Classification Of Drivers Based On Their Driving Patterns

Rutul Patel*, Saumil Patel†, Kunj Kanzariya‡, and Ayush Patel§
*†‡§Computer Science and Engineering, Ahmedabad University, Ahmedabad, India
{rutul.p, saumil.p, kunj.k1, ayush.p3}@ahduni.edu.in

***Abstract*—Driving behavior analysis is crucial for enhancing road safety and developing intelligent transportation systems. Traditional evaluation methods often fall short in capturing the complexity of driving patterns. However, recent advancements in data analytics and sensor technologies enable the collection of rich datasets, facilitating more comprehensive analyses. In this study, we propose a methodology that leverages dynamic time warping (DTW) and K-means clustering to categorize driving behavior based on risk patterns derived from UAV-collected data. By representing driving patterns as variable-length time series, we create classification models that capture subtleties and temporal relationships. The K-means algorithm, guided by DTW similarity metrics, partitions driving patterns into clusters, revealing distinct behavior profiles. Statistical techniques aid in determining the optimal number of clusters, ensuring effective characterization of driving patterns. Our results demonstrate the effectiveness of the proposed approach in uncovering valuable insights for road safety and transportation decision-making. Personalized approaches derived from identified behavior clusters can significantly contribute to improving road safety and driving experiences.**

***Index Terms*—Unsupervised Learning, Clustering, Driving risk, Time-series**

I. Introduction

A. Background

Driving behavior is a complex interaction of various factors, including individual habits, external conditions, and situation. Understanding and analyzing driving patterns have become important for enhancing road safety, designing intelligent transportation systems, and developing advanced driver-assistance systems (ADAS). Conventional techniques for evaluating driving behavior frequently depend on subjective observations or constrained metrics, which may not adequately represent the entire range of driving behaviors. But because to developments in data analytics and sensor technologies, it's now possible to collect rich datasets encompassing diverse driving scenarios and behaviors.

B. Motivation

To utilize the abundance of driving data accessible today to create advanced and accurate models for categorizing driving behavior. Our goal is to uncover valuable insights that can enhance road safety and enable better decision-making in transportation. The dynamic and evolving nature of driving behavior is acknowledged by presenting driving patterns as variable length time-series. By using this method, we are able to produce classification models that are more resilient by capturing the subtleties and temporal relationships present in driving data. Furthermore, by using machine learning methods to examine these patterns, we can find undiscovered correlations, pinpoint risk factors, and eventually aid in the creation of intelligent systems that can foresee and stop unfavorable driving outcomes.

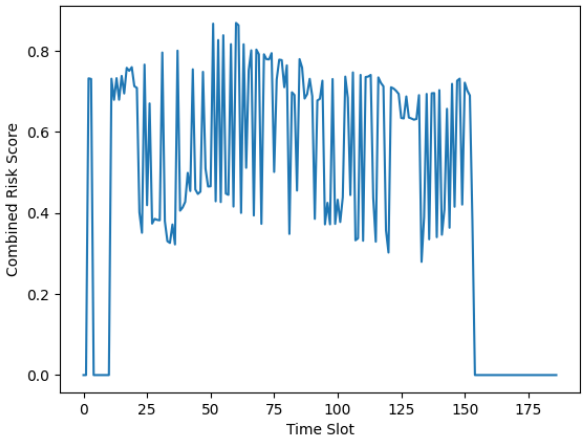

Fig. 1. Risk vs Time slot graph of an example vehicle

II. Dataset Description

The Data is collected using UAV. An camera mounted Drone was used to get top view of the road. YOLOv5 was used to detect the position of moving vehicles on road. Where each data point was captured every 33ms. So, If a car is going fast it will have lower number of time-series data compared to a slow moving vehicle. Each data point contains the coordinates of the car at that particular time instant, velocity and direction. Then a risk value was assigned to each car at each instant. In Fig. 1 we visualize the risk vs time of a particular car.

III. Problem formulation

Our task is to find similar drivers based on their driving pattern. Each driving pattern is associated with a risk value at every time instance, which is labeled by hand. These risk values form a time-series for each driver observed in the dataset. Importantly, each time-series will vary in length depending on the duration the driver remains within the frame, leading to variable-length time series data.

We denote the time series for a driver i as $D^i = \{d_{1i}, d_{2i}, d_{3i}, ..., d_{Ti}\}$, where T represents the number of time instances, and d_{ti} is the risk value at time t.

A. Euclidean Distance

Euclidean distance provides a linear measure of similarity between two time series of equal length. The Euclidean distance between two time series D_i and D_j, each of length N, is calculated as:

$$\text{Euclidean}(D_i, D_j) = \sqrt{\sum_{n=1}^{N} (d_{in} - d_{jn})^2}$$

This measure is simple but requires that the time series have the same length.

B. Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is used for measuring the similarity between two time series that may not align in time, speed, or length. DTW finds the optimal alignment by minimizing the distance between corresponding points. The distance is computed as:

$$\text{DTW}(D_i, D_j) = \min \sqrt{\sum_{(t_k, t'_k) \in \text{path}} (d_{it_k} - d_{jt'_k})^2}$$

C. DTW Barycenter Averaging (DBA) K-means

DBA K-means is an algorithm that extends K-means clustering by using Dynamic Time Warping Barycenter Averaging (DBA) to compute the average of a cluster. This method allows for averaging time series of variable lengths by computing a consensus series that minimizes the DTW distance to the series in the cluster. The centroid of a cluster C_i in DBA K-means, represented by μ_i, is computed as:

$$\mu_i = \text{DBA}(C_i) = \arg\min_{\mu} \sum_{x \in C_i} \text{DTW}(x, \mu)^2$$

DBA provides a centroid that is a better representative for clusters of time series data, accommodating the variabilities in length and alignment.

D. Objective Function for K-means Clustering

The objective function for k-means clustering, now incorporating Euclidean distance, DTW, and DBA as similarity metrics, is formulated as:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \text{Metric}(x, \mu_i)^2$$

Here, Metric can be Euclidean, DTW, or DBA depending on the clustering approach. The number of clusters is k, C_i is the ith cluster, x is a time series in cluster C_i, and μ_i is the centroid of cluster C_i, which may be the arithmetic mean or a DTW barycenter depending on the chosen metric.

IV. Methodology

We use K-means clustering approach to find similar driving patterns. We start by pre-processing the data. In which we remove all the time-series with zero risk at all time instances. Also handled any missing values, outliers, or noise, ensuring data uniformity in format and scale through normalization or standardization techniques.

Next, we employ statistical techniques such as the elbow method or silhouette score (Table 1) to determine the optimal number of clusters (k) for the k-means algorithm. Through experimentation and evaluation of clustering performance, we identify the most suitable k that effectively captures distinct driving behavior patterns. Utilizing the k-means algorithm, we initialize k centroids randomly and proceed to partition the driving patterns into k clusters based on similarity. This iterative process involves assigning data points to the nearest centroid and updating centroids until convergence.

We leverage the capabilities of the tslearn[1] framework, a powerful toolkit designed specifically for time series analysis and clustering tasks. By integrating tslearn into our workflow, we gain access to a wide range of functionalities tailored for handling variable-length time-series data effectively.

A. Data Preprocessing and Enhancement

To handle the variability and complexity of the collected driving behavior time-series data, a preprocessing technique was employed. The goal was to refine the dataset for improved clustering performance. It was observed that the time-series data often contained leading and trailing zeros, indicating periods of inactivity or negligible risk. To address this, a truncation method was implemented. This method aimed to remove extraneous zeros from the beginning and end of each series, standardizing data lengths and emphasizing periods of significant driving activity.

This truncation not only streamlined the dataset by eliminating redundant data points but also preserved the critical temporal sequences reflective of driver risk behaviors. It was imperative that this refinement process safeguard the zeros integral to the active risk periods, ensuring the fidelity of each driver's risk profile remained intact. Consequently, the dataset was distilled to encompass only meaningful risk evaluations, significantly enhancing the granularity and comparability of the time-series data.

B. Implications of Truncated Data on Clustering

Our preprocessing work prepared the data for advanced clustering methods. We got rid of unnecessary zeros and removed any series that were all zeros, which helped us see the subtle differences in driving risk more clearly. By using Dynamic Time Warping (DTW) along with K-means clustering, we found distinct groups of driving behaviors, each showing different kinds of risk. This success shows how important it is to properly prepare your data before analyzing it. Doing this helped us understand driving patterns better, which is a big step forward for improving road safety and smart transportation systems.

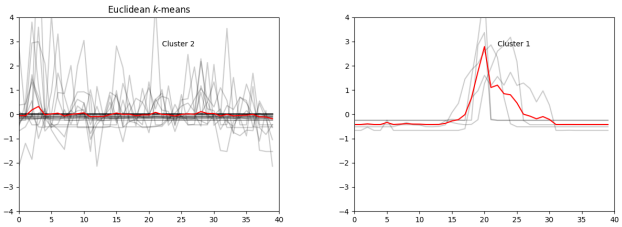

Fig. 2. Eucledian Distance Clustering

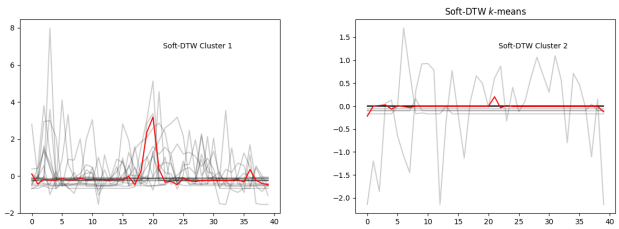

Fig. 3. Soft DTW Clustering

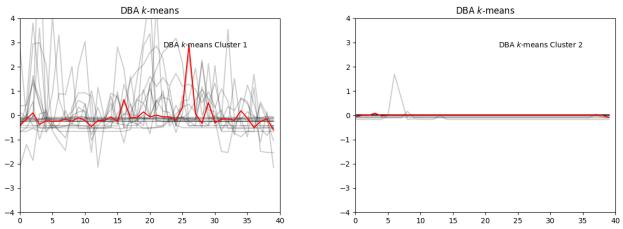

Fig. 4. DBA KMeans Clustering

V. Results/Inferences

TABLE I
Comparison of Silhouette Scores for Different Clustering Methods

Cluster Number	CRS	WSDsRS	WSDbRS	WRARS
2	0.3683	0.1933	**0.8677**	0.1661
3	0.3538	0.1954	**0.7642**	0.1688
4	0.3189	0.2741	**0.8042**	0.1109
5	0.1422	0.2648	**0.5308**	0.1012

The WSDb clustering method performed better than the combined Risk scores as well as individual risk score approaches. [CRS - Combined Risk Score, WSDsRS - WSD Risk Score, WSDbRS - WSDb Risk Score.]

Figure 2 illustrates the clusters formed by using the Euclidean distance within the K-means algorithm. First image captures a prominent peak in driving risk, suggesting an instance of aggressive driving, while the second image shows a cluster characterized by a lower risk profile.

Similarly, the utilization of Soft-DTW within K-means clustering is visualized in Figure 3. These clusters reveal nuanced driving patterns, with first figure showing sporadic risk patterns and second image displaying a more uniform risk distribution.

Figure 4 shows the clusters obtained from the DBA K-means algorithm. In these figures, the grey lines represent individual driving patterns, while the red line indicates the average pattern or the centroid of the cluster. First image in that demonstrates a more volatile driving pattern, indicative of higher variability in driving behavior, whereas second depicts more stable behavior with less deviation from the mean.

Figure 5 presents a comparison of silhouette scores, contrasting the combined approach with the truncated WSDb-only approach. The line chart demonstrates a marked improvement in silhouette scores when the WSDb-only approach is employed, particularly evident in clusters with two and three instances. This implies that the WSDb approach, especially when truncated to eliminate irrelevant data, is more effective in discerning distinct driving patterns compared to the combined approach.

As illustrated in Figures 2-5, the identified clusters provide valuable insights into the diverse driving behaviors observed within the dataset. These figures visually represent the distinct characteristics and trends within each cluster. Understanding this information enables targeted interventions and personalized approaches for improving road safety, enhancing traffic management strategies, and developing effective driver assistance systems.

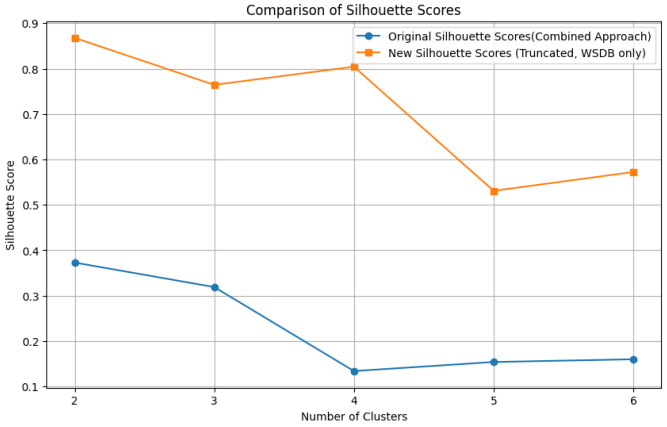

Fig. 5. Comparison of silhouette scores from the Combined Approach and the WSDb Approach.

VI. Current Limitations and Future Work

Despite the advances demonstrated in our clustering methodology, there are limitations that need further investigation. One of the challenges lies in the weighted sum approach, which did not perform as expected. The inherent complexity and multi-dimensional nature of driving patterns could be contributing factors that diminish the efficacy of this method. Future work may explore more sophisticated multi-dimensional modeling techniques that may more accurately capture the complex interplay of risk factors in driving behaviors.

Another area of focus for future work is the enhancement of real-time processing capabilities. As data collection methods grow more sophisticated, the volume and velocity of incoming data will increase. Developing algorithms that can process and analyze this data efficiently will be crucial for the immediate application of our findings in real-time systems.

Additionally, we aim to enrich our dataset with more diverse driving scenarios and conditions. This will allow our clustering algorithms to identify and adapt to a wider range of driving behaviors, potentially leading to more robust and versatile classification models.

VII. Conclusion

The present study has established a methodological framework that effectively clusters driving behaviors using UAV-collected time-series data. The integration of DTW and K-means clustering has allowed us to discern distinct patterns in driving risk, laying the groundwork for developing intelligent transportation systems tailored to individual behaviors. However, the journey toward an exhaustive understanding and prediction of driving behavior is ongoing. Future work will seek to refine these clustering techniques further and explore their integration into real-time systems for enhancing road safety. Ultimately, the goal is to facilitate the development of advanced driver-assistance systems and contribute to the broader objective of ensuring a safer driving experience for all road users.

References

[1] R. Tavenard et al., "Tslearn, A Machine Learning Toolkit for Time Series Data," *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1-6, 2020.

[2] B. -c. Guo, Y. -l. Wang, M. Gao, J. Lu, G. -s. Han and L. -b. Zhang, "End to End Autonomous Driving Behavior Prediction Based on Deep Convolution Neural Network," 2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI), Boston, MA, USA, 2022, pp. 1-6, doi: 10.1109/DTPI55838.2022.9998956.

[3] B. I. Kwak, M. L. Han and H. K. Kim, "Driver Identification Based on Wavelet Transform Using Driving Patterns," in IEEE Transactions on Industrial Informatics, vol. 17, no. 4, pp. 2400-2410, April 2021, doi: 10.1109/TII.2020.2999911.

[4] Fu, X., Meng, H., Wang, X., Yang, H., & Wang, J. (2022). A hybrid neural network for driving behavior risk prediction based on distracted driving behavior data. PloS one, 17(1), e0263030. https://doi.org/10.1371/journal.pone.0263030

[5] Dynamic time warping (DTW). File Exchange - MATLAB Central. (n.d.). https://www.mathworks.com/matlabcentral/fileexchange/43156-dynamic-time-warping-dtw

[6] K-means clustering algorithm - javatpoint. (n.d.). https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning