Intelligent Decision Support to Determine the Best Sensory Guardrail Locations

Noelia Rico, Irene Díaz, José R. Villar*, Enrique de la Cal

Computer Science Department, University of Oviedo, Oviedo, Spain

Abstract

Introducing intelligent safety devices in the roads would lead to enhance both the time of reaction and the traffic safety. Nevertheless, these intelligent devices are expensive, so choosing their location should be done carefully. This research is focused a decision support system to decide the placement of a specific safety device designed in a research project. This approach includes a feature selection stage, a model learning stage and the deployment stage. Decision models learn from real datasets with information related with accidents, classifying the samples as Fatal, Severe or Slight injury. Also, a case based risk index is proposed, so samples within the same label can be sorted. Therefore, in the deployment stage, each possible location is ranked and the user gets a feedback of the suitability of each of them to be considered for placing the intelligent safety device. The experimentation shows the proposal is valid provided the dataset for training includes enough granularity. However, it is shown that specific risk index should be designed for each road type and fork.

Keywords: Road Safety, Machine Learning, Sensory Guardrails

1. Introduction

This study focuses on the design of a decision support tool for the assessment where safety intelligent road barriers should be located. These intelli-

*Corresponding author: José R. Villar, villarjose@uniovi.es

Email addresses: UO230790@uniovi.es (Noelia Rico), sirene@uniovi.es (Irene Díaz), villarjose@uniovi.es (José R. Villar*), delacal@uniovi.es (Enrique de la Cal)

Figure 1: An impact attenuator. These type of guardrail is placed at diversions in relatively high speed roads and motorways. *Image source: http://www.hiasa.com/es/cargarFichaProducto.do?identificador=79*

gent road barriers include sensory systems, on-crash video recording among them, so they are complex and expensive; its location must be decided with some sensibility. The focused road barriers -Fig. 1- are specifically designed for intersections and diversions, and are usually known as impact attenuators. Nevertheless, the solution to the problem was focused on a more generic way, aiming to develop a tool that could propose the most interesting spots for each type of road geometry.

The design of roads and the location of safety devices are two important factors in reducing the accident rate in the network. As stated in [1, 2], a safety focused road design leads to better performances in terms of traffic security without high impact in the total budget. Concerning the safety devices, the problem of deciding the correct spot to locate them is two-fold: there are decisions to be made at design time and at the road exploiting time [3, 4], when accident blackspots are found. In this latter case, the decision is made based on the available traffic data; therefore, the examples of traffic and crash data should be ranked and, somehow, a decision must be draw. For sure, all these interventions and the decisions need to be evaluated [5, 6].

Ranking indexes have been widely use either in the design of guardrails and their evaluation [7], in the design of road segments, the identification of accident blackspots [8] or in the selection of road safety decisions. Some studies focus on the maintenance point of view, that is, the data from the road maintenance companies; some others based their studies on crash experimental data, while other studies focuses on the data available from the National Traffic Agency (NTA). Therefore, the ranking indexes vary enormously according to the specific sub-field of interest. For instance, data coming from crash experiments include physical variables, which are the basis for the in-

dexes [7]: the occupant impact velocity, the acceleration severity index or the maximum change in the vehicle velocity, among others. These indexes could be reused in other areas if there is mapping from other measures to those physical variables. As an example, in those cases where the driver was able to break marks on the asphalt can be used to estimate those physical variables. However, these marks are not always measurable; therefore, this type of indexes cannot be used as general measurements of the severity of an accident.

The research on suitable ranking indexes shows that the most suitable rank measurement directly relies on the available data. And each collection may or not include a severity index. For instance, the KABCO injury severity scale is used by the police in the United States to classify each accident as one of the five defined labels since 1966 [9]. However, the interpretation of each label varies from one state to another [10]. Different indexes based on the accident records were proposed in [5]. Chen et al. [11] proposed the Road Safety Risk Index as a merged index of qualitative and quantitative variables, including data from either the travellers, the roads, the vehicles, the environment, the traffic fines, and the traffic accidents. These variables were aggregated to obtain a single value that somehow reflects the risk of each road; the entropy was proposed for determining the weights of each aggregation. Therefore, there are almost as many risk indexes different approaches as contributions to the literature; each of them is valid for the specific focused problem.

Data coming from NTA have been analyzed in several studies. For instance, [12] proposed to cluster the data from the Indian NTA and to use Association Rule Mining. This study shows the typical rules that can be extracted due to the inherent categorical type of the provided variables and its relatively reduced granularity and cardinality. What the studies in the literature clearly remark is the need of more detailed data, so better data mining can be performed on the given collections [8, 12]. Additionally, the merge of datasets coming from the NTA with the datasets coming from other sources -like the road maintenance agencies- can improve the benefits of applying these methods. Similar studies can be analysed from the literature [13, 14].

Al-Badairi et al analysed the relationships between the input features of the dataset from the Oregon Department of Traffic and the casualties of the big truck crashes in run-off the road accidents [15]. They proposed ordered probit models to try to find out the relevant knowledge. Basically, the main

factors were mainly due to the familiarity of the drivers with the network, but no real hint concerning safety devices were found. Similarly, Mussone et al studied the relationships among the variables in the urban traffic in the city of Turin (Italy) [16] using traffic and weather data. After a pre-processing and feature selection process based on correlations and SOM clustering, PCA was performed to extract up to 8 features to a cumulative representation of 93%; the most linked original features were then chosen as the input variables for the second stage. This second step made use of Artificial Neural Networks to predict the level of severity of an accident.

From both analysis -design of ranking indexes and learning and deploying the decision models-, it is clear that the most important thing is to obtain high quality data. As stated in [17], 'The most serious data quality issues appear to be: inaccuracies in crash location and time, difficulties in data linkage (e.g. with traffic data) due to inconsistencies in databases, severity misclassification, inaccuracies and incompleteness of involved users demographics and inaccurate identification of crash contributory factors.' With all these lacks in the available data, obtaining suitable decision models to locate safety devices becomes a real challenge.

This work was initially inspired in [13], where the Andalusia Region database of Susceptible Elements of Improvement were used together with the database of accidents from the Spanish Traffic National Agency (DGT [18], Dirección General de Tráfico) to extract rules in order to find relationships between the crashes and those Susceptible Elements of Improvement. The authors proposed several machine learning methods for that task. However, no results were reported, mainly because there are no correspondences between the data from both databases.

This research tackles with a decision support system to assess the location of intelligent barriers and safety devices. A method is proposed to learn models able to classify and sort the traffic locations according to the predicted risk of accidents and their severity if data is provided with the enough granularity. The study proposes a generalized solution for the problem, although it has to be adapted to the specific data available for each case. To our knowledge, this is the first hybridized solution for this type of problems.

Therefore, the safety devices -the impact attenuators among them- locations can be sorted, so an assessment in the maintenance investment can be delivered. To do so, a four stage process is detailed, including the database generation, the modeling stage, the ranking of the locations stage and the final deployment. The current implementation has been developed for the

available datasets from the DGT. Results show that, despite having datasets with a very poor granularity, the models can perform properly and that the suggested rank function fits to those constrains.

This paper is organized as follows. Next section details the four stage method, including the theoretical backgrounds and the specific adaptations to the available data. In Section 3 details the experimentation and the parameter setting, discussing the obtained results as well. Finally, conclusions are drawn.

2. Materials and Methods

This solution hybridizes two different artificial intelligent techniques. On the one hand, a C5.0 decision tree is performed to label new location candidates with the more suitable crash severity label. On the other hand, a risk index, computed using a retrieval and a reuse stages from Case Based Reasoning (CBR), is proposed to sort the location candidates as a function of the similar spots in the historical database.

Besides, the performance of the final solution completely relies on the data quality. Thus, the first subsection describes the available dataset and its main features. The design of the classifier is detailed in subsection 2.2, followed by the CBR based index computation. Finally, the integration of both solutions is explained.

2.1. A description of the DGT dataset

From the study published in [13], we contacted the DGT, which is the Spanish Agency for guaranteeing the traffic security. The DGT publishes every year a report of the accidents in the Spanish roads; these data sets are publicly available at [18]. In this work we have considered the data of accidents in Spain from 2008 to 2013 (both included), Table 1 shows the number of accidents per year.

Year	2008	2009	2010	2011	2012	2013
Number of Accidents	93161	88251	85503	83027	83115	89519

Table 1: Accidents per year, from the DGT data set.

There are 36 features included in the main DGT dataset, plus 10 features related with the vehicles involved in the crash and 26 features related with

the injuries. The main part of these features are categorical, that is, there are just a finite set of values allowed for each of them. The road location is barely identified; although there keep a field for the council, it is rarely filled: the location of the crash is identified by the road and the province in the vast majority of the examples. Of course, there are also numerical values: the number of vehicles involved or the number of fatal injuries are clear examples. Table 2 describes the most important variables.

Table 2: Description of the variables in the datasets.

Variable	**Values**
Hour	1 h periods from a day
Week day	Monday(1) to Sunday(7)
Province	52 Spanish provinces
Region	18 Spanish regions
Area	Road(1), Urban Area (2), Side Street (3),Detour (4)
Grouped area	Intercity road(1),City road (2)
Road	Road identifier
Road owner	National(1),Regional(2),Provincial(3),Municipal(4),Other(5)
Road kind	Motorway(1),Highway(2),Motored-vehicle road(3) Road with slow lane(4),Road without slow lane(5) Byway(5), Side road(7), Road fork(8), other(9)
Road elements	No available data(0), Nothing remarkable(1), Zebra-cross or island(3), Middle road island(4),Central stop lane(5), Left-turn traffic circle(6),Other (7)
Priority	No available data(0), Traffic officer(1), Traffic Lights(2), STOP sign(3), GIVE WAY sign(4), Road markings only(5), Zebra-crossing sign(6), Other signalization(7), Other(8)
Roadbed Conditions	Dry and Clean(1), Shaded(2), Damped(3), Frozen(4), Snowed(5), Muddy(6), Loose gravel(7),Oily (8), Other(9)
Road Stretch	Straight line(1), Soft curves(2), Strong curves without traffic signs(3)
Luminosity	Day light(1), Twilight(2), Night: good lighting(3), Night: bad lighting (4), Night: No lighting (5)
Traffic volume	Low(1), High(2), Congested(3), None(4)

Continued on next page

Table 2 – continued from previous page

Variable	Values
Special action	Reversible lane(1), Roadside set up(2), Other measure(3), None (4)
Sidewalk	No, Yes
Additional context	14 different contexts
Atmospheric conditions	Good weather(1), Dripping fog(2), Fog(3), Rain(4), Strong rain(5), Snowing(6), Strong wind(7)
Restricted visibility	Buildings(1), Layout of nature(2), Vegetation(3), Blinding(5), Dust or smoke(6), A different constraint(7), No restriction(8)

After analyzing the data, it was found several irrelevant variables with respect to the accident rate or the security device needed. Nevertheless, the worst result was that the dataset itself was rather incomplete and suffering from granularity. Consequently, the dataset was filtered, keeping only those features for which there was no evidence of being irrelevant.

As this research is focused on learning the severity of a location, the locations provided by variable *Road fork kind* are considered. Thus, the goal of this learning process is to study the severity for each road fork identified by the variable *Road fork kind* (See Table 3).

Road fork type
T or Y shape
X or + shape
Acceleration lane
Diverting lane
Roundabout

Table 3: Different road forks considered in the dataset.

To analyze the severity associated to each road fork we must identify the variable associated to severity. According to [19], it is possible to classify accidents as *slight injury*, *serious injury* and *fatal injury*. Fatal injury includes the cases where death occurs in less than 30 days as a result of the accident. Serious injuries are those where either immediate or later detention

in hospital as an in-patient, was required. Data provided by DGT includes three variables related to accident severity: *Number of deaths (D), Number of serious injured (SE), Number of slightly injured (SI).*

Using these three variables it is possible to construct a new variable called *Accident severity* in the following way.

- Fatality: $D > 0$.
- Serious injury: $SE > 0$ and $D = 0$.
- Slight injury: $SI > 0$, $SE = 0$ and $D = 0$

Accident severity restricted to each road fork, which takes the values {Fatality, Serious injury, Slight injury}, is the goal of this research.

2.2. Learning the severity of a location

At a first sight, the problem seemed to be easily addressable using frequent patterns and association rule mining (ASM). Therefore, the initial stage of this phase was performing ASM on the data [20]. This method produced a fairly vast amount of rules, typical for the ASM, that need further filtering and processing. However, either no suitable method for filtering and finding meaningful rules was found, or the set of rules included only elemental ones. Consequently, an alternative based on machine learning is proposed.

The strategy followed to extract knowledge from the accident databases is based on first studying the main factors affecting accident severity at an intersection year by year. In a second step, the rules associated to each different intersection across the years are mined in order to obtain the most frequent rule sets using a voting strategy across years. The basic algorithm is described below.

```
For each road fork
    For each year from 2008 to 2013
        Return the best classifier
    End for
    Return the most frequent rule set per road fork
End for
```

2.2.1. Classifier selection

The prediction of accident severity on crosses and intersections is modeled in this problem as a Machine Learning (ML) problem. As the final goal of the work is to provide an understandable output to be used by traffic analysts, we focus our attention on approaches based on tree models, due to their performance and great interpretability. At this concern, tree based models has been chosen to build classification trees.

In general, tree-based ML models and algorithms work in the following procedure [21]. They grow a tree using forward selection using a top-down approach from root to leaves until some stopping condition is reached. At each step they find the best split according to some impurity measure. The node associated to the maximal impurity reduction is then selected. Finally, most method prune the tree back and obtain a rule per path from the root to each leaf.

Different combinations of metrics, splits, stopping conditions and pruning methods lead to different approaches. In this work, considering both interpretability and performance, C5.0 ([22]), recursive partitioning (PART) and Random Forests ([23] are selected as classifiers. The performance of these classifiers is measured in terms of the well-known measures Precision, Recall, F_1 and Accuracy [24]. Precision is the fraction of relevant examples among the retrieved instances. Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. F-score is the harmonic mean of Precision and Recall. It is a quite common measure used to weight the existing trade-off between Precision and Recall.

2.3. Measuring road hazardousness

Several different indexes have been proposed in the literature to rank the hazardousness of either roads or current driving conditions, etc. As an example, [25] proposed a ranking index of the segments of road in order to alert drivers about arising contingencies in the traffic or in the road.

The problem of designing a ranking index was analyzed in [11], finding that they are problem specific and also data specific. This means that designing a generalizable ranking index is a real challenge, which should be kept open to fit the specificity of the problem and the data available. There are several reasons for this challenge. There is no clear relationship among the different factors involved in road hazardous indexes; for instance, it has been found that there is no clear relationship between the crash frequency

and the traffic flow [26]. Furthermore, there are differences in the relationships when data from different countries are analyzed [27]. Therefore, each problem would lead to the most suited set of factors and their corresponding aggregation [28].

Examples of such indexes are those propose in [29, 13]. In [29], three different indexes are proposed to evaluate the design and maintenance strategies of roads. One index is focused on density of crashes on each road segments, a second index focuses on the number of injuries in the road sections, while a third one deals with the number of fatalities occurred in the segments. These indexes were estimated for the design of the roads; however, they can also be obtained from historical data. With the density of crashes, injuries and fatalities for each section, and the length of the segment, the three indexes are computed for each road segment and -by means of aggregation- for the complete road.

Similarly, Martin et al. [13] proposed two indexes -one referred to the number of crashes and the second referred to the number of injuries and fatalities- computed using historical data from i) 5 previous years, ii) last two years, iii) two years back and iv) last year. With all these indexes, the authors proposed a set of four rules to classify a road section as hazardous or not.

In this study a different approach is proposed considering not the historical data of the segment but from similar segments. The idea underneath is to apply similar retrieval concepts than those used in Case Based Reasoning (CBR) [30] to find those road segments that better match the current position, using these cases as the historical data for computing the hazardousness index. CBR typically includes four stages: Retrieval, Reuse, Revise and Retain. The retrieval is concerned with fining the most similar cases, assigning a similarity degree to each one. Reuse deals with the selection of the most interesting retrieved cases, even partially, to generate a new outcome. Revise makes use of reasoning to generate completely new proposals based on the retrieved and chosen information. Finally, Retain estimates when a new generated case is found worth to be included in the case base for future reuse. For the extend of this study, only the first two stages are needed.

Each segment must be described not only using the geometry design information -like the slope, cant, radious, gps location, etc.- but also including the environmental factors -tar type, surrounding nature, etc.-. All these factors must be assessed, that is, the implication of each factor in the similarity of the cases is fundamental to obtain good results. This assessment can be

implemented by means of weights, whose values can be fixed a priory [29, 13] or using any other method [11]. Actually, if variables related to the environmental factors are introduced then it may be interesting to cluster these factors and to define different set of weights to each cluster.

Unfortunately, only the DGT statistical crash data collection was available for this research. As mentioned before, this dataset has a very poor granularity, and there were no information concerning with the factors mentioned above. The granularity was given in terms of road identification, crashing council and the type of segment where the crash took place -straight segment, Y or X intersection, etc.- Therefore, for experimentation purposes, similarity of the cases were restricted to these variables, assigning them the same weight. The reason for this weight selection is the poor granularity of the DGT dataset; a candidate is compared to all the accident cases for the same road, council and segment type. Nevertheless, these variables do not tell much about the similarities in the context of these spots, so there is no much reason to think a better set of weights can be set.

Once the similar segments have been retrieved, then they are reused. Reusing cases means considering the factors that are found relevant in the severity of a crash. These factors are the number of fatalities (fn), the number of severe injured individuals (si), the number of lightly injured individuals (li) and the number of involved vehicles (iv); all these factors are integer numbers. All of them have to be aggregated in order to obtain a single scalar ranking index, so we propose to scale each factor to the interval $[0.0, 1.0]$ and then to aggregated them using a weighted sum.

Each factor is lineally scaled from 0 to an upper limit, corresponding from 0.0 to 1.0. However, the factors have no real upper limit, so the scaled factors may surpass the 1.0 scaled limit. To illustrate this problem, let's focus on the number of fatalities; let's say the upper limit is 2, so this factor is scaled 0 to 2 to the interval [0.0, 1.0]. Sadly, in an accident there can be more than two fatalities, meaning the scaled factor surpasses the value of 1.0.

The weights, for this research, have been manually chosen according to the relevance we think each one has. For instance, the number of fatalities has the higher weight ($w_{fn} = 0.4$), then the number of severe injuries ($w_{si} = 0.3$); the remaining weights are $w_{li} = 0.1$ and $w_{iv} = 0.2$ for the light injuries and the number of involved vehicles, respectively. The weight selection must be carefully defined according to the problem to solve and the data quality.

The index is computed in three steps. Firstly, compute the risk index for each of the retrieved cases; let NC be the number of retrieved cases. Each

case includes four mentioned factors. These factors represent the information the dataset from the DGT includes for each crash sample. Each of these feature values is mapped to the $[0.0, 1.0]$ interval using linear functions, a 0.0 in each feature value was assigned to a 0.0 in the mapping. An upper limit (UL) was fixed to each feature, so this UL was mapped to a 1.0. The UL for fn, si, li and iv were set to $UL_{fn} = 2$, $UL_{si} = 2$, $UL_{li} = 5$ and $UL_{iv} = 3.0$. In the case of iv an extra linearity segment was introduced, so a value of 2.0 for iv was mapped to 0.75. Finally, an aggregation of the scaled factors is obtained as a weighted sum with a-priori fixed weights.

Secondly, the maximum (MAX) and the mean ((MN) values of this index among all the retrieved cases should be calculated. Thirdly, in order to shift the final index to the worst scenario, the average of MAX and MN is powered to $\frac{1}{NC}$. This last computed value is the risk index of the current sample based on the most similar cases in the database.

2.4. Ranking the locations and deployment

Finally, the two approaches need to be merged so the analyzed locations could be sorted. In this approach we perform the classification and the ranking independently. Then, the calculated risk index is assigned to each classified sample. Finally, the examples are sorted by the class severity and then by the risk index.

3. Experiments and Results

3.1. Selection of the classifier and parameter tuning

Before starting with parameter tuning, it is necessary to focus on the distribution of *Accident severity* variable. As Figure 2 shows, this variable is extremely imbalanced. In fact the number of slight injuries is almost 9 times of the total amount of serious injuries and this is 10 times the number of fatal accidents (independently on the year and road fork kind considered). That makes the problem difficult to solve so that different re-sampling as well as learning strategies have been tested.

Thus, the default configuration of the algorithms selected in Section 2.2.1 was considered to select the most suitable learning strategy among multi-category, one to all and one to one learning strategy. In addition, given that the dataset is quite unbalanced, resampling must be considered. Under-sampling, Over-sampling, SMOTE and ROSE strategies were also tested and compared to no resampling. Thus, each classifier was trained using cross

Figure 2: Box-plot graph for the different classes of accident severity

validation with k folds for each learning strategy and resampling technique, obtaining that the best combination in terms of F_1 is one to one as learning strategy and ROSE ([31]) as resampling strategy. ROSE provides a unified framework to deal simultaneously with the problem of model estimation and accuracy evaluation in imbalanced learning. It builds on the generation of new artificial examples from the classes, according to a smoothed bootstrap approach.

Once the learning (one to one) and resampling (ROSE) strategies are set, each classifier is optimized via parameter tuning. In particular, the parameters analyzed are detailed below.

- Random forests: Number of predictors sampled for splitting at each node (ranging from 1 to 6).

- C5.0. In this case several parameters have been studied: Feature selection, number of boosting iterations (trials), ranging in the set $\{1, 2, 5, 10\}$ and output model (tree or a rule set).

- Recursive partitioning: Complexity parameter (ranging from 1 to 10).

Again, the learning process is tested using cross validation with 10 folds. R version 3.3.3 as well as the caret package [32] were the tools used to perform these experiments.

Table 4 shows the method that obtains the best F-score. Each value represents the percentage of experiments the method obtained the best F-score. As it can be seen, C5.0 is the method performing better most of times. Thus, it is the one selected for predicting accident severity.

Road Fork	Serious vs Slight	Fatality vs Slight	Fatality vs Serious
T or Y	C5.0 (100%)	C5.0 (100%)	C5.0 (100%)
X or +	C5.0 (100%)	C5.0 (100%)	C5.0 (100%)
Acceleration lane	C5.0 (100%)	rf (50%) C5.0 (33%) rpart (17%)	rf (68%) C5.0 (16%) rpart (16%)
Diverting lane	C5.0 (100%)	C5.0 (50%) rf (33%) rpart (17%)	rpart (50%) rf (33%) C5.0 (17%)
Roundabout	C5.0 (100%)	C5.0 (50%) rf (33%) rpart (17%)	C5.0 (100%)

Table 4: Methods obtaining the best performance for each combination of road fork and crash severity. In brackets the percentage of times that each method was the best

3.2. Training and validation of the proposal

Thus, C5.0 is selected as base classifier to predict accident severity. According to preliminary experiments detailed in Section 3.3, the best configuration for C5.0 is winnowing and a tree based structure instead of rules. The number of boosting iterations depends on the data set. The performance of C5.0 measured in terms of F-score is shown in Table 5. F-score ranges from 0.62 to 0.8. As it can be seen, the algorithm performance is quite similar across years and road forks kinds. F-score is about 0.7 in average. Although it is not so bad, it is a challenge to improve classifier efficiency.

When studying accident severity at a certain road fork, the decision trees provided by C5.0 are different across years as the data sets are obviously different. However, it is interesting to check if there is any consistent rule to predict accident severity across the years. To do that, tree models are translated into rules, obtaining 6 different rule bases for each road fork. Thus, to obtain the final rule base, only rules occurring half of years are considered. Table 6 shows the number of rules obtained by road fork and accident severity.

Year	T or Y	X or +	Accelerating lane	Diverting lane	Roundabout
2008	0.67	0.71	0.67	0.70	0.69
2009	0.67	0.74	0.73	0.73	0.73
2010	0.68	0.74	0.70	0.62	0.72
2011	0.66	0.71	0.74	0.65	0.75
2012	0.64	0.75	0.72	0.70	0.73
2013	0.65	0.76	0.80	0.65	0.77

Table 5: F-score obtained when accident severity is predicted with C5.0

The rule base associated to accident severity in roundabouts is shown in Table 7 as example. As it can be seen, accident severity depends on priority signs, existence of sidewalks, road luminosity, traffic volume and other road properties.

Road fork type	Fatal	Serious injury	Slight injury
T or Y shape	19	2	5
X or + shape	6	4	4
Acceleration lane	11	2	2
Diverting lane	1	0	6
Roundabout	6	6	7

Table 6: Number of rules occurring during at least 3 years.

3.3. Measuring risk

As the final goal of this research is to predict where to place sensory Guardrail Locations, it is necessary to establish a priority among the different places or road profiles. Thus, once a test example is classified as *fatal*, *serious injury* or *slight injury* the risk index is computed and the examples are ranked according to it. As the risk factor is computed over the data set, we have checked the risk index for the different kinds of road forks studied in the classification problem. The obtained factor risks are shown in Figure 3 where each risk value is the average of the factor risks obtained during the different years. As it can be seen, the riskiest forks are roundabouts while acceleration lanes are the least ones. Note that this is a measure independent of the accident severity classification.

Factors for predicting Slight injuries
{Roadbed conditions != Dry and clean, Luminosity !=Twilight, Sidewalk=No, Traffic volume=Low}
{Road kind != Highway, Roadbed conditions = Dry and clean, Priority=Give Way sign, Sidewalk=No, Traffic volume=Low}
{Priority=STOP sign, Traffic volume=Low}
{Luminosity=Night: No lighting, Traffic volume=Low}
{Restricted visibility != No restriction, Traffic volume=Low}
{Road kind != Motored-vehicle road, Priority= Other signals, Traffic volume=Low}
{Priority != Give way sign OR Other signals, Grouped area= City road, Luminosity= Night: good lighting, Traffic volume=Low}
Factors for predicting Serious injuries
{Luminosity =Night: good lighting, Sidewalk=Yes, Special action=None}
{Priority=Give Way sign, Traffic volume=Low}
{Priority!=Zebra-crossing sign, Traffic volume!=Low, Sidewalk=Yes}
{Roadbed conditions != Dry and clean, Traffic volume=Low}
{Sidewalk = No, Traffic volume!=Low}
{Sidewalk = No}
Factors for predicting Fatal injuries
{Priority != GIVE WAY sign Luminosity !=Night: good lighting, Roadbed conditions = Dry and clean, Sidewalk=Yes}
{Luminosity!= Night: No lighting, Traffic volume!=Low}
{Priority != STOP or Give way signs or Zebra-crossing sign, Roadbed conditions = Dry and clean}
{Priority != STOP sign, Road kind=Highway, Roadbed conditions = Dry and clean, Traffic volume=Low, Sidewalk=No, Restricted visibility != No}
{Priority != STOP sign or or Zebra-crossing sign, Roadbed conditions = Dry and clean, Traffic volume=Low, Sidewalk=Yes, Restricted visibility != No}
{Traffic volume!=Low, Sidewalk=No}

Table 7: Factors for predicting accident severity for a X or + Road Fork

As can be seen, the proposed risk index lacks in generalization as the values obtained clearly differs from one type of crossing to another. This problem is present in any ad-hoc risk index because the same set of variables are used in computing the risk index for all type of crash.

For instance, the number of fatalities in accidents in a roundabout is higher because of the speed and the traffic density. However, the number

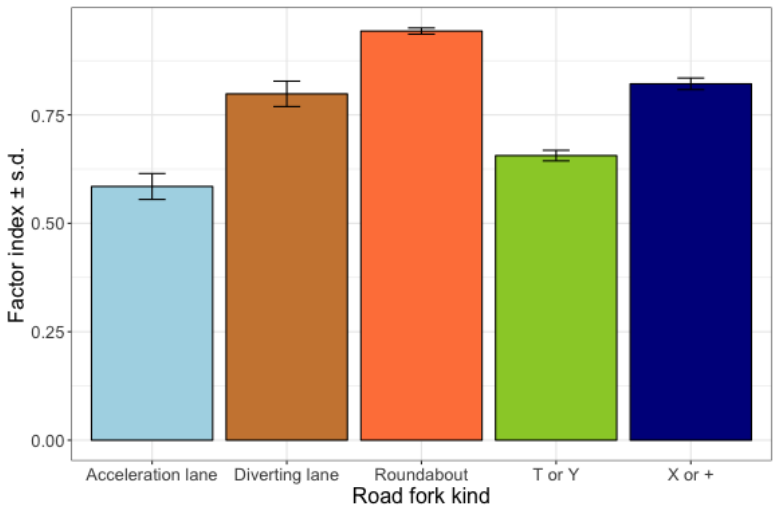

Figure 3: Factor risk by Road Fork. Each bar represents the factor risk in average associated to each road fork during the years under study (2008 to 2013)

of fatalities when diverting of a motorway is smaller because the font-lateral crashes might produce more injuries and less fatalities.

The obtained results suggest that a different risk index should be developed for each type of fork and road. Therefore, more research is needed in order to discriminate the most relevant features involved in the risk index calculation for each type of fork and road.

4. Conclusions

In this research, a proposal of a decision support system to help in choosing the location where to place an intelligent sensory guardrail among several candidates. The decision support system suggests an order of the candidates according to a label of the accident severity and a risk factor. The intelligent model use to classify the candidates is learned based on historical data from a accident database. Moreover, the risk index is also obtained based on similarly retrieved cases within the historical data, in a case-based reasoning fashion. The experimentation carried out in this research made use the public accident dataset published by the DGT. This dataset includes plenty of nominal and discrete features. A study of the relationships and the most interesting features has been performed before the process of learning models for the classification.

A comparison of the performance of different models, C5.0 has been found the most interesting and robust model for each type of crossing type. Together with the risk index, the proposal has been found interesting and with a good performance on the data. However, one of the main problems found is the high data granularity. For instance, the risk index might be in compromise due to this reason: the available dataset does not include information about the kilometer, so the similarity is not as precise as it would be needed.

Nevertheless, the whole solution allows the user to sort the candidates, producing labeled candidates together with the available risk index. If better data is provided, the procedure described in this research would allow to obtain finest models and a more precise risk index. A careful selection of the most suitable features should be performed in order to obtain a risk index for each road and crossing types.

As future work we plan to perform a deeper study about risk factors and study if there is any relation between it and the accident severity that it is predicted by a classifier. In addition, it is also necessary to introduce more information to the system for representing forks in a more accurate way. On the other hand, feature selection techniques as well as other parameter settings (minimum number of examples per leaf or pruning level) can be deeper studied in order to improve the efficiency of the method.

Acknowledgements

Funding: This work was supported by the Innterconecta call of the European Union Structural Funds (FEDER) INTERCONECTA CDTI project ABECATIM (SOL-00082271 / ITC-20151039) and by the Spanish Government through the MINECO project TIN2017-87600-P.

References

[1] M. Cornelissen, P. Salmon, N. A. Stanton, R. McClure, Assessing the 'system' in safe systems-based road designs: using cognitive work analysis to evaluate intersection designs, Accident Analysis & Prevention 74 (2015) 324–338.

[2] D. J. Gabauer, X. Li, Influence of horizontally curved roadway section characteristics on motorcycle-to-barrier crash frequency, Accident Analysis & Prevention 77 (2015) 105–112.

[3] J. Strandroth, Validation of a method to evaluate future impact of road safety interventions, a comparison between fatal passenger car crashes in sweden 2000 and 2010, Accident Analysis & Prevention 76 (2015) 133–140.

[4] Q. Yu, Z. Guo, Z. Zhang, J. Wang, Assistant decision-making system for road safety strategy, Procedia-Social and Behavioral Sciences 96 (2013) 320–328.

[5] O. Basile, L. Persia, Tools for assessing the safety impact of interventions on road safety, Procedia-Social and Behavioral Sciences 53 (2012) 682–691.

[6] C. Roque, F. Moura, J. L. Cardoso, Detecting unforgiving roadside contributors through the severity analysis of ran-off-road crashes, Accident Analysis & Prevention 80 (2015) 262–273.

[7] D. J. Gabauer, H. C. Gabler, Comparison of roadside crash injury metrics using event data recorders, Accident Analysis & Prevention 40 (2008) 548 – 558.

[8] P. K. Agarwal, P. K. Patil, R. Mehar, A methodology for ranking road safety hazardous locations using analytical hierarchy process, Procedia-Social and Behavioral Sciences 104 (2013) 1030–1037.

[9] B. Burdett, Improving Accuracy of KABCO Injury Severity Assessment by Law Enforcement Officers, Master's thesis, Civil and Environmental Engineering School, University of Wisconsin-Madison, 2014.

[10] F. H. A. (FHWA), Kabco injury classification scale and definitions, `https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf`, 2017.

[11] F. Chen, J. Wang, Y. Deng, Road safety risk evaluation by means of improved entropy topsis–rsr, Safety science 79 (2015) 39–54.

[12] S. Kumar, D. Toshniwal, Analysing road accident data using association rule mining, in: Computing, Communication and Security (ICCCS), 2015 International Conference on, IEEE, pp. 1–6.

[13] L. Martín, L. Baena, L. Garach, G. López, J. de Oña, Using data mining techniques to road safety improvement in spanish roads, Procedia-Social and Behavioral Sciences 160 (2014) 607–614.

[14] S. Kumar, D. Toshniwal, A data mining framework to analyze road accident data, Journal of Big Data 2 (2015) 26.

[15] N. S. S. Al-Bdairi, S. Hernandez, An empirical analysis of run-off-road injury severity crashes involving large trucks, Accident Analysis & Prevention 102 (2017) 93–100.

[16] L. Mussone, M. Bassani, P. Masci, Analysis of factors affecting the severity of crashes in urban road intersections, Accident Analysis & Prevention 103 (2017) 112–122.

[17] M. Imprialou, M. Quddus, Crash data quality for road safety research: current state and future directions, Accident Analysis & Prevention (2017).

[18] D. G. de Tráfico, Dirección general de tráfico, portal estadístico, accidentes 2015, `https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/`, 2017.

[19] P. Michalaki, M. A. Quddus, D. Pitfield, A. Huetson, Exploring the factors affecting motorway accident severity in england using the generalised ordered logistic regression model, Journal of Safety Research 55 (2015) 89 – 97.

[20] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases VLDB '94, pp. 487–499.

[21] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[22] R. J. Quinlan, Data Mining Tools See5 and C5.0, 2000.

[23] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[24] I. Díaz, J. Ranilla, E. Montañés, J. Fernández, E. F. Combarro, Improving performance of text categorisation by combining filtering and support vector machines, Journal of the American Society for Information Science and Technology 55 (2004) 579–592.

[25] V. Rosolino, I. Teresa, A. Vittorio, F. D. Carmine, T. Antonio, R. Daniele, Z. Claudio, Road safety performance assessment: a new road network risk index for info mobility, Procedia-Social and Behavioral Sciences 111 (2014) 624–633.

[26] C. Roque, J. L. Cardoso, Investigating the relationship between run-off-the-road crash frequency and traffic flow through different functional forms, Accident Analysis & Prevention 63 (2014) 121–132.

[27] C. Roque, J. L. Cardoso, Safeside: a computer-aided procedure for integrating benefits and costs in roadside safety intervention decision making, Safety science 74 (2015) 195–205.

[28] F. Russo, M. Busiello, G. Dell'Acqua, Safety performance functions for crash severity on undivided rural roads, Accident Analysis & Prevention 93 (2016) 75–91.

[29] K. Jamroz, M. Budzyński, W. Kustra, L. Michalski, S. Gaca, Tools for road infrastructure safety management–polish experiences, Transportation Research Procedia 3 (2014) 730–739.

[30] J. Kolodner, Case-based reasoning, Morgan Kaufmann, 2014.

[31] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Mining and Knowledge Discovery 28 (2014) 92–122.

[32] M. Kuhn, The caret package, `http://topepo.github.io/caret/index.html`, 2017.