SCHOLARLY REVIEW ONLINE

Scholarly Review Online - Winter 2024/2025

Aggressive Driving Behavior Recognition based on K-means Clustering

By Green Cai

AUTHOR BIO

Green Cai is a 12th-grade homeschooled student from Beijing. She began coding at age 10, discovered a passion for AI in 10th grade, and devoted her 11th-grade year to studying machine learning. As an enthusiastic flutist, sailor, and barista, Green has a strong passion for interdisciplinary learning. Her goal is to utilize machine learning to make music production accessible to everyone.

ABSTRACT

Accurate identification of aggressive driving behavior is crucial for improving traffic safety and promoting social responsibility, and it is also a key component in the realization of intelligent transportation systems. This paper proposes a method for aggressive driving behavior recognition and data modeling based on K-means clustering. First, the NGSIM (Next Generation Simulation) vehicle trajectory dataset is used for preprocessing the raw data, and effective behavioral feature representations are obtained through feature extraction. Next, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data, extracting three main components to simplify the data structure. Subsequently, the K-means clustering algorithm is employed to classify the reduced data and identify potential patterns of aggressive driving behavior. Experimental results demonstrate that the proposed method effectively identifies aggressive driving behavior, providing technical support for improving traffic system safety and raising awareness of social responsibility.

Keywords: Aggressive driving behavior, K-means clustering, Principal Component Analysis

INTRODUCTION

Aggressive driving behavior is a significant contributor to traffic accidents and road safety issues globally. Timely detection and prediction of such behavior can substantially reduce accidents and enhance road safety by enabling drivers or autonomous systems to take appropriate actions [2]. According to the National Highway Traffic Safety Administration (NHTSA), approximately 95% of road traffic accidents are caused entirely or in part by human factors [2]. Aggressive actions, such as sudden lane changes and rapid acceleration or deceleration, introduce turbulence into traffic flow, increasing speed variability and reducing headway [8]. Identifying aggressive driving can assist in real-time interventions within Intelligent Transportation Systems (ITS), such as Advanced Driver Assistance Systems (ADAS), to correct improper maneuvers and reduce accidents [3].

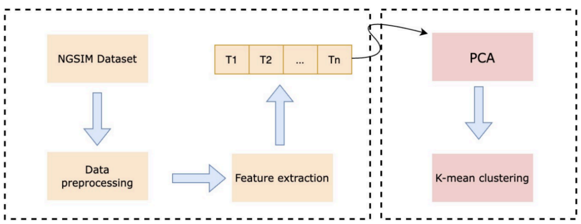

Figure 1.*Pipeline of the proposed methods.*

Existing methods for detecting aggressive driving often rely on subjective definitions and manually set thresholds based on specific driving parameters [6, 8]. For example, Lee [8] used a 1-second Time-to-Collision (TTC) threshold to identify aggressive driving on highways, while other studies employed acceleration-related thresholds [6, 7]. These approaches suffer from inconsistencies and lack generalizability across different driving scenarios. Moreover, manually defining thresholds can be both subjective and error-prone, limiting the effectiveness of these methods in real-world applications. Despite this success, there is a growing interest in utilizing large-scale vehicle trajectory data to identify driving behaviors [1]. Clustering algorithms, particularly K-means, have shown promise in grouping similar driving behaviors without prior labeling [3]. However, high-dimensional and correlated data can pose challenges for clustering algorithms, potentially leading to reduced accuracy and increased computational complexity.

To overcome these challenges, we propose a method for aggressive driving behavior recognition and data modeling based on K-means clustering. We demonstrate the proposed framework in Figure 1. Firstly, we utilize the NGSIM (Next Generation Simulation) vehicle trajectory dataset [2] for preprocessing the raw data and extract effective behavioral feature representations through comprehensive feature extraction. We extract 13 parameters, including mean, standard deviation, and median values of longitudinal and transverse velocities and accelerations, which capture key characteristics essential for identifying aggressive driving behaviors. Next, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data, extracting three principal components that retain 87% of the data variance. This step simplifies the data structure, mitigates feature redundancy, and reduces computational complexity, thereby enhancing the efficiency and effectiveness of the clustering process. Subsequently, the K-means clustering algorithm is employed to classify the reduced data and identify potential patterns of aggressive driving behavior. By clustering the driving behavior data without the need for labeled samples or subjective thresholds, our method provides a data-driven approach to detect aggressive driving patterns.

Our experimental results demonstrate that the proposed method effectively identifies aggressive driving behavior, achieving a high Silhouette coefficient of 0.84. This indicates robust clustering performance and clear separation between aggressive and non-aggressive driving behaviors. The visualization of clustering results further confirms the effectiveness of our approach. Our contributions are summarized as follows:

1. We propose a method for aggressive driving behavior recognition based on PCA and K-means clustering, effectively identifying aggressive driving patterns without relying

on subjective thresholds.

2. We extract comprehensive behavioral features from the NGSIM dataset and apply PCA for dimensionality reduction, retaining 87% of data variance while improving computational efficiency. We demonstrate through experiments that our approach achieves a high Silhouette coefficient of 0.84, indicating strong clustering performance and validating the effectiveness of our method.

3. Our method provides a scalable and generalizable solution for detecting aggressive driving behavior, with potential applications in intelligent transportation systems and traffic safety management.

METHODS

Overview

This work proposes a data-driven approach for identifying aggressive driving behaviors by integrating feature extraction, dimensionality reduction, and clustering techniques. The method consists of three major components: data preprocessing and feature extraction from the NGSIM vehicle trajectory dataset, dimensionality reduction using Principal Component Analysis (PCA), and clustering with an enhanced K-means algorithm. Our approach identifies aggressive driving patterns without relying on predefined thresholds or labeled data, ensuring scalability and adaptability across various driving environments.

Data Preprocessing and Feature Extraction

The NGSIM dataset, containing detailed vehicle trajectory data, captures key driving metrics such as position, velocity, and acceleration at a sampling rate of 10 Hz. To ensure the data is suitable for analysis, we perform several preprocessing steps. Missing data points are imputed using linear interpolation to maintain the temporal continuity of the data. Outliers, defined as data points with a z-score exceeding 3, are removed to prevent skewing the analysis. Finally, each feature is standardized to zero mean and unit variance to ensure compatibility with clustering algorithms, which are sensitive to the scale of input features.

We extract thirteen features to represent both longitudinal and lateral driving dynamics. These features include the mean, median, and standard deviation of velocity and acceleration in both directions, as well as maximum deceleration to capture sudden braking events. These features provide a comprehensive view of the vehicle's driving behavior and form the basis for identifying aggressive driving patterns.

Dimensionality Reduction with PCA

Given the high dimensionality of the extracted features, Principal Component Analysis (PCA) is applied to reduce the feature space while retaining the most informative components. PCA transforms the original, correlated features into a set of uncorrelated principal components that maximize variance in the data.

Let $X \in \mathbb{R}^{n\times p}$ represent the standardized feature matrix, where n is the number of samples, and p is the number of features. The covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ is computed as:

$$\Sigma = \frac{1}{n-1} X^T X$$

where X^T is the transpose of the matrix X. The principal components are derived by solving the eigenvalue problem:

$$\Sigma v_i = \lambda_i v_i$$

where v_i represents the i^{th} eigenvector, and λ_i is the corresponding eigenvalue. The eigenvectors v_i define the directions of the principal components, while the eigenvalues λ_i indicate the variance captured by each principal component.

The original feature matrix X is then projected onto the top k principal components, where k is the number of components retained. This results in a reduced feature matrix

$Z \in \mathbb{R}^{n \times k}$:

$$Z = XW_k$$

where $W_k \in \mathbb{R}^{p \times k}$ is the matrix of the top k eigenvectors. In our study, the first three principal components, which together explain 87% of the total variance, are selected for further analysis. This reduction simplifies the clustering process while preserving the most important aspects of driving behavior.

K-means Clustering with Improved Initialization and Weighted Distance

We employ K-means clustering to group the driving data into distinct clusters, representing different driving behaviors, including aggressive patterns. Traditional K-means clustering is sensitive to the initialization of cluster centers, which can lead to suboptimal clustering. To mitigate this, we adopt the K-means++ initialization method, which improves the initial selection of cluster centers by maximizing their separation. This is achieved by selecting the first cluster center randomly, and then selecting each subsequent centerc_j with probability proportional to its squared distance from the nearest center already chosen. The probability for each data point x_i to be selected as a center is given by:

$$P\left(c_j = x_i\right) = \frac{d\left(x_i, c\right)^2}{\Sigma_{x_i \in X} d\left(x_i, c\right)^2}$$

where $d(x_i, c)$ is the Euclidean distance between the point x_i and the nearest already selected center c.

After initialization, K-means iteratively assigns each data point x_i to the nearest cluster centerc_k, minimizing the following objective function:

$$L = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k} \|x_i - c_k\|^2$$

where $z_{i,k} = 1$ if point x_i is assigned to cluster$k$, and $z_{i,k} = 0$ otherwise. The cluster centers are updated by computing the mean of the points assigned to each cluster. This iterative process continues until convergence, meaning that the cluster assignments no longer change. We further improve the clustering by introducing a weighted distance metric. Instead of using the standard Euclidean distance, we weigh the contribution of each principal component based on the amount of variance it explains. Let ω_n represent the weight for the n-th principal component, where ω_n is proportional to the variance explained by that component. The weighted distance metric is defined as:

$$d\left(x_i, c_k\right) = \sqrt{\sum_{n=1}^{k} \omega_n \left(x_{i,n} - c_{k,n}\right)^2}$$

where $x_{i,n}$ is the n-th principal component of the point $x_{i,n}$, and $c_{k,n}$ is the n-th component of the cluster center $c_{k,n}$. This ensures that the clustering process emphasizes the most informative features, improving the accuracy of detecting aggressive driving behavior.

Computational Complexity and Efficiency

The computational complexity of the K-means algorithm per iteration is $O(NKd)$, where N is the number of data points, $K$is the number of clusters, and d is the number of dimensions after PCA. By reducing the dimensionality from 13 features to 3 principal components, we significantly reduce the computational load. Furthermore, K-means++ initialization accelerates the convergence of the algorithm, leading to fewer iterations required to achieve stable cluster assignments.

EXPERIMENTS AND RESULTS

Datasets

For our experiments, we utilized the publicly available NGSIM (Next Generation

Simulation) dataset, which provides detailed vehicle trajectory data collected from real-world traffic scenarios. The dataset includes high-resolution information on vehicle positions, speeds, and accelerations, with a sampling rate of 10 Hz. In this study, we focus on a subset of the NGSIM dataset, specifically designed for aggressive driving behavior analysis. The selected subset contains $n = 200$ driving segments, divided into two categories: normal driving behavior (type 0) and aggressive driving behavior (type 1). The dataset is balanced, with an equal number of samples for each driving behavior type.

Experimental Setup

After data preprocessing, we extracted 13 key features from the driving behavior data. These features capture both longitudinal and transverse driving dynamics, which are critical for identifying aggressive driving patterns. The extracted features are as follows:

1. v^{+}_{ang}: Mean longitudinal speed, representing the average speed in the direction of travel.
2. v_{std}: Standard deviation of longitudinal speed, measuring the variability in speed.
3. v^{+}_{m}: Median longitudinal speed.
4. v_{ang}: Mean transverse speed, representing the average lateral speed (side-to-side movement).
5. v_{m}: Median transverse speed.
6. a^{+}_{ang}: Mean longitudinal acceleration, reflecting the average change in speed over time in the direction of travel.
7. a_{std}: Standard deviation of longitudinal acceleration.
8. a_{med}: Median longitudinal acceleration.
9. a_{ang}: Mean transverse acceleration, capturing the average lateral acceleration.
10. a_{std}: Standard deviation of transverse acceleration.
11. a_{med}: Median transverse acceleration.

These features allow us to analyze both the speed and acceleration profiles of the vehicle in both the longitudinal (forward/backward) and transverse (sideways) directions. Aggressive driving is typically characterized by higher variability in these features, such as rapid changes in speed or erratic lane changes. After feature extraction, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the feature space. The first three principal components, which together explain 87% of the variance in the data, are selected for clustering using the K-means algorithm. The K-means algorithm is initialized using K-means++ and configured to group the driving data into two clusters: normal driving behavior and aggressive driving behavior.

Evaluation Metrics

To evaluate the quality of the clustering results, we use the Silhouette coefficient, which measures how similar each point is to its assigned cluster compared to other clusters. The Silhouette coefficient $S(i)$ for a data point i is defined as:

$$S(i) = \frac{b(i) - a(i)}{max(a(i),\ b(i))}$$

where $a(i)$ is the average intra-cluster distance (the average distance between i and all other points in the same cluster), and $b(i)$ is the average inter-cluster distance (the average distance between i and all points in the nearest different cluster). The Silhouette coefficient ranges from -1 to 1, with values close to 1 indicating well-separated clusters, values close to 0 suggesting overlapping clusters, and negative values indicating poor clustering. In this study,

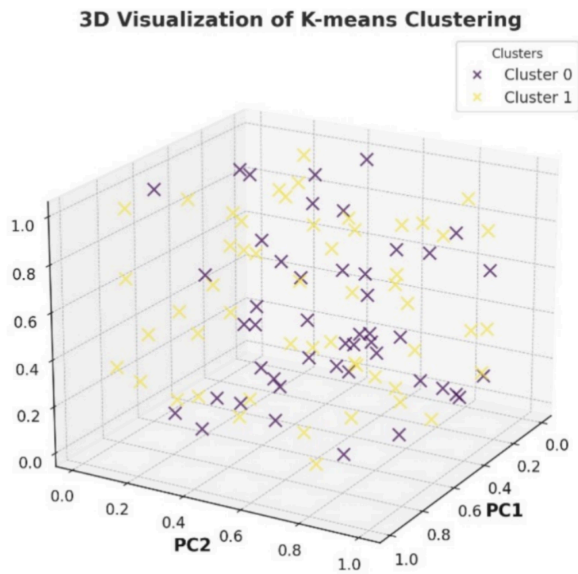

Figure 2. *3D visualization of K-means clustering*

we achieved a Silhouette coefficient of 0.84, indicating strong clustering performance.

Results

Figure 2 presents a 3D visualization of the clustering results in the PCA-reduced feature space. The samples are color-coded according to their assigned clusters: purple represents normal driving behavior (Cluster 0), and yellow represents aggressive driving behavior (Cluster 1). The distinct separation between the two clusters visually demonstrates the effectiveness of the K-means clustering algorithm in distinguishing between different driving patterns. The quality of the clustering is further validated by the Silhouette coefficient, which quantifies how well-separated and coherent the clusters are. In this study, we achieved a Silhouette coefficient of 0.84, indicating robust clustering performance. A value close to 1 suggests that the clusters are well-separated, and each point is assigned to the correct cluster with minimal overlap.

Driving Behavior Analysis Based on Clustering Results.

From the clustering results and the feature analysis in Figure 3, it is evident that aggressive drivers (Cluster 1) exhibit significantly different behaviors compared to normal drivers (Cluster 0) across several key features.

1. Mean Longitudinal Speed (v^{+}_{ang}): Aggressive drivers maintain a higher average speed (28.93 m/s) compared to normal drivers (23.93 m/s), which is expected since aggressive driving is typically associated with higher speeds.
2. Speed Variability (v_{std}): The standard deviation of longitudinal speed is also greater for aggressive drivers (8.00 m/s) than for normal drivers (4.90 m/s), indicating that aggressive drivers tend to fluctuate their speed more frequently, likely due to rapid accelerations and decelerations.
3. Transverse Speed (v_{ang}): Similarly, aggressive drivers exhibit higher mean transverse speeds (25.72 m/s) compared to normal drivers (21.18 m/s), suggesting that they make more lateral movements, such as lane changes, which are typical of aggressive driving.
4. Acceleration Patterns: Both longitudinal and transverse acceleration metrics show higher variability and magnitudes for aggressive drivers. For instance, the mean longitudinal acceleration (a^{+}_{ang}) is higher for aggressive drivers (0.53 m/s²) than for normal drivers (0.27 m/s²), and the standard deviation of longitudinal acceleration (a_{std}) is 5.20 m/s² for aggressive drivers compared to 4.27 m/s² for normal drivers. These higher values reflect the more erratic and dynamic nature of aggressive driving, which often involves sudden acceleration and deceleration events.

These metrics confirm the clear distinction between aggressive and normal driving behaviors. The clustering results not only visually separate the driving behaviors but are also supported by the quantitative differences in key driving parameters. Aggressive driving is characterized by higher speeds, greater variability in both speed and acceleration, and more frequent lateral

Figure 3. *Mean Feature Values of Different Driving Behaviors*

movements, as captured by the extracted features.

Evaluation of Clustering Model Performance.

The Silhouette coefficient, which measures how well-separated the clusters are, provides further validation of the clustering performance. In this study, a Silhouette coefficient of 0.84 was achieved, indicating that the clusters are well-formed and distinct. This value, close to 1, suggests that the majority of the data points are correctly assigned to the respective clusters with minimal overlap between the two driving behavior categories. The effectiveness of the clustering model can be attributed to several factors:

1. PCA-Based Dimensionality Reduction: By reducing the dimensionality of the feature space, PCA helps focus on the most important components of the data, making it easier for the clustering algorithm to identify patterns.
2. Feature Importance: The selected features, such as speed and acceleration, are highly relevant to distinguishing between different driving behaviors, making the clustering process more robust and accurate.
3. K-means++ Initialization: The use of K-means++ for initializing the cluster centroids ensures better starting conditions for the algorithm, leading to faster convergence and improved final clustering outcomes.

The high Silhouette coefficient, combined with the visual separation of clusters, demonstrates that the proposed method is highly effective in identifying and clustering driving behaviors. This result is crucial for applications in real-time traffic monitoring, where accurately detecting aggressive driving can contribute to safer road conditions and better traffic management.

DISCUSSION AND CONCLUSIONS

The results of our study clearly demonstrate the effectiveness of a K-means clustering approach in identifying and distinguishing aggressive driving behaviors based on key driving features such as speed and acceleration. The high Silhouette coefficient of 0.84 indicates that our model successfully captures the inherent differences between normal and aggressive driving patterns, and the visual separation of clusters further supports the robustness of the proposed method.

One of the strengths of our approach lies in the feature extraction process, where we carefully selected features that are directly related to driving behavior. By analyzing metrics such as longitudinal and transverse speed, as well as acceleration variability, we were able to build a model that effectively separates different driving styles. This not only allows for accurate identification of aggressive driving behaviors but also provides valuable insights into how these behaviors differ from normal driving patterns.

However, several limitations should be acknowledged. First, while our approach performed well on the NGSIM dataset, the generalization of this model to other datasets remains an open question. Different traffic environments, vehicle types, and driving conditions may require additional tuning of the feature extraction and clustering process. Second, our method relies on a fixed number of clusters ($K = 2$), which limits the ability to capture more nuanced driving behaviors, such as mildly aggressive driving or defensive driving styles.

Looking forward, future work will focus on enhancing the generalization capability of the model across diverse datasets and traffic environments. This includes exploring adaptive clustering techniques, expanding the feature set to capture more complex driving behaviors, and developing real-time processing capabilities. Additionally, we plan to investigate the integration of this approach into real-world applications, such as Advanced Driver Assistance Systems (ADAS) and autonomous driving platforms, where real-time detection of aggressive driving could significantly improve road safety.

REFERENCES

Areen Alsaid, Lee, J. D., Noejovich, S. I., & Chehade, A. (2023). The Effect of Vehicle Automation Styles on Drivers' Emotional State. *IEEE Transactions on Intelligent Transportation Systems (Print)*, *24*(4), 3963–3973. https://doi.org/10.1109/tits.2023.3239880

Chen, X., Wang, Y., Sun, X., & Cai, Y. (2023). Unsupervised Cross-Scenario Abnormal Driving Behavior Recognition Using Smartphone Sensor Data. *IEEE Internet of Things Journal*, *11*(8), 14698–14709. https://doi.org/10.1109/jiot.2023.3344482

Chhabra, R., Verma, S., & C. Rama Krishna. (2017). A survey on driver behavior detection techniques for intelligent transportation systems. *International Conference on Cloud Computing*, *Confluence*(pp. 36–41). https://doi.org/10.1109/confluence.2017.7943120

Davis, G. A., Hourdos, J., Xiong, H., & Chatterjee, I. (2011). Outline for a causal model of traffic conflicts and crashes. *Accident Analysis & Prevention*, *43*(6), 1907–1919. https://doi.org/10.1016/j.aap.2011.05.001

Deng, Z., Chu, D., Wu, C., Liu, S., Sun, C., Liu, T., & Cao, D. (2022). A Probabilistic Model for Driving-Style-Recognition-Enabled Driver Steering Behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(3), 1838–1851. https://doi.org/10.1109/tsmc.2020.3037229

Ferreira, J., Carvalho, E., Ferreira, B. V., de Souza, C., Suhara, Y., Pentland, A., & Pessin, G. (2017). Driver behavior profiling: An investigation with different smartphone sensors and machine learning. *PLOS ONE*, *12*(4), e0174959. https://doi.org/10.1371/journal.pone.0174959

Lee, S. E., Erik, Wierwille, W. W., & Polytechnic, V. (2004, March). *A comprehensive examination of naturalistic lane-changes*. Bts.gov; United States. Department of Transportation. National Highway Traffic Safety Administration. https://rosap.ntl.bts.gov/view/dot/4191

National Highway Traffic Safety Administration. (2014). *National Motor Vehicle Crash Causation Survey: Report to Congress. | National Technical Reports Library - NTIS*. Ntis.gov. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB2009103608.xhtml

Saunier, N., & Sayed, T. (2008). Probabilistic Framework for Automated Analysis of Exposure to Road Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, *2083*(1), 96–104. https://doi.org/10.3141/2083-11