulation.
SPEED	Instantaneous speed of the desired vehicle.
STEERING WHEEL	Instantaneous angle of the steering wheel.
LINEARITY	First derivative of the steering wheel angle.
HEADING ERROR	Difference between the actual heading and the desired road heading.
LATERAL POSITION	Distance between the desired vehicle center and the road center.
ROAD SLOPE	Slope of the road.



(a) Urban Train (b) Urban Test

(c) Interurban Train (d) Interurban Test

Fig. 7. Time periods were the risky situations were induced along the four driving sessions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

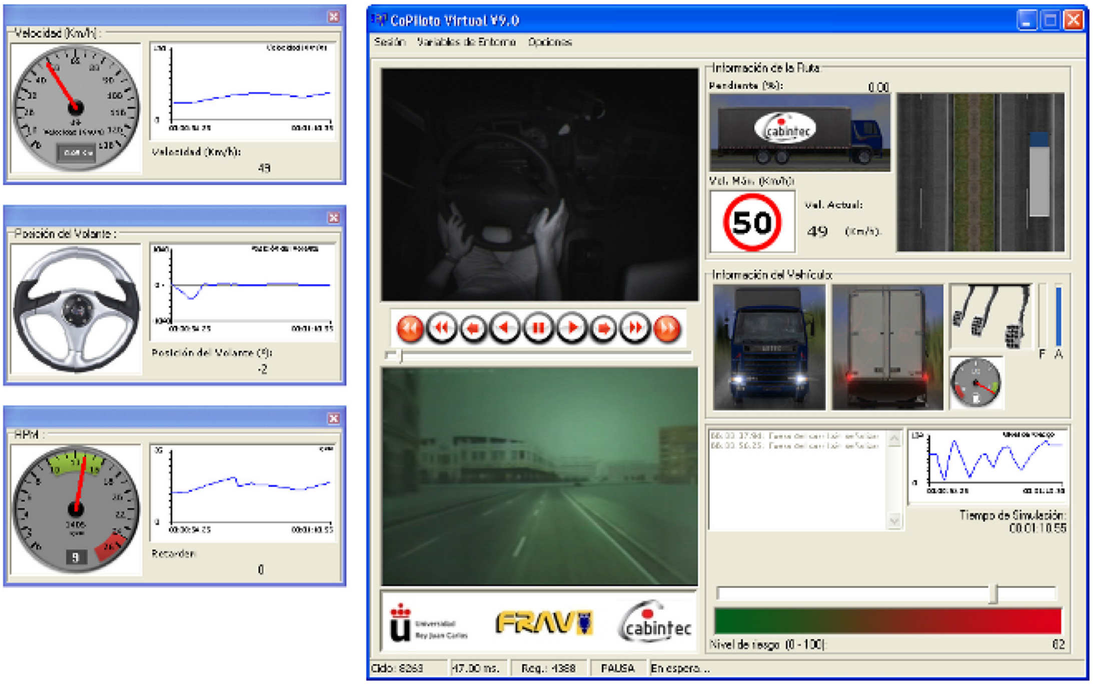


Fig. 8. Graphical User Interface of the Virtual Co driver tool.

driving task during the induced risky situations. On the other side, the driver could generate unexpected risky situations related to the decision making. Therefore, the risk level must depend on the performance of the driving task. Also, the driving risk should not be assessed with only a binary value but the assessment of human experts is needed to produce a driving risk ground truth.

4.2. Knowledge acquisition

In order to generate a proper ground truth for the development of a driving risk detection system, a group of traffic safety experts have evaluated the driving risk in each driving session. All the traffic safety experts are participants of the Master of Urban Mobility, or the Master in Data Science, at the Rey Juan Carlos University (CETINIA, 2018) with driving license and at least ten years of experience in driving tasks.

An effective knowledge acquisition tool called Virtual Co driver (Siordia, de Diego, Conde, & Cabello, 2012) (see Fig. 8) has been used to collect the experts knowledge. This tool is able to reproduce the simulated exercises with a high fidelity using all the data acquired in each driving session.

The Virtual Co driver system allows the evaluation of the driving risk through a Visual Analog Scale (VAS) (Couper, Tourangeau, Conrad, & Singer, 2006) in a range from 0 to 100, where 100 refers to the highest driving risk level. This method has been considered the best for subjective measurements (see, for instance, Cork et al., 2004).

Regarding the individuals in charge of the assessments, a group of 46 experts have evaluated the risk of the driving sessions recorded on the urban scenario. In the case of the interurban scenario, a group of 17 experts have evaluated the risk presented in the different driving sessions. Fig. 9 illustrates the VAS evaluations generated by the traffic safety experts.

The evaluations of the experts present a high heterogeneity. This leads to generate difficulties to identify agreement zones between the series. Thus, it is not clear whether the experts were able to detect the induced risky situations, or the specific risk level assigned to each risky situation.

This is a typical situation when working with subjective data. Going deeper, it can be identified two specific problems. The first one is related to the conflictive levels where the data differ significantly in the level assigned at a specific time. The second one is related to the conflictive evaluations (i.e. opinions) where the data show contradictions (e.g. two segments with different slopes but with the same risk level).

4.3. Clustering

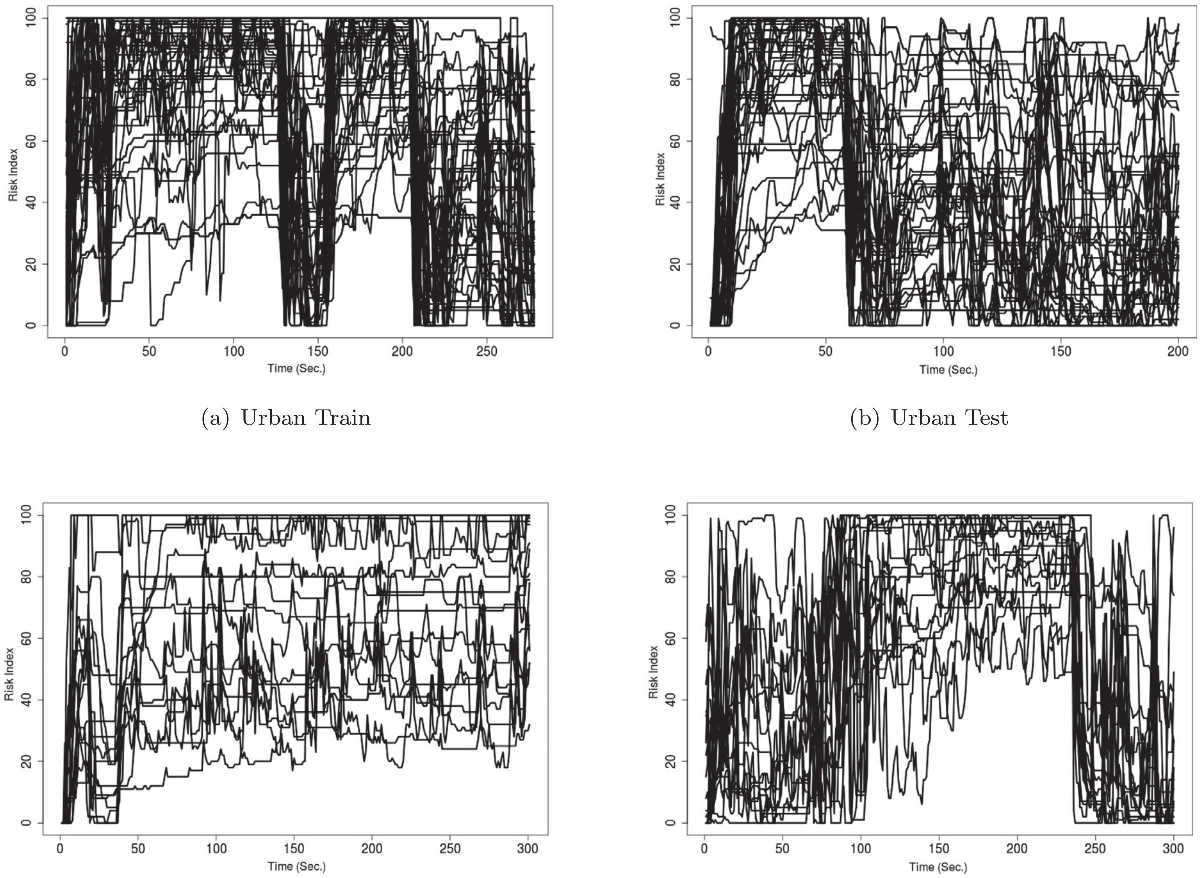
In order to facilitate the analysis of the evaluations made by the experts, it is necessary to arrange the subjective data into groups. These latter must be organized in order to filter and classify the homogeneous evaluations (i.e. grouping similar evaluations) of the experts. For this purpose, the characteristics of each generated group should be analyzed separately.

The proposed methodology is used to extract homogeneous clusters from a set of subjective data collected from a group of driving experts. First, the VAS evaluations generated by the experts were characterized using the TSA algorithm. Then, the *Clustering Similarity* proposed in Section 3.1.3 is calculated following formula (8). After that, the *Agnes* hierarchical clustering algorithm is applied using the former similarity. Thus, the experts are grouped into several clusters.

5. Experiments

The proposed methodology has been applied to the practical case presented in the previous section in two different experiments. Section 5.1 describes the first experiment where the dataset was collected from an urban scenario while Section 5.2 illustrates the second experiment where the dataset was obtained from an interurban scenario.

As explained in Section 4.1, two sessions were collected in each scenario. The first session in each scenario is used in the training step, and the second session in each scenario is used for testing purposes.



(a) Urban Train

(b) Urban Test

(c) Interurban Train

(d) Interurban Test

Fig. 9. VAS evaluations (Risk Index) of the four recorded driving sessions acquired from the human experts.

In the training step, the clustering technique groups homogeneous evaluations. In each cluster, a specific risk model has been trained. The average of all the evaluations in each cluster and the standard deviation of these evaluations have been employed as a driving risk ground truth for the generation of rules for the ES. The aim of this step is to discover the main variables that each group of traffic safety experts were observing while evaluating the risk in the driving session. In order to generate a buffered risk signal as similar as possible to the driving risk ground truth (mean evaluation of each cluster), the parameters (rules) of the risk model have been tunned with the application of a Genetic Algorithm following the work presented in de Diego, Siordia, Crespo, Conde, and Cabello (2013). Thus, it will be possible to detect those relevant variables for each group of experts. Then, each risk model (one per cluster) generated in the training step will be used to generate a risk signal in the testing step. Thus, given a new evaluation related to a new driving session, it is assigned to the nearest cluster. Next, the trained model in the cluster is used to predict a risk signal for the driving session.

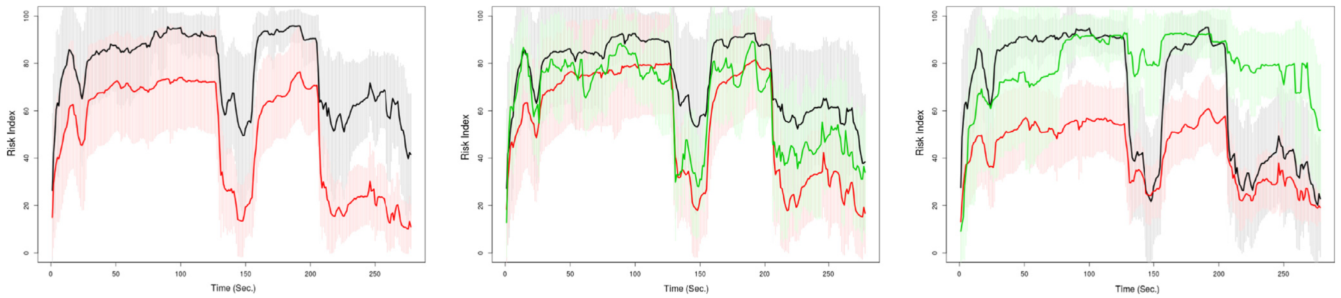
For comparison reasons, three alternative models have been considered: a General Model, Logistic Regression (LR) (Menard, 2018) and Support Vector Machine (SVM) (Tsochantaridis et al., 2004). The General Model is a risk model that uses the evaluations of all the traffic safety experts involved on the experiment. That is, no cluster analysis is considered. In the case of LR and SVM, no evaluations from experts are used. Instead, to train these models, the induced risky situations define the response variable. The response is 1 during the risky situations and 0 out of those periods of time (see Fig. 7). The LR model uses the binary induced experimental risk signal for training purposes. The SVM model also uses the binary induced experimental risk signal for training, considering only the most significant variables in Table 1. The LR and SVM models will predict binary risks in the testing step.

5.1. Urban experiment

In this experiment, the 46 risk evaluations from an urban scenario have been considered (see Section 4 and Fig. 9 for a complete description). Fig. 10 shows the mean and standard deviation of the risk evaluations that belong to each produced cluster.

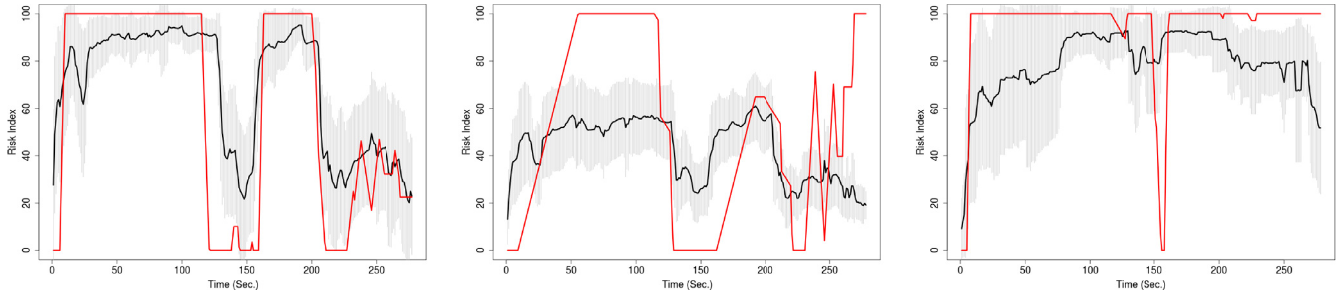
For the *Level Similarity* (see Fig. 10(a)), two different groups have been identified. Although both groups present similar opinions, the range used for the risk assessment of each group is different. The first one is a group of 24 traffic safety experts that performed their risk evaluations with a level in a range from 0 to 95 (black line). The second one is group of 22 traffic safety experts with risks evaluations with a mean level in a range from 0 to 76 (red line).

For the *Angle Similarity* (see Fig. 10(b)), three different groups have been found. These groups, of 17, 26, and 6 traffic safety experts, respectively, present similar risk levels with different behaviors. The first group (black line) presents a mean level risk evaluation in a smaller range. In general, the risk level descends in lower runs when it is compared with the other two groups. That is, the risk assessment of the traffic safety experts who belong to this group presents a long-term memory effect. The second group (red



(a) Mean Similarity (b) Angle Similarity (c) Clustering Similarity

Fig. 10. Obtained clusters in the Urban experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 11. Risk buffers generated for each cluster in the training step for the Urban experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

line) presents a set of risk evaluations with the highest ascending and descending runs. The risk assessment presents a short-term memory effect. The third group (green line) contains characteristics of the other groups. Further, the mean risk evaluation presents a higher variability. Thus, they show an irregular roughness in the evaluations made by the traffic safety experts who belong to this group.

For the *Clustering Similarity* (see Fig. 10(c)), three groups have been obtained. As in the synthetic example presented in Section 3.2, two groups with pretty similar evaluations have been identified (red and green lines). Both groups present a long-term memory effect at different levels (high and low, respectively). The other group (black line), presents a mean level evaluation with a higher range. In this case, the risk assessment shares the minimum and maximum values of the other two clusters. This group presents a short-term memory effect allowing a wider range for the risk assessment. This provides richer information about the driving risk along the driving sessions.

Fig. 11 shows the risk signals generated by the buffered risk models learned for each cluster in the training step. In all the cases, the values of the buffered risk signals (red lines) have been pretty similar to the mean risk signals used as the driving risk ground truth (black lines).

For the first group (i.e. Cluster 1), the three main generated risk peaks have been correctly detected (see Fig. 7(a)). The maximum values reached by the buffered risk signal (red line) are similar to the ones reached by the driving risk ground truth (black line).

For the second group (i.e. Cluster 2), the maximum levels mismatched at different time lapses along the driving session. These results show that the traffic safety experts grouped in this cluster were not consistent with the criteria used for the risk assessment. In this case, the three main risk peaks were evaluated by the traffic safety experts applying different approaches and it was not possible to adjust a buffered risk model properly.

Table 2
Risk model generated with the evaluations of the Cluster 1 in the Urban experiment.

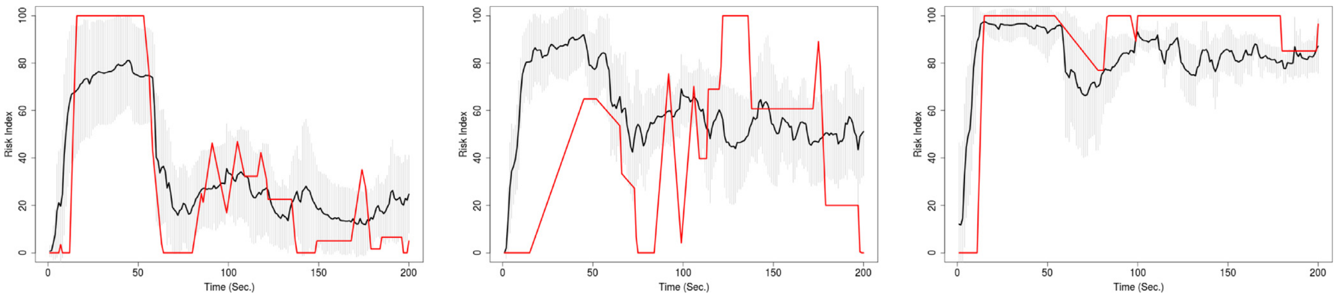
Hands Code	Slope		L. Time
2-0	−7.9		7.4
1-1	12		0.21
1-0	5		1.7
0-1	13.9		1.6
0-0	26		4
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Heading Error	−8.7	< > 0.98	4.5

Table 3
Risk model generated with the evaluations of the Cluster 2 in the Urban experiment.

Hands Code	Slope		L. Time
2-0	−0.87		6.3
1-1	8.6		9.5
1-0	9.4		5.1
0-1	10.2		8.7
0-0	2.2		4.7
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Heading Error	−20	< > 0.86	5.2
Vehicle Speed	29.31	> 89	4.5

For the third group (i.e. Cluster 3), as the risk assessment (black line) has been made at high risk levels at most of the time, the buffered risk signal (red line) follows a similar behavior providing poor information about the three main risky situations.

As explained before, a Genetic Algorithm is used to learn the most relevant variables to generate the risk signal in each cluster. The variables of the buffered risk models learned for the three clusters are shown in Tables 2, 3, and 4, respectively. Two param-



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 12. Risk buffers generated for each cluster in the testing step for the Urban experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Risk model generated with the evaluations of the Cluster 3 in the Urban experiment.

Hands Code	Slope		L. Time
2-0	-0.96		8.9
1-1	18		2.6
1-0	11		2.8
0-1	22		3.7
0-0	28		2.6
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Heading Error	-15	$<>0.88$	9.9

eters have been considered for each variable: the slope, that represents the change per second in the risk buffer, and the latency time, that represents the delay before that change.

The first cluster obtained the lowest latency times allowing a faster response when detecting risky situations. It leads to a short-term memory effect. Thus, the most significant positive slope (penalizing slope) has been given by the situation of driving with no hands on the steering wheel (26). In this situation, after a latency time of 4 seconds the risk buffer would reach its maximum value in less than 4 seconds (0 to 100 in 8 seconds). Moreover, when the vehicle is almost parallel to the road (Heading Error $<0.98°$) and the driver has two hands on the steering wheel, the risk buffer would reach its minimum value in less than 12 seconds (100 to 0 in 12 seconds).

The second and third clusters have been obtained the highest latency times (see Tables 3 and 4). This is a consequence of the long-term memory effect shown by the traffic safety experts bunched in these two groups. In both cases, the action of driving with two hands on the steering wheel reduces the risk level in less than one risk level per second after a long latency time. Further, the most significant prizing slope was given by the heading error in both cases. For these groups of traffic safety experts, the vehicle position has been the most relevant factor for a safe driving.

Once the training step is completed, the buffered risk models learned during this step has been applied to the testing sessions recorded in the urban scenario. In order to evaluate the performance of the learned risk models, the evaluations of the traffic safety experts bunched on the three groups have been considered (see Fig. 9). Fig. 12 shows the buffered risk signals generated by the buffered risk models (red lines) and the mean risk evaluations acquired from the traffic safety experts (black lines).

For the first group (i.e. Cluster 1), significant changes on the level of the buffered risk signal have been found at the three main risky situations (see Fig. 7(b)). Thus, the level assigned to each situation by the buffered risk signal is pretty similar to the one assigned by the traffic safety experts.

For the second group (i.e. Cluster 2), the risk signal generated by the buffered risk model differs significantly at the second half of the driving session (second 100 to 200). This is a consequence of the inconsistency presented by this group of traffic safety experts when evaluating the driving risk.

For the third group (i.e. Cluster 3), the risk signal generated by the buffered risk model shows a similar behavior to the risk evaluation made by the traffic safety experts. Nevertheless, as in the training session, the high levels given by both risk signals provide poor information about the three main risky situations.

For this experiment, it can be concluded that the buffered risk model generated with the knowledge of the 17 traffic safety clustered in the first group shows the best performance in both training and testing steps.

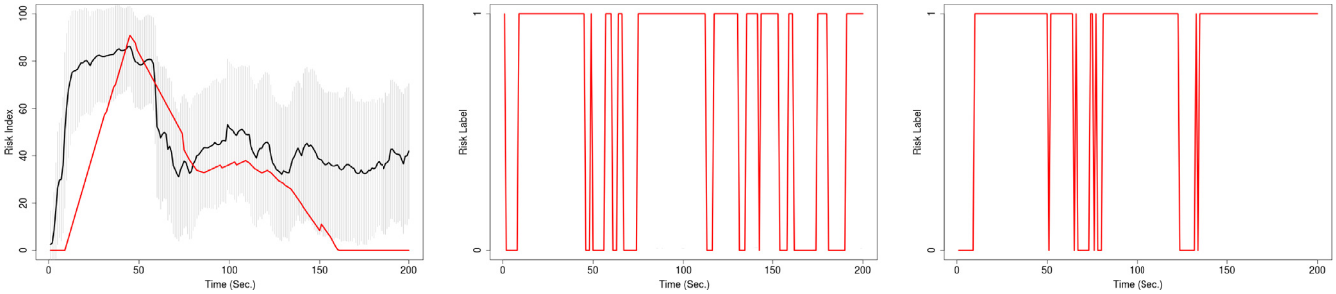
Fig. 13 shows the risk signals generated by the alternative models (General Model, LR and SVM models, respectively) on the testing sessions of the urban scenario.

The General Model uses the risk evaluations of all the 46 traffic safety experts involved. In this case, the buffered risk signal presents only significant changes on the first risky situation and it has not been able to detect the second and third risky situations. Therefore, the use of traffic safety experts' evaluations with different criteria on the generation of a risk model leads to a loss of information.

For the LR model the most significant variables acquired from the driver, the vehicle and the road have been used as predictors. Here, the generated risk signal could produce a lot of false alarms. Further, as the risk signal is built with binary values (0 or 100), the predicted risk provides poor information about the real risky situations, failing to assess each risky situation with a specific risk level as the traffic safety experts do.

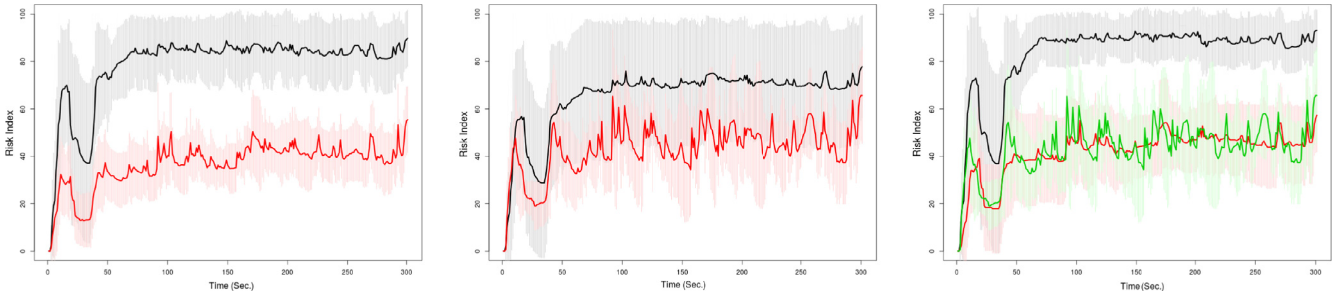
Regarding the SVM model, it detects the three main risky situations of the testing driving session with few false alarms. However, as in the previous case (i.e. LR model), the predicted risk provides poor information about the detected risky situations as all the risky situations are assessed with the same risk level.

In conclusion, in all the cases the buffered risk models trained considering homogeneous information (clusters 1, 2, and 3) have performed better than the buffered risk model trained with all the evaluations (i.e. General Model). The results show that the risk level assigned by the traffic safety experts to each of the main risky situations is very important for the risk assessment. It is clear that the risk involved in each risky situation must be evaluated depending on the performance of the driving task. Thus, the binary experimental risk signal would not provide enough information to be used as a driving risk ground truth.



(a) General Model (b) Logistic Regression (c) SVM

Fig. 13. Risk buffers generated in the testing step for the Urban experiment.



(a) Mean Similarity (b) Angle Similarity (c) Clustering Similarity

Fig. 14. Obtained clusters in the Interurban experiment.

5.2. Interurban experiment

In this experiment, the 17 risk evaluations from the interurban scenario have been considered (see Section 4 and Fig. 9 for a complete description). Fig. 14 shows the mean and standard deviation of the risk evaluations that belong to each cluster.

For the *Level Similarity*, two main groups with similar evaluations have been identified. In both cases, after the second risky situation, the driving risk mean level remains constant showing a long-term memory effect. The first group (black line), presents a stable risk evaluation in a high risk level (mean over 50). The second group (red line), presents a risk evaluation in a medium range (mean under 50). In this case, the variability of the risk evaluations made by the traffic experts bunched in this group show an irregular roughness.

For the *Angle Similarity*, two main groups have been obtained. A group of traffic safety experts that performed a stable risk evaluation along the whole driving session (black line). Another group of traffic safety experts that show a strong irregular roughness in their evaluations (red line). Both groups show a long-term effect.

For the *Clustering Similarity*, three groups have been identified. The first two groups (red and green lines) show a long-term memory effect at different levels (high and low, respectively). The third group (green line) presents a long-term memory effect with high variability.

Following the same process described in the previous Urban experiment, the usefulness of each cluster produced with the combination of both the *Level Similarity* and the *Angle Similarity* has been evaluated. Each mean evaluation and its standard deviation have been used as a driving risk ground truth for the generation of a set of buffered driving risk models. Fig. 15 shows the buffered risk signals generated by the buffered risk models learned at the training step.

In all the cases, the buffered risk signals (red lines) show similar values to the mean evaluations used as driving risk ground truth (black lines).

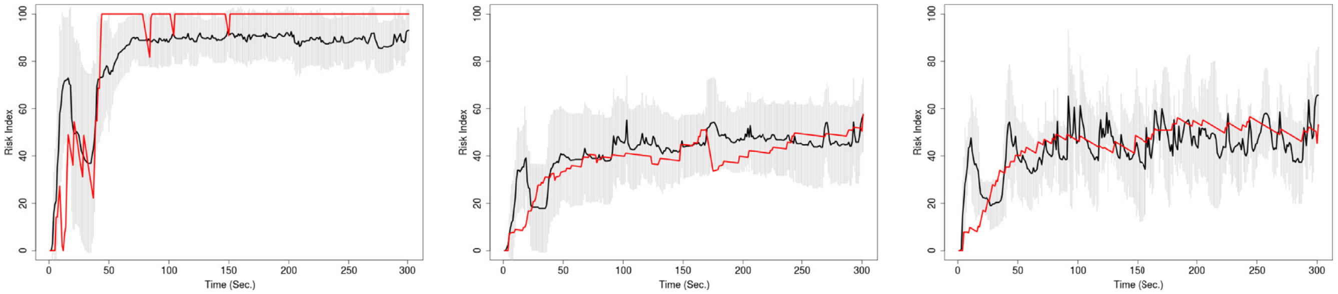
Table 5
Risk model generated with the evaluations of the Cluster 1 in the Interurban experiment.

Hands Code	Slope		L. Time
2-0	−26		7.3
1-1	12		6.4
1-0	20		1.5
0-1	22		5.3
0-0	15		6.2
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Lane Invasion	−3	= 0	7.1
Speed	−5.9	< 18	2.1
Heading Error	−13	< > 0.22	7.7
Security Distance	18.5	< 25	7.1

As in the Urban experiment, a Genetic Algorithm is used to learn the significant variables and parameters for each detected group of experts. For the first group (i.e. Cluster 1), the high risk levels used by the traffic safety experts, produced a buffered risk model with very high slope values and latency times (see Table 5). The buffered risk signal shows a very long-term memory effect, providing poor information about some induced risky situations.

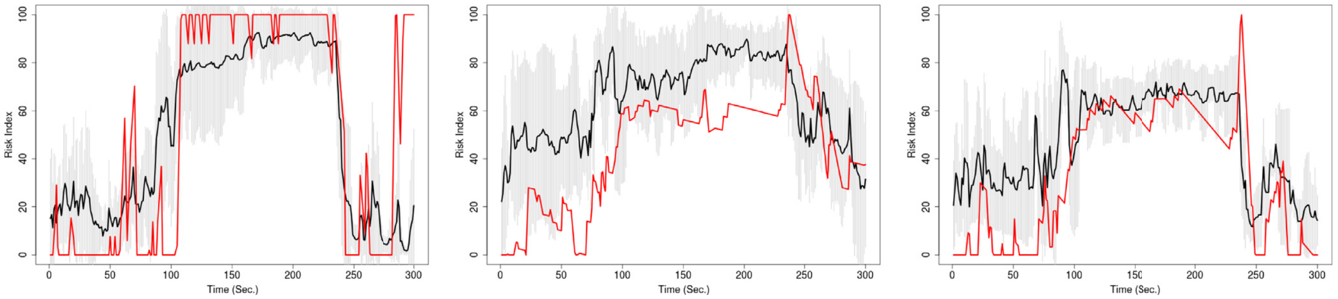
For the rest of groups (i.e. Cluster 2 and Cluster 3), the optimal parameters present a set of low slope values and latency times (see Tables 6, and 7, respectively). In both cases, the buffered risk models generate smoothed signals (red lines). This properly fits the ones used as driving risk ground truth (black lines). In the third group, the strong irregular roughness, presented by the variability of the mean risk evaluation, has been smoothed keeping its main trend.

In general, the traffic safety experts bunched in the first group (i.e. Cluster 1) have penalized harder the action of driving with only one hand over the steering wheel regardless of the performance of the driving task (0 to 100 in less than 7 seconds). Also, the traffic safety experts bunched in the other groups (i.e. Cluster 2



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 15. Risk buffers generated for each cluster in the training step for the Interurban experiment.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Fig. 16. Risk buffers generated for each cluster in the testing step for the Interurban experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Risk model generated with the evaluations of the Cluster 2 in the Interurban experiment.

Hands Code	Slope		L. Time
2-0	−1.1		0.12
1-1	1.2		1.3
1-0	0.16		0.41
0-1	1.1		1.2
0-0	1.5		0.28
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Lane Invasion	−0.22	= 0	2.5
Lateral Position	−3.1	< 0.88	1.9

Table 7
Risk model generated with the evaluations of the Cluster 3 in the Interurban experiment.

Hands Code	Slope		L. Time
2-0	−4.6		0.32
1-1	1.8		0.18
1-0	0.14		0.17
0-1	1.7		0.20
0-0	2.6		0.12
Actuators (Inhibitors and Activators)			
Variable	**Slope**	**Condition**	**L. Time**
Lane Invasion	−0.31	= 0	2.3

and Cluster 3) get used to the one hand driving performed by the driver and penalized harder the action of driving with no hands on the steering wheel.

Next, the buffered risk models learned have been applied to the testing sessions recorded in the interurban scenario. Fig. 16 shows the buffered risk signals generated by the buffered risk models (red lines) and the mean risk evaluations acquired from the traffic safety experts (black lines).

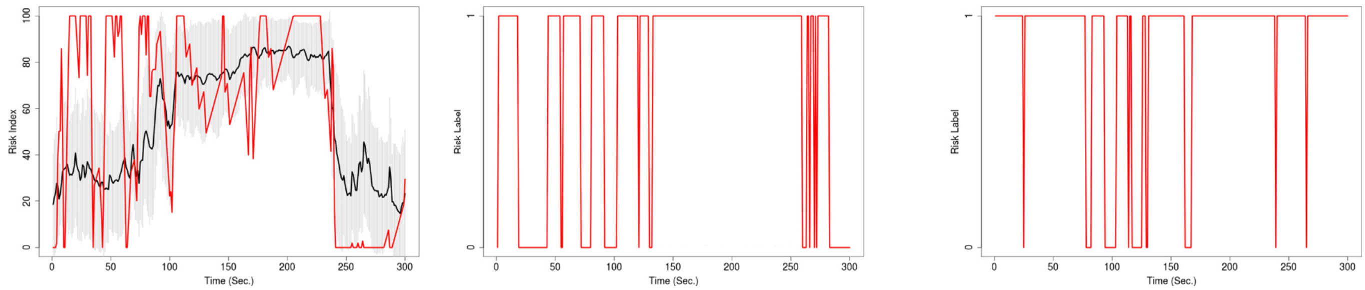
In all the cases, significant changes on the level of the buffered risk signal have been found on the main risky situations induced along the driving session (see Fig. 7(d)). The best results have been obtained by the buffered risk models of the first two groups (i.e. Cluster 1 and Cluster 2). In both cases, the buffered risk signals (red lines) fit properly the mean risk signals acquired from the traffic safety experts (black lines).

Regarding the three alternative models (General Model, LR and SVM), they have been considered in order to compare the performance of the generated buffered risk models. Fig. 17 shows the risk signals generated by these three models on the testing sessions of the interurban scenario. In all the cases several false alarms were produced out of the time lapses were the risky situations were induced.

In the General Model, the buffered risk signal provides poor information about the risky situations. In this case, the risk signal generated by the buffered model shows almost a binary behavior. These results show that the consideration of non-homogeneous information on the training of a risk model leads to a bad performance. Thus, the mean evaluation used as driving risk ground truth was very noisy and uninformative.

Regarding the LR and SVM models, they do not provide specific information about each of the four induced risky situations. Here, all the risky situations have been assessed with the same risk level regardless of the driving performance and of the driver workload.

Summed up briefly, in the context of the interurban scenario, the proposed methodology has been able to detect and to characterize different homogeneous groups of experts when generating a ground truth. Alternative methods have provided poor information about the detected risk situations generating several false alarms.



(a) GeneralModel (b) LR Model (c) SVM Model

Fig. 17. Risk buffers generated in the testing step for the Interurban experiment.

6. Conclusions

This paper has introduced a novel methodology for the selection of subjective sequential data for the training of ES. The methodology is based on the arrangement of homogeneous information acquired from a group of human experts. It has been proposed two different similarity measures between linearized sequential data, and a novel similarity measure that uses their combination using cluster information.

Several experiments have been achieved in order to illustrate the performance of the methodology. Three of the most representative ones have been included. The first example uses synthetic data to present the methodology. The methodology has been applied to a practical case of the ITS domain where an ES for driving risk prediction has been trained and evaluated through risk evaluations acquired from a group of traffic safety experts. An experiment has been focused on an urban scenario, and another experiment makes use of data collected from an interurban scenario.

The obtained results from these experiments have shown the relevance of selecting homogeneous information for the generation of a reliable ground truth. Also, it could be concluded that the ES trained with homogeneous evaluations performed better when predicting the driving risk. Moreover, these results show the relevance of the use of subjective sequential data when dealing with phenomena where a set of rules could not be easily acquired from human experts, such as risk assessment. In this case, the rules have been properly learned from a set of homogeneous evaluations arranged with the presented methodology obtaining outstanding results.

Notice that this work will be extended with future enhancements. In order to automatize the clustering process, the optimal number of clusters could be calculated using a combination of validation indices (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). Alternative similarity measures could be defined regarding the specific domain. ITS and specifically driving risk prediction have shown an adequate environments for the proposed methodology. Nevertheless, other domains will be considered in order to illustrate the general viability of the approach. For instance, the proposed methodology could be used in Quality of Experience domain. In this case, the quality of a network service is evaluated from a set of users. The opinion of the users is used to train a Machine Learning model. Thus, a previous arrangement of users is mandatory in order to build proper models.

Acknowledgments

Research supported by grants from the Spanish Ministry of Economy and Competitiveness, under the Retos-Colaboración program: UNIKO (Ref: RTC-2015-3521-7) and PPI (Ref: RTC-2015-3580-7). Supported by the Minister for Science and Innovation of Spain: CABINTEC (PSE-37010-2007-2) and VULCANO (TEC2009- 10639-C04-04). The authors would like to thank to the Cátedra de Ecotransporte, Tecnología y Movilidad and to the Urban Mobility Master of the Rey Juan Carlos University. Special thanks to RACE, CEIT and CTAG colleagues for its support. Thanks to CONACYT and CONCYTEY from México for supporting the project through their scholarship programs.

References

Agarwal, M., & Goel, S. (2014). Expert system and it's requirement engineering process. In *Recent advances and innovations in engineering (ICRAIE), 2014* (pp. 1–4). IEEE.

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer Publishing Company, Incorporated.

Alam, M., Ferreira, J., & Fonseca, J. (2016). Introduction to intelligent transportation systems. In *Intelligent transportation systems* (pp. 1–17). Springer.

Attwell, K., Leask, J., Meyer, S. B., Rokkas, P., & Ward, P. (2017). Vaccine rejecting parents engagement with expert systems that inform vaccination programs. *Journal of Bioethical Inquiry, 14*(1), 65–76.

Bone, S. A., & Mowen, J. C. (2006). Identifying the traits of aggressive and distracted drivers: A hierarchical trait model approach. *Journal of Consumer Behaviour, 5*(5), 454–464.

Brazalez, A., Ares, J., & Matey, L. (2006). Driving simulators: Past present and future. *Euromech colloquium 476, real-time simulation and virtual reality applications of multibody systems*.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in medicine, 20*(6), 825–840.

CETINIA, U. D. S. L. (2018). Master data science. http://www.masterdatascience.es/. [Online: accessed 01-Apr-2018].

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust package for determining the best number of clusters. R package version, 2(3).

Cheng, S. Y., Park, S., & Trivedi, M. M. (2007). Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis. *Computer Vision and Image Understanding, 106*(2–3), 245–257.

Cork, R. C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., & Alexander, L. (2004). A comparison of the verbal rating scale and the visual analog scale for pain assessment. *The Internet Journal of Anesthesiology, 8*(1), 23–38.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review, 24*(2), 227–245.

Daza, I. G., Hernández, N., Bergasa, L. M., Parra, I., Yebes, J. J., Gavilán, M., et al. (2011). Drowsiness monitoring based on driver and driving data fusion. In *Intelligent transportation systems (ITSC), 2011 14th international ieee conference on* (pp. 1199–1204). IEEE.

Dia, H. (2002). An agent-based approach to modelling driver route choice behaviour under the influence of real-time information. *Transportation Research Part C: Emerging Technologies, 10*(5–6), 331–349.

de Diego, I. M., Crespo, R., Siordia, O. S., Conde, C., & Cabello, E. (2011). Automatic driving risk detection based on hands activity. In *Intelligent transportation systems (ITSC), 2011 14th international ieee conference on* (pp. 1033–1038). IEEE.

de Diego, I. M., Muñoz, A., & Moguerza, J. M. (2010). Methods for the combination of kernel matrices within a support vector framework. *Machine learning, 78*(1–2), 137.

de Diego, I. M., Siordia, O. S., Conde, C., & Cabello, E. (2011). Combining experts knowledge for driving risks recognition. In *Vehicular electronics and safety (ICVES), 2011 ieee international conference on* (pp. 59–64). IEEE.

de Diego, I. M., Siordia, O. S., Conde, C., & Cabello, E. (2012). Optimal experts' knowledge selection for intelligent driving risk detection systems. In *Intelligent vehicles symposium (IV), 2012 ieee* (pp. 896–901). IEEE.

de Diego, I. M., Siordia, O. S., Crespo, R., Conde, C., & Cabello, E. (2013). Analysis of hands activity for automatic driving risk detection. *Transportation Research Part C: Emerging Technologies, 26*, 380–395.

Djamal, B., Ernesto, D., Grosky, W., Abdelkader, H., Amit, S., Wagner, R., et al. (2017). Database and expert systems applications. *Lecture Notes in Computer Science, 10438*.

Fastenmeier, W., & Gstalter, H. (2007). Driving task analysis as a tool in traffic safety research and practice. *Safety Science, 45*(9), 952–979.

Gaines, B. R. (1987). An overview of knowledge-acquisition and transfer. *International Journal of Man-Machine Studies, 26*(4), 453–472.

Greenberg, H. (1959). An analysis of traffic flow. *Operations Research, 7*(1), 79–85.

Guerraz, A., Privault, C., Goutte, C., Gaussier, E., Pacull, F., & Renders, J.-M. (2010). Hierarchical clustering with real-time updating. US Patent 7,720,848.

Healey, J. (2011). Recording affect in the field: towards methods and metrics for improving ground truth labels. In *International conference on affective computing and intelligent interaction* (pp. 107–116). Springer.

Hodson, R. F. (2018). *Real-time expert systems computer architecture*. CRC Press.

Hua, J. (2008). Study on knowledge acquisition techniques. In *Intelligent information technology application, 2008. IITA'08. second international symposium on: 1* (pp. 181–185). IEEE.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*: 344. John Wiley & Sons.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. In *Data mining in time series databases* (pp. 1–21). World Scientific.

Kircher, K., & Ahlström, C. (2009). Issues related to the driver distraction detection algorithm attend. *First international conference on driver distraction and inattention. Gothenburg, Sweden*.

Lachaud, J.-O., Vialard, A., & De Vieilleville, F. (2005). Analysis and comparative evaluation of discrete tangent estimators. In *International conference on discrete geometry for computer imagery* (pp. 240–251). Springer.

Lin, Y., & Zhang, Y. (2012). Credit risk assessment based on neural network. In *Natural computation (ICNC), 2012 eighth international conference on* (pp. 402–404). IEEE.

Liou, Y. I., & Nunamaker, J. F. (1990). Using a group decision support system environment for knowledge acquisition: A field study. In *System sciences, 1990., proceedings of the twenty-third annual hawaii international conference on: 3* (pp. 40–49). IEEE.

MacAdam, C. C. (1981). Application of an optimal preview control for simulation of closed-loop automobile driving. *IEEE Transactions on Systems, Man, and Cybernetics, 11*(6), 393–399.

Malta, L., Miyajima, C., Kitaoka, N., & Takeda, K. (2011). Analysis of real-world driver's frustration. *IEEE Transactions on Intelligent Transportation Systems, 12*(1), 109–118.

Menard, S. (2018). *Applied logistic regression analysis*: 106. SAGE publications.

Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. *Expert Systems with Applications, 72*, 335–343.

Prelec, D. (2004). A bayesian truth serum for subjective data. *science, 306*(5695), 462–466.

Rakha, H., El-Shawarby, I., & Setti, J. R. (2007). Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger. *IEEE Transactions on Intelligent Transportation Systems, 8*(4), 630–640.

Ranney, T. A., Mazzae, E., Garrott, R., & Goodman, M. J. (2000). NHTSA driver distraction research: Past, present, and future. *Driver distraction internet forum*: 2000.

Sathyanarayana, A., Boyraz, P., & Hansen, J. H. (2008). Driver behavior analysis and route recognition by hidden markov models. In *Vehicular electronics and safety, 2008. ICVES 2008. IEEE international conference on* (pp. 276–281). IEEE.

Schneider, M., & Kiesler, S. (2005). Calling while driving: effects of providing remote traffic context. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 561–569). ACM.

Simons-Morton, B., Lerner, N., & Singer, J. (2005). The observed effects of teenage passengers on the risky driving behavior of teenage drivers. *Accident Analysis & Prevention, 37*(6), 973–982.

Siordia, O. S., de Diego, I. M., Conde, C., & Cabello, E. (2011a). Combining traffic safety knowledge for driving risk detection. In *Intelligent transportation systems (ITSC), 2011 14th international IEEE conference on* (pp. 564–569). IEEE.

Siordia, O. S., de Diego, I. M., Conde, C., & Cabello, E. (2011b). Section-wise similarities for clustering and outlier detection of subjective sequential data. In *International workshop on similarity-based pattern recognition* (pp. 61–76). Springer.

Siordia, O. S., de Diego, I. M., Conde, C., & Cabello, E. (2012). Accident reproduction system for the identification of human factors involved on traffic accidents. In *Intelligent vehicles symposium (IV), 2012 IEEE* (pp. 987–992). IEEE.

Siordia, O. S., de Diego, I. M., Conde, C., & Cabello, E. (2014). Subjective traffic safety experts' knowledge for driving-risk definition. *IEEE Transactions on Intelligent Transportation Systems, 15*(4), 1823–1834.

Siordia, O. S., de Diego, I. M., Conde, C., Reyes, G., & Cabello, E. (2010). Driving risk classification based on experts evaluation. In *Intelligent vehicles symposium (IV), 2010 IEEE* (pp. 1098–1103). IEEE.

Sjöberg, L., Moen, B.-E., & Rundmo, T. (2004). Explaining risk perception. *An Evaluation of the Psychometric Paradigm in Risk Perception Research, 10*(2). 665–612

Sriram, S., & Yuan, X. (2012). An enhanced approach for classifying emotions using customized decision tree algorithm. In *Southeastcon, 2012 proceedings of IEEE* (pp. 1–6). IEEE.

Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y., Schuurmans, G. D., et al. (2004). Support vector machine learning for interdependent and structured output spaces. In *Twenty-first international conference on machine learning (ICML 2004)* (pp. 1–8). AAAI Press.

Turban, E. (1991). Managing knowledge acquisition from multiple experts. In *Developing and managing expert system programs, 1991., proceedings of the IEEE/ACM international conference on* (pp. 129–138). IEEE.

Van Do, T., Le Thi, H. A., & Nguyen, N. T. (2018). *Advanced computational methods for knowledge engineering*. Springer.

Wagner, W. P. (2017). Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert Systems with Applications, 76*, 85–96.

Wang, H.-C., Lee, C.-S., & Ho, T.-H. (2007). Combining subjective and objective qos factors for personalized web service selection. *Expert Systems with Applications, 32*(2), 571–584.

Wang, J., Zhu, S., & Gong, Y. (2010). Driving safety monitoring using semisupervised learning on time series data. *IEEE Transactions on Intelligent Transportation Systems, 11*(3), 728–737.

Wang, J.-S., Knipling, R. R., Goodman, M. J., et al. (1996). The role of driver inattention in crashes: New statistics from the 1995 crashworthiness data system. In *40th annual proceedings of the association for the advancement of automotive medicine: 377* (p. 392).

Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications, 69*, 29–39.

WenBin, C., XiaoLing, L., YiJun, L., & Yu, F. (2010). A machine learning algorithm for expert system based on mycin model. In *Computer engineering and technology (ICCET), 2010 2nd international conference on: 2* (pp. V2–262). IEEE.

Wick, M. R., & Slagle, J. R. (1989). An explanation facility for today's expert systems. *IEEE Expert: Intelligent Systems and Their Applications, 4*(1), 26–36.

Yau, C., & Sattar, A. (1994). Developing expert system with soft systems concept. In *Expert systems for development, 1994., proceedings of international conference on* (pp. 79–84). IEEE.

Zakaria, M., Abdel-Moneim, T., Abdin, H., Hafez, A. E.-D., & Darwish, S. (2017). Optimization and mechanical simulation of a pursuit-evader scenario using genetic algorithm and stewart platform. In *Mechanical and aerospace engineering (ICMAE), 2017 8th international conference on* (pp. 314–319). IEEE.

Zhang, H., Schreiner, C., Zhang, K., & Torkkola, K. (2007). Naturalistic use of cell phones in driving and context-based user assistance. In *Proceedings of the 9th international conference on human computer interaction with mobile devices and services* (pp. 273–276). ACM.

Zhu, Y., Wu, D., & Li, S. (2007). A piecewise linear representation method of time series based on feature points. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 1066–1072). Springer.