Driving Behavior Safety Levels: Classification and Evaluation

Kui Yang
Department of Civil, Geo and Environmental Engineering, Technical University of Munich
Munich, Germany
kui.yang@tum.de

Christelle Al Haddad
Department of Civil, Geo and Environmental Engineering, Technical University of Munich
Munich, Germany
christelle.haddad@tum.de

George Yannis
School of Civil Engineering National Technical University of Athens (NTUA)
Athens, Greece
geyannis@central.ntua.gr

Constantinos Antoniou
Department of Civil, Geo and Environmental Engineering, Technical University of Munich
Munich, Germany
c.antoniou@tum.de

***Abstract*— Driving simulator and naturalistic driving studies are often used to understand driving behavior characteristics. It is essential to evaluate the traffic safety of driving behavior in real time, which is helpful to trigger interventions of Advanced Driver Assistance Systems (ADAS) to ensure the driving safety. Therefore, this paper aims to propose a framework of driving behavior safety level classification and evaluation in real time, which was validated by a case study based on a driving simulation experiment. The proposed methodology focuses on finding the optimal number of safety "levels" or "zones" for driving behavior, classifying the safety levels with the help of different clustering techniques, and evaluating the driving safety levels based on developed classification models in real-time. Three clustering techniques were applied, including k-means clustering, hierarchical clustering and model-based clustering. The optimal number of clusters was found to be four using k-means), and the clusters of safety levels will be labelled as "normal" driving, "low risk" driving, "middle risk" driving and "high risk" driving. A Support Vector Machine (SVM) and a decision tree were thereafter developed as the classification model. The accuracy of the combination of model-based clusters and SVM models proved to be the best with four clusters, yet no significant difference to other models was found.**

Keywords—driving behavior safety levels, driving simulation, clustering, SVM, decision trees

I. Introduction

Road safety improvement is one of the goals introduced by EU in its "Zero Vision" [1]. Specifically, the aim is to cut European road fatalities and serious injuries down to zero. Following this aim, driving simulator studies and naturalistic driving studies are designed to better understand drivers' behavior, as it is one of the main factors impacting road safety. With advanced technologies set to improve data collection, driving behavior is postulated to belong to one or more safety levels or zones, ranging from "normal" to "dangerous" driving. The goal of this paper is to propose a real-time classification and evaluation framework of driving behavior safety levels, which was validated by a case study based on a driving simulation experiment. Three clustering algorithms are proposed including k-means clustering [2], hierarchical clustering [3] and a model-based clustering [4], and the optimal number of clusters for each is also identified. The obtained clusters can be well visualized using advanced machine learning algorithms such as T-distributed Stochastic Neighbor Embedding (t-SNE) [5]. Support Vector Machines (SVM) and decision trees are then used to develop models for real-time safety level classification for new observations. The contents of the paper will be structured as follows. First, the overall methodology will be introduced, including the overall framework, and the formulation of the used clustering and classification (modeling) algorithms. Then the simulation design environment, and the data collection and variables of interest are presented. Afterwards, the results are described and analyzed. Finally, a conclusion is given, focusing on the main findings but also limitations and future work needed.

II. Methodology

A. Overall Framework

After improving the methodology presented by [6], this paper outlines the overall framework in Fig. 1 . It includes the main methodological components along with the information flow. Generally, each observation may hold multiple attributes, such as driving speed, headway, and lateral location in the lane.

The methodology includes training and application steps. During the training step, archived surveillance data are used to (i) find the optimal number of clusters, presenting the ideal number of driving behavior safety zones or levels; (ii) identify the various driving behavior safety levels through clustering the available observations; and (iii) estimate the transition processes between these regimes. Finally, the information is stored into a knowledge base and further supports the application of the framework. During the application step, the appropriate classification model was selected to evaluate the driving safety levels with the input of the real-time surveillance data.

In this study, three clustering algorithms including k-means clustering, hierarchical clustering and a model-based clustering, were employed to find the optimal number of clusters. These three algorithms were then used to cluster the available observations. Finally, support vector machines (SVM) and decision trees were developed with the input of the labeled datasets based on three clustering algorithms to evaluate driving behavior safety levels and further to test the performance of developed models and clustering algorithms.

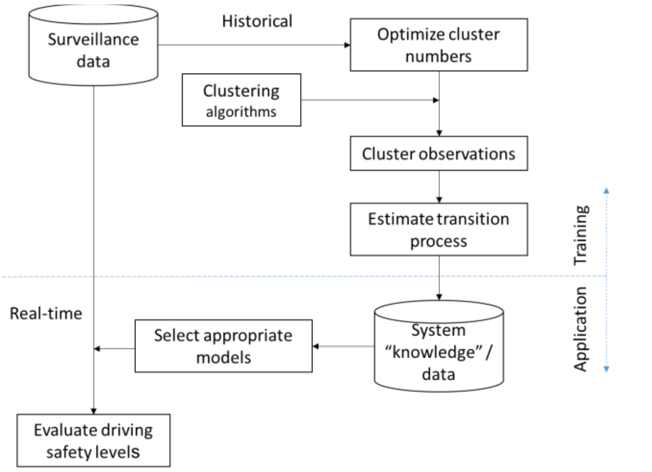

Fig. 1 Overall framework of driving behavior safety level classifications and estimations.

B. Clustering Algorithms

a) K-means clustering

K-means clustering is a popular unsupervised learning algorithms that solves the clustering problem. The Elbow method is one of the most popular methods for determining the optimal number of clusters in k-means clustering [6]. The basic idea in the k-means clustering to define clusters is that the total intra-cluster variation (known as total within-cluster variation) is minimized. The total within-cluster variation is popularly defined as the sum of squared Euclidean distances between items [7], and can be formulated as follows.

$$Total.D = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (1)$$

where x_i is a driving behavior data observation belonging to the cluster C_k, and μ_k is the mean value of the observations assigned to the cluster C_k.

The total within-cluster variation measures the compactness (i.e. goodness) of the clustering and it should be as small as possible. Each observation x_i is assigned to the closest cluster based on the Euclidean distance between the object and the centroid (Eq. 1). There are five important steps to identify the optimal cluster number. They are (i) computing the clustering algorithm for different values of k; (ii) for each k, measuring the cost of the optimal quality solution (e.g., $Total.D$); (iii) plotting the curve of $Total.D$ according to the results from (ii); (iv) the location of a bend (knee) in the plot is generally considered as the appropriate number of clusters.

b) Hierarchical clustering

Hierarchical clustering can create a hierarchy of clusters, and presents the hierarchy in a dendrogram to cluster multidimensional data sets, by evaluating dissimilarities of objects in the variables space, or similarities of variables in the objects space [8]. Some studies (i.e., [3][8]), describe in detail the hierarchical clustering methods. Hierarchical clustering adopts either an agglomerative technique, which is a series of fusions of the n objects into groups, or a divisive technique, which separates n objects successively into finer groups, to build a hierarchy of clusters. Since agglomerative techniques are more commonly used [9], they are used in this paper. Agglomerative hierarchical clustering methods are characterized by *the distance metric* and *the linkage method.*

The distance metric presents the similarity between each cluster. Euclidean distance whose equation is $d = \sum_{x_i \in C_k} (x_i - \mu_k)^2$, is used in this paper. *The linkage method* determines how the distance between two clusters is defined. Common linkage methods include single linkage, complete linkage, and ward linkage. After comparing these methods during preliminary analysis, we decided to use the complete linkage in our final analysis for hierarchical clustering since it could best fit our dataset. The complete linkage refers to the longest distance between two observations in each cluster, and its equation is $D_{12} = max_{ij}(X_i, Y_j)$ where $X_i \ and \ Y_j$ are two observations. The distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster.

c) A model-based clustering

A model-based clustering assumes a data model and applies an expectation-maximization (EM) algorithm to find the most likely model components and the number of clusters. In the model-based clustering literature, the Gaussian Mixture Model (GMM) is most commonly used, such as [4][10]. GMM attempts to optimize the fit between the observed data and some mathematical model using a probabilistic approach. First, a specific-form mixture of Gaussians is assumed, and the density of the Gaussian mixture model [4] is:

$$f(x|\theta) = \sum_{m=1}^{M} \pi_m \varphi(x|\rho_m, \Sigma_m)$$

where $\varphi(x|\rho_m, \Sigma_m)$ is the density of a multivariate Gaussian random variable $X$with mean ρ_m and covariance matrix Σ_m, and $\theta = (\pi_1, \cdots, \pi_M, \rho_1, \cdots, \rho_M, \Sigma_1, \cdots, \Sigma_M)$**.**

Second, the parameters (i.e., the mean and the standard deviation) of this model are estimated by the Expectation Maximization (EM) algorithm. EM starts with a random or heuristic initialization and then iteratively uses two steps to resolve the circularity in computation: (i) E-Step, which determines the expected probability of data assignment to clusters with the help of current model parameters. (2) M-Step, which determines the optimum model parameters of each mixture by using the assignment probabilities as weights [11].

C. Evaluation Models of Driving Behavior Safety Levels

In order to develop better evaluation models for driving behavior safety levels, support vector machines (SVM) and decision trees (DT) are used in this paper as well as the parameter fine-tune of developed models. SVM was originally designed based on statistical learning theory and structural risk minimization (e.g., [12][13]), and it will be used to classify driving behavior safety states. Among various types of SVM models, the C-support vector machine (C-SVM) was used in this study due to its most common use [13]. SVM models were developed in R® 3.5.3, using Package '*e1071*' [15]. Given training vectors $\boldsymbol{x_i} \in R^n, i = 1, \cdots, l$, in two classes and an indicator vector $\boldsymbol{y} \in R^l$ such that $y_i \in \{1, -1\}$. C-SVM [13] solves the following primal optimization problem.

$$\begin{aligned} \underset{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi}}{min} \quad & \tfrac{1}{2} w^T \boldsymbol{w} + C \textstyle\sum_{i=1}^{l} \xi_i \\ s.t \quad & y_i(\boldsymbol{w}^T \emptyset(\boldsymbol{x_i}) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \cdots, l \end{aligned} \quad (2)$$

where $\emptyset(\boldsymbol{x_i})$ maps $\boldsymbol{x_i}$ into a higher-dimensional space and $C > 0$ is the regularization parameter. Due to the possible high dimensionality of the vector variable $\boldsymbol{w}$, the following dual problem is solved [14].

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{min} \quad & \tfrac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\alpha} - \boldsymbol{e}^T \boldsymbol{\alpha} \\ s.t \quad & \boldsymbol{y}^T \boldsymbol{\alpha} = 0, \\ & 0 \leq \alpha_i \leq \mathrm{C}, \ i = 1, \cdots, l, \end{aligned} \quad (3)$$

where $\boldsymbol{e} = [1, \cdots, 1]^T$ is the vector of all ones, $\boldsymbol{Q}$ is an l by l positive semi definite matrix, $Q_{ij} \equiv y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j})$, and $K(\boldsymbol{x_i}, \boldsymbol{x_j}) \equiv \emptyset(\boldsymbol{x_i})^T \emptyset(\boldsymbol{x_j})$ is the kernel function. The optimal $\boldsymbol{w}$ satisfies $\boldsymbol{w} = \sum_{i=1}^{l} y_i \alpha_i \emptyset(\boldsymbol{x_i})$ and the decision function is

$$\mathrm{sgn}(\boldsymbol{w}^T \emptyset(\boldsymbol{x}) + b) = \mathrm{sgn}(\sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x_i}, \boldsymbol{x}) + \mathrm{b})$$

SVM models could also handle multi-classification problems. In this paper two kernel functions were considered:

a) Radial Kernel: $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \exp(-\gamma|\boldsymbol{x}_i - \boldsymbol{x}_j|^2)$

b) Linear Kernel: $K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \boldsymbol{x}_i{}^T \boldsymbol{x}_j$

A decision tree (DT) is a decision support algorithm that uses a tree-like model of decisions and their possible consequences to perform both classifications and regressions. The detailed knowledge of DT is introduced in previous papers, such as [16]. In this paper, the package 'rpart' [17] was used to develop the DT in R® 3.5.3.

III. Data and Experimental Design

The driving simulator experiment was conducted in the Department of Transportation Planning and Engineering of the School of Civil Engineering of the National Technical University of Athens (NTUA), where the FOERST Driving Simulator FPF is located. The driving simulator consists of 3 LCD wide screens 40''(fullHD), a total angle view of 170 degrees, a driving position, and a support base.

The simulated road environment is an undivided two-lane rural road, which is a single carriageway with length 2,1 km, width 3m, with zero gradient, and mild horizontal curves. In the driving simulations, two traffic scenarios (i.e., moderate traffic conditions and high traffic conditions) and three distraction conditions (i.e., no distraction, cell-phone conversation and conversation with passenger) were examined in a full factorial within-subject design. During each trial of the experiment, two unexpected incidents that are the sudden appearance of an animal (deer or donkey) on the roadway were scheduled to occur at approximately fixed points along the drive. The driving simulator provides a "Free Driving" scenario that familiarizes the participants with the demands of an everyday drive. After a familiarization drive and a necessary short brake, each participant has only one chance to drive approximately 12,6km within about 20min in total. Finally, the sample of participants is a total of 260 individuals.

TABLE I. VARIABLES FOR ANALYSIS

No	***Variables***	***Description***
1	age	Driver's age
2	LateralPosition	Average distance to the right road board (m)
3	StdevLateralPos	Standard deviation of distance to the right road board (m)
4	AverageSpeed	Average speed (km/h)
5	StdevSpeed	Standard deviation of speed (km/h)
6	RspurAverage	Average track of the vehicle from the middle of the road (m)
7	StdRspur	Standard deviation of track of the vehicle from the middle of the road (m)
8	RalphaAverage	Average direction of the vehicle compared to the road direction in degrees
9	StdRalpha	Standard deviation of direction of the vehicle compared to the road direction in degrees
10	BrakeAverage	Average brake pedal position (%)
11	StdBrake	Standard deviation of brake pedal position(%)
12	GearAverage	Average chosen gear (0 = idle, 6 = reverse)
13	StdGear	Standard deviation of chosen gear (0 = idle, 6 = reverse)
14	RpmAverage	Average motor revolutions in 1/min
15	StdRpm	Standard deviation of motor revolutions (1/min)
16	HWayAverage	Average headway, distance to the ahead driving vehicle (m)
17	StdHWay	Standard deviation of headway, distance to the ahead driving vehicle (m)
18	DleftAverage	Average distance to the left road board (m)
19	StdDleft	Standard deviation of distance to the left road board (m)
20	WheelAverage	Average steering wheel position in degrees
21	StdWheel	Standard deviation of steering wheel position in degrees
22	TheadAverage	Average time to headway, i.e. to collision with the ahead driving vehicle (s)
23	StdThead	Standard deviation of time to headway, i.e. to collision with the ahead driving vehicle (s)
24	TTLAverage	Average time to line crossing, time until the road border line is exceeded (s)
25	StdTTL	Standard deviation of time to line crossing, time until the road border line is exceeded (s)
26	TTCAverage	Average time to collision (all obstacles) (s)
27	StdTTC	Standard deviation of time to collision (s)
28	Crash_number	Number of crashes during the driving interval

The simulator records data at intervals of 33 to 50 milliseconds, including at first, 33 variables in each session. In order to explore driving behavior safety level classification and estimation, 28 variables were further aggregated and collected, including the driving characteristics in normal driving scenario and the driving characteristics prior to the lowest speed during event scenarios. The variables are listed in TABLE I. It is noted that the crash could only happen at unexpected incidents. And 0.65 crashes as an average happened for each driver during the driving simulation.

IV. Results and Analysis

A. *Optimizing Driving Behavior Safety Levels*

K-means clustering, hierarchical clustering and a model-based approach were used to identify the optimal levels of driving behavior safety. Fig. 2 illustrates the optimal number of clusters based on Elbow method. According to the location of the bend (knee) in the plot, the optimal number of driving behavior safety levels is four.

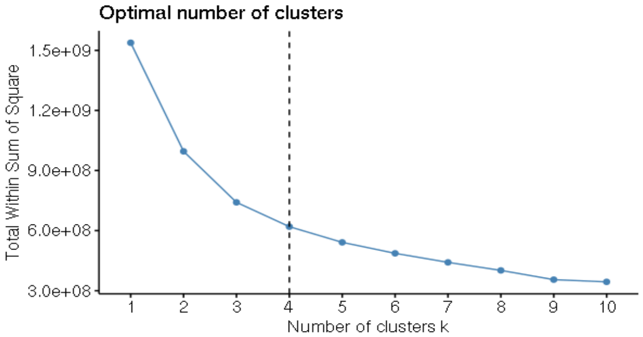

Fig. 2 Optimal number of clusters based on k-means clustering.

Fig. 3 shows the cluster dendrogram of hierarchical cluster analysis. The Ward's minimum variance method to perform agglomerative clustering. In the dendrogram, each leaf corresponds to one observation, and we can see the hierarchy of clusters. As we move up the tree, observations that are similar to each other are combined into branches. However, we can determine the number of clusters within the dendrogram and cut the dendrogram at a certain tree height to separate the data into different groups. The optimal number of state levels of driving behavior safety was found to be three. The red rectangle borders show the three clusters in Fig. 3 .

As a parametric method that uses the Gaussian distribution,

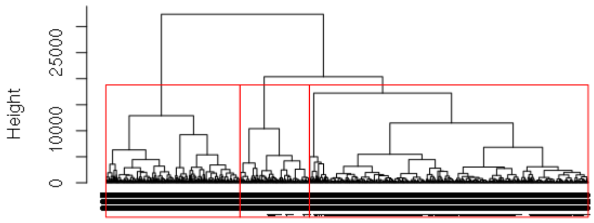

Fig. 3 Cluster dendrogram of hierarchical cluster analysis.

the Gaussian mixture model (GMM) is a widely used model-based approach. Bayesian Information Criterion (BIC) is an important index to find the number of clusters by selecting the best clustering model and it uses the likelihood and a penalty term to guard against overfitting. The bigger the BIC is, the better the number of clusters. Fig. 4 shows optimal number of clusters based on GMM. Therefore, the optimal number of driving behavior safety levels is four based on GMM, where the –BIC, which is 179507, is smallest.

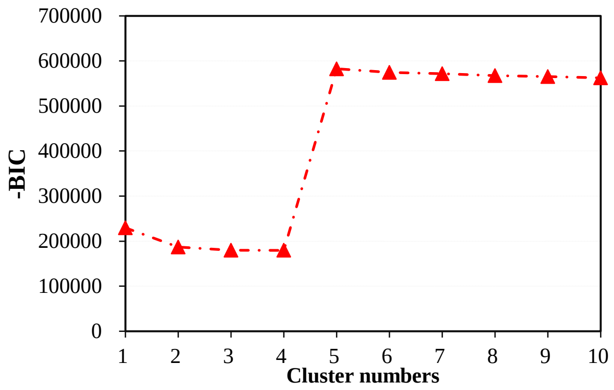

Fig. 4 Optimal number of clusters based on GMM.

B. *Clustering and Classification*

After defining the optimal driving behavior safety levels, the observations in the dataset were further clustered and classified into different clusters. K-means clustering, hierarchical clustering and a model-based approach were used for this purpose, as presented in Fig. 5 , Fig. 6 and Fig. 7 .

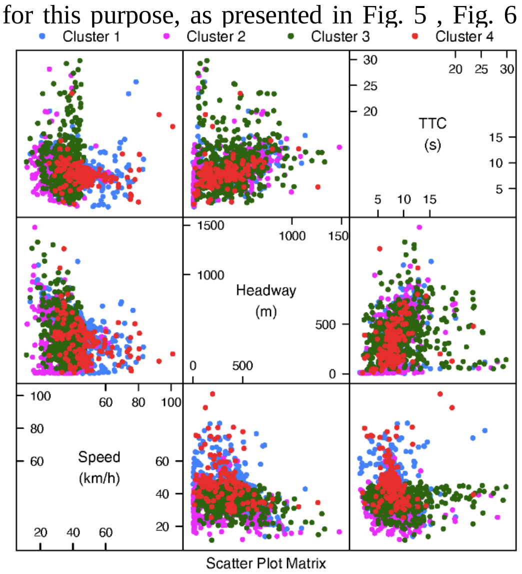

Fig. 5 Different clustering scenarios based on k-means clustering.

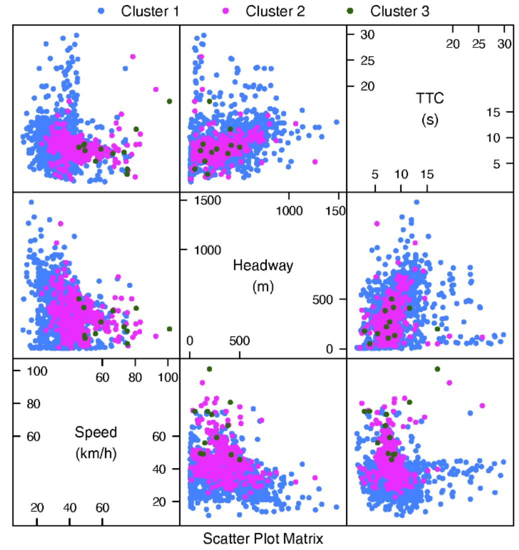

Fig. 6 Different clustering scenarios based on hierarchical clustering.

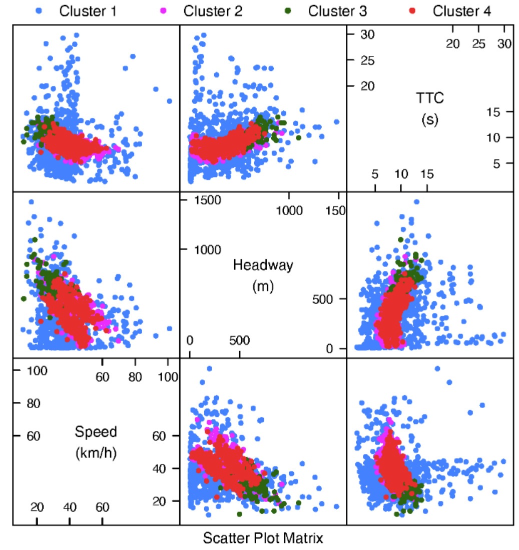

Fig. 7 Different clustering scenarios based on GMM.

Three variables, namely average speed, average headway and average TTC, were selected as examples. It is interesting to note that the resulting sets of clusters based on these three clustering algorithms have similar geometries.

T-SNE [5] was used to visualize the clustering algorithms. It is extremely useful for visualizing high-dimensional data [19], and it has a dimensionality reduction method to visualize data embedded in a lower number of dimensions, to see patterns and trends in the data. It can deal with more complex patterns of Gaussian clusters in multidimensional space compared to Principal Component Analysis. T-SNE results are shown in Fig. 8 , Fig. 9 and Fig. 10 . These visualizations show that driving behavior is well clustered into several levels.

In order to identify the best clustering algorithm, four widely used indices, i.e., the within clusters sum of squares, the average silhouette width, Dunn index and Calinski-Harabasz index, were used. The within clusters sum of squares is a measurement showing how closely related objects are in a cluster. The smaller the value, the more closely related objects are within the cluster. The average silhouette width is a measurement considering how closely related objects are within the cluster and how clusters are separated from each other. The silhouette value ranges from 0 to 1, and a value closer to 1 suggests that the data is better clustered [20]. The Dunn index [21] is an internal evaluation scheme, where the result is the ratio of minimum separation and maximum diameter for all clusters based on the clustered data itself. The higher the Dunn index value is, the better the clustering is. The Calinski-Harabasz index [22] also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters. The higher the Calinski-Harabasz index is, the better the performance is. The results are listed in TABLE II.

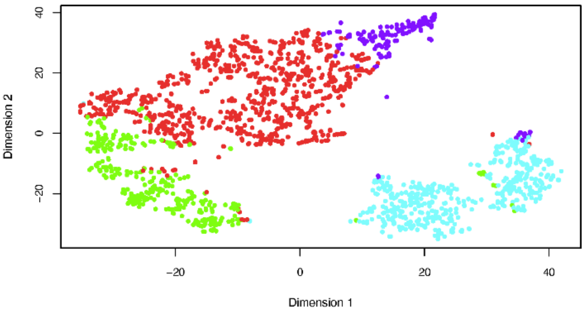

Fig. 8 Visualization of clustering results based on k-means: elbow method.

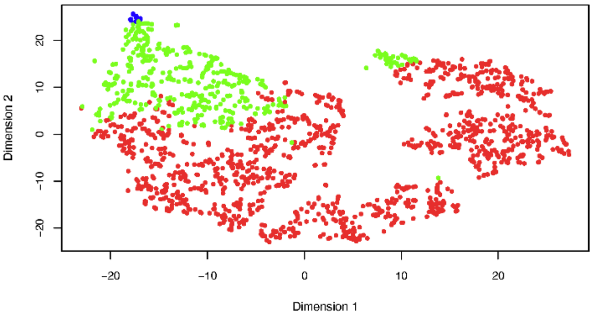

Fig. 9 Visualization of clustering results based on hierarchical clustering.

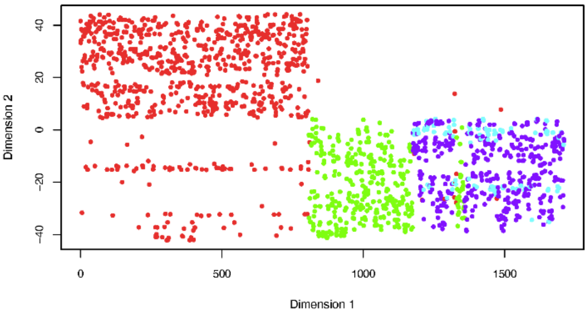

Fig. 10 Visualization of clustering results based on GMM.

TABLE II. COMPARING THREE CLUSTERING ALGORITHMS

Index	*K-means*	*Hierarchical cluster*	*GMM*
The within clusters sum of squares	6.2E8	11.3E8	11.6E8
The average silhouette width	0.341	0.176	-0.0239
Calinski-Harabasz index	843.4	310.2	182.9
Dunn index	0.0266	0.0240	0.0071

The indices show that the k-means algorithm is the best since its within clusters sum of squares is the smallest and its average silhouette width, the Calinski-Harabasz index and Dunn index are the biggest among the three.

C. Driving Behavior Safety Level Evaluations

After clustering the driving behavior safety levels, classification methods were used to evaluate the crash risk of driving behavior and further identify the safety levels. For this purpose, the widely used support vector machine (SVM) and decision trees were used.

The original dataset with clustered labels was divided randomly with the help of the stratified sampling technique into training data and test data, with 1197 observations (i.e., 70.0%) and 514 observations (i.e., 30.0%), respectively. The training data was applied to train SVM models and decision trees. Firstly, 80 SVM models, with different key parameters (the kernel function, the gamma and the cost), were developed to identify the best SVM model. Eight different gammas (i.e. 0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10) and five different costs (i.e. 0.01, 0.01, 1, 10, and 100) were considered for each of two kernel functions (i.e. radial and linear). Finally, the best model was identified for different clustering algorithms. TABLE III. lists the results of SVM models. The total accuracy of the three best SVM models are quite high (> 93.0%). It means that the developed SVM models can well identify the driving behavior safety levels in the data based on the three clustering methods.

TABLE III. RESULTS OF SVM MODELS

Parameters	*K-means*	*Hierarchical cluster*	*GMM*
Kernel	Linear	Radial	Linear
Gamma	0.001	0.01	0.001
Cost	10	10	100
Number of Support Vectors	110	165	107
Best performance	0.0251	0.0485	0.0169
Total Accuracy	97.3%	93.4%	98.7%

The test data was further used to test the developed SVM models and decision trees. The results are listed in TABLE IV. , TABLE V. and TABLE VI. The total accuracy of SVM model in the k-means clustering scenario is (95+126+232+35) / 514 = 94.9%, and the percentages of true predictions for each traffic safety levels are higher than 84.0%. Similarly, the total accuracy of SVM models in hierarchical clustering and GMM scenario are 95.5% and 97.9%, respectively, whereas the total accuracy of decision trees in k-means clustering scenario, hierarchical clustering and GMM scenario are 92.0%, 93.8% and 95.9%, respectively. Therefore, it can be found that the SVM models perform better than the decision trees. For each model, there are no significant differences between the accuracy from the training data and the test data in the three clustering algorithms. This indicates that the developed SVM model and decision trees are reasonable and well developed. Besides, the safety levels of driving behaviors are all well identified. Importantly, the total accuracy in GMM scenario is the highest among the three scenarios. By ignoring the performance difference between the developed models, we can conclude that the GMM can slightly improve the clustering performance of the safety level of driving behaviors. This can also reflect that the optimal safety level / cluster is four.

TABLE IV. THE EVALUATED LEVELS OF SVM AND DECISION TREE IN K-MEANS CLUSTERING SCENARIO

Model	***Clustered***	***Predicted***					
		1	***2***	***3***	***4***	***True***	***False***
SVM	1	95	7	2	8	84.8%	15.2%
	2	1	126	3	0	96.9%	3.1%
	3	0	2	232	0	99.1%	0.9%
	4	0	0	3	35	92.1%	7.9%
	Total					94.9%	5.1%
Decision tree	1	43	0	0	0	100.0%	0.0%
	2	15	106	0	15	77.9%	22.1%
	3	8	0	84	3	88.4%	11.6%
	4	0	0	0	238	100.0%	0.0%
	Total					92.0%	8.0%

TABLE V. THE EVALUATED LEVELS OF SVM AND DECISION TREE IN HIERARCHICAL CLUSTERING SCENARIO

Model	***Clustered***	***Predicted***				
		1	***2***	***3***	***True***	***False***
SVM	1	374	14	0	96.4%	3.6%
	2	8	114	1	92.7%	7.3%
	3	0	0	3	100.0%	0.0%
	Total				95.5%	4.5%
Decision tree	1	371	11	0	97.1%	2.9%
	2	20	107	1	83.6%	16.4%
	3	0	0	4	100.0%	0.0%
	Total				93.8%	6.2%

TABLE VI. THE EVALUATED LEVELS OF SVM AND DECISION TREE GMM SCENARIO

Model	***Clustered***	***Predicted***					
		1	***2***	***3***	***4***	***True***	***False***
SVM	1	252	0	0	0	100.0%	0.0%
	2	2	117	0	0	98.3%	1.7%
	3	1	0	24	3	85.7%	14.3%
	4	1	0	4	110	95.7%	4.3%
	Total					97.9%	2.1%
Decision tree	1	240	2	0	2	98.4%	1.6%
	2	0	120	0	0	100.0%	0.0%
	3	0	0	24	3	88.9%	11.1%
	4	0	0	14	108	88.5%	11.5%
	Total					95.9%	4.1%

V. CONCLUSION

The findings of this paper proved that driving safety could be clustered into several levels: ideally four. They can be labelled as “normal” driving, “low risk” driving, “middle risk” driving and “high risk” driving. Among k-means clustering, hierarchical clustering, and a model-based clustering (i.e., GMM), k-means clustering gave the optimal number of clusters. The combination of developed SVMs and GMM outperformed the other combined algorithms; yet, the difference was not significant. This further supports the hypothesis that the driving data is well clustered in various levels, and that models could be developed for safety level classifications. Additionally, it is also found in this paper that the SVM models perform better than the decision trees. Still, this research does not come without limitations. The dataset did not include existing variables on drivers’ demographics and attitudes and perceptions. Future work should also focus on identifying the factors (variables) of importance associated with one or the other driving levels and their relationships.

ACKNOWLEDGMENT

This work has been funded from the European Union’s Horizon 2020 research and innovation programme i–DREAMS under grant agreement No 814761.

REFERENCES

[1] European Comission. Directorate-General for Mobility and Transport. (2011). White Paper on Transport: Roadmap to a Single European Transport Area: Towards a Competitive and Resource-efficient Transport System. Publications Office of the European Union.

[2] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

[3] Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2019). Hierarchical clustering: Objective functions and algorithms. Journal of the ACM (JACM), 66(4), 1-42.

[4] McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics, 26(21), 2705-2712.

[5] Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature methods, 16(3), 243-245.

[6] Antoniou, C., Koutsopoulos, H. N., & Yannis, G. (2013). Dynamic data-driven local traffic state estimation and prediction. Transportation Research Part C: Emerging Technologies, 34, 89-107.

[7] Zhang, J., Chen, W., Gao, M., & Shen, G. (2017). K-means-clustering-based fiber nonlinearity equalization techniques for 64-QAM coherent optical communication system. Optics express, 25(22), 27570-27580.

[8] Smoliński, A., Walczak, B., & Einax, J. W. (2002). Hierarchical clustering extended with visual complements of environmental data set. Chemometrics and Intelligent Laboratory Systems, 64(1), 45-54.

[9] Balcan, M. F., Liang, Y., Gupta, P. (2014). Robust hierarchical clustering. The Journal of Machine Learning Research, 15(1), 3831-3871.

[10] McNicholas, P. D., & Murphy, T. B. (2010). Model ‐ based clustering of longitudinal data. Canadian Journal of Statistics, 38(1), 153-168.

[11] Model-based clustering and Gaussian mixture model in R. https://en.proft.me/2017/02/1/model-based-clustering-r/

[12] Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention, 51, 252-259.

[13] Yang, K., Wang, X., & Yu, R. (2018). A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transportation research part C: emerging technologies, 96, 192-207.

[14] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[15] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. Misc functions of the department of statistics, Probability Theory group, TU Wien, v R package version 1.6-8. 2017.

[16] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[17] Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). Package 'rpart'. Available online: https://cran.r-project.org/web/packages/rpart/rpart.pdf.

[18] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[19] Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-SNE effectively. Distill, 1(10), e2.

[20] Christian Hennig (2020). fpc: Flexible Procedures for Clustering. R package version 2.2-5. https://CRAN.R-project.org/package=fpc

[21] Garay, A. B., Contreras, G. P., & Escarcina, R. P. (2011). A GH-SOM optimization with SOM labelling and dunn index. 11th International Conference on Hybrid Intelligent Systems (HIS) (pp. 572-577).

[22] Łukasik, S., Kowalski, P. A., Charytanowicz, M., & Kulczycki, P. (2016). Clustering using flower pollination algorithm and Calinski-Harabasz index. In 2016 IEEE Congress on Evolutionary Computation (CEC) (pp. 2724-2728).