Feature importance estimation using clustering and classification for risky driving behaviour

Tuo Mao[a*], Adriana-Simona Mihaita[a], Yuming Ou[a], Fang Chen[a]

[a] University of Technology Sydney, Faculty of Engineering and IT, School of Computer Science, 61 Broadway Ultimo, NSW, Australia.
tuo.mao@uts.edu.au*, adriana-simona.mihaita@uts.edu.au, yuming.ou@uts.edu.au, fang.chen@uts.edu.au.

* Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)

September 02-03, 2024, Crete, Greece

April 14, 2024

Keywords: Risky driving behaviour, Clustering, Classification, Feature importance calculation

1 INTRODUCTION

Dangerous driving behaviour, encompassing speeding, reckless driving, driving under the influence, and distractions, contributes significantly to global traffic accidents and fatalities (Sarkar and Andreas, 2004). Near misses, situations narrowly avoiding accidents, expose passengers to g-forces with detrimental effects on the body.

Studies traditionally rely on traffic violation and crash data, which fail to capture near-miss events, a common consequence of risky driving behaviours. Such behaviours, including speeding and tailgating, increase the likelihood of near misses (Elvik et al.).

Observing road near misses presents challenges due to limited data. Methods include self-reported questionnaires, simulations, and onboard vehicle sensors. Environmental factors like traffic congestion and poor road conditions exacerbate risky behaviours, while peer pressure and social norms also play roles (Ivers et al., Tao et al.).

Individual characteristics like age and experience influence risky driving. Young drivers exhibit higher risk due to factors like impulsivity, while older drivers may experience declines in cognitive and physical abilities. Targeted interventions like driver education programs aim to mitigate these risks (Tao et al., Ivers et al.).

Technology, notably mobile phones and in-vehicle systems, contributes to distracted driving, amplifying accident risks. G-force, experienced during sudden stops or collisions, correlates with injury severity, underlining the importance of curbing dangerous driving (Klauer et al.).

CompassIoT data provides g-force metrics for road near misses, offering valuable insights into risky driving behaviours, including GPS location, vehicle dynamics, and timing.

Key techniques for quantifying feature importance in clustering include the Centroid Variance Method (Sujatha et al., 2013), focusing on feature variance to distinguish clusters. Unsupervised to Supervised Conversion treats clusters as classes, leveraging supervised classification for feature importance. ANOVA/Chi-Square Tests (McHugh, 2013) measure feature distribution differences, while Leaving-One-Out Testing (Wong, 2015) evaluates individual feature impact directly.

This paper's major contributions include:

- It utilizes new vehicle telematics and IoT data, offering broad coverage and frequent updates, providing comprehensive records of risky driving dynamics. This dataset is less biased than surveys, with larger sample sizes.
- It introduces an algorithm assessing each feature's impact, enhancing model transparency and applicability to diverse datasets. This framework requires no preset models, offering richer insights than traditional analyses.
- It pioneers analysing risky driving behaviour location distributions, validating findings and exploring absolute and relative locations to intersections.

2 METHODOLOGY

The proposed framework (Figure 1) begins with raw data comprising various numeric and non-numeric features, with only numeric features retained after the initial step. Data cleaning ensues, addressing missing values, formatting issues, outliers, and noise.

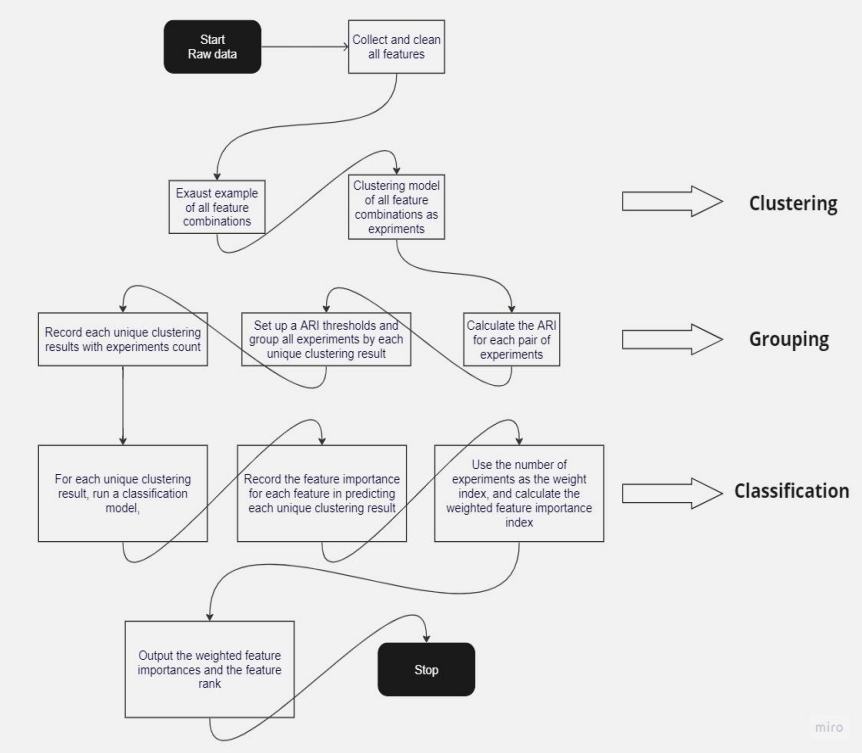

Figure 1 The proposed framework flowchart.

Next, the clustering model processes each combination of collected features, employing exhaustive sampling to identify optimal clustering of the dataset across diverse feature sets. The primary clustering model utilized is K-means, an unsupervised method known for its stability and interpretability. While other models like Spectral clustering and Gaussian mixture models are tested, K-means consistently yield stable and understandable results, thus exclusively presented due to space constraints.

In K-means, centroids representing cluster centres are randomly selected, with the model iteratively optimizing cluster assignments to minimize the sum of squared errors (SSE). Multiple trials explore varying cluster numbers, plotting SSE against clusters to discern an elbow point indicating optimal cluster count, typically achieved programmatically using the "kneed" Python package.

Following clustering, experiments with similar results are grouped, employing the Adjusted Rand Index (ARI) to assess clustering similarity. Experiments exceeding a predefined ARI threshold are

clustered together, while those below are considered separate. Each group's representative experiment yields a unique clustering result, merged with the raw data's features.

To elucidate feature importance, a classification model is applied to the raw data, with the unique clustering result as labels. Feature importance indexes are computed and recorded in a data frame (Table 1), weighted-averaged by experiment count. This finalizes the feature importance index calculation, crucial for subsequent analysis.

Table 1 The output feature importance indexes table

Group	Feature 1 importance	...	Feature V importance	Experiment Count
Group 1	I11	...	I1V	N1
Group 2	I21	...	I2V	N2
...	...	...	...	...
Group Z	IZ1	...	IZV	NZ

3 RESULTS

The main dataset collected in this research is provided by CompassIoT using vehicular sensor devices that stream data to their cloud servers and are available via licensed CompassIoT services. There are 9 parameters collected for this accident dataset, such as longitude and latitude coordinates, timestamp of the accident occurred, X-acceleration, Y-acceleration, speed (km/h), max speed (km/h), lane count, and classification of the accident recorded. A total of 6699 incidents are collected over 3 years period (2020 to 2023). We further processed the data and generated the cumulative turning angle and maximum turning angle from the incidents' nearest 10-meter road sections.

Table 2 shows the Weighted Feature Importance of all the input features. Distance to the nearest intersection and distance to the nearest road curb are the top most important features whose feature importance index is greater than 0.18. Other features are less significant since their feature importance is less than 0.1.

Table 2 The WFI results in ARI threshold = 0.9

Feature	Distance to intersection	Distance to curb	longitude	speed	Max speed	Cumulative angle	latitude	Max angle	xacc	yacc	lane count
Weighted Feature Importance	0.332	0.185	0.096	0.073	0.070	0.059	0.041	0.036	0.019	0.018	0.011

4 DISCUSSION

By comparing all the feature importance outcomes, several typical findings can be summarized. Distance to the nearest intersection and distance to the nearest road curb are the most critical features in identifying risky driving behaviour.

Compared to the other feature importance models (Centroid Variance Method, ANOVA/Chi-Square Tests, and Leaving-One-Out Testing), the proposed model has the following advantages:

1. Most feature importance metrics come from known clustering labels and calculate the feature importance contributing to a specific cluster (k-means clusters feature importance and random forest classification feature importance). The proposed model tests all combinations of features and concludes the most popular cluster patterns (groups) which is more generalized than most existing models.

2. Most feature importance methods are used for feature selection (such as the Centroid Variance Method embedded in the k-means clustering model) and feature validation (such as the leaving-one-out cross-validation model) and do not reveal the quantified index of feature importance.

2. Few feature importance models consider the importance of the combinations of features like the proposed model. For example, when identifying the harsh right-turn behaviour that happens on right-turn road geometry, we observe the feature combinations of y acceleration, cumulative turning angle and maximum turning angle and discover only 4% of total risky driving behaviours are caused by right-turn road geometry.

3. The interpretability of the proposed model is very good, and the model is transparent. You can check any clusters and any feature's importance contributes to any cluster pattern.

The major drawbacks of the proposed model are as follows:

1. Due to the exhausts exampling of all combinations of features, the model requires a huge amount of computation power. With the increment of input features, the computation load increases exponentially. Luckily, this computation difficulty can be released by utilizing parallel computing since the clustering and classification between each combination of features are independent. We can also make this algorithm linearly scalable by adapting the leave-one-out method when we leave the most important feature out after examining a fixed number of feature combinations.

5 CONFLICT OF INTEREST

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

6 REFERENCES

ELVIK, R., HYE, A., VAA, T. & SRENSEN, M. 2009. The Handbook of Road Safety Measures.

IVERS, R., SENSERRICK, T., BOUFOUS, S., STEVENSON, M., CHEN, H.-Y., WOODWARD, M. & NORTON, R. 2009. Novice drivers' risky driving behavior, risk perception, and crash risk: findings from the DRIVE study. *American journal of public health,* 99**,** 1638-1644.

KLAUER, S. G., GUO, F., SIMONS-MORTON, B. G., OUIMET, M. C., LEE, S. E. & DINGUS, T. A. 2014. Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine,* 370**,** 54-59.

MCHUGH, M. L. J. B. M. 2013. The chi-square test of independence. 23**,** 143-149.

SARKAR, S. & ANDREAS, M. 2004. Acceptance of and engagement in risky driving behaviors by teenagers. *Adolescence,* 39.

SUJATHA, S., SONA, A. S. J. I. J. O. E. R. & TECHNOLOGY 2013. New fast k-means clustering algorithm using modified centroid selection method. 2**,** 1-9.

TAO, D., ZHANG, R. & QU, X. 2017. The role of personality traits and driving experience in self-reported risky driving behaviors and accident risk among Chinese drivers. *Accident Analysis and Prevention,* 99**,** 228-235.

WONG, T.-T. J. P. R. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. 48**,** 2839-2846.