A Study of Bad Driving Behavior Based on Improved K-Means Clustering and Neural Network

Fujun Yang, Yi Zhang, Junpeng Chen, Qiang Yang and Zhqiang Zhou

ABSTRACT

With the popularization of the Internet of Vehicles technology, excavating useful information from a large amount of driving behavior data to evaluate the driver's safe driving behaviors in real time, accurately and efficiently, is of great significance to improve the safety management level and transportation efficiency of road transportation process. Based on the driver's driving behavior characteristics and driving industry management standards, this paper proposes a bad driving behavior evaluation method based on improved k-means clustering and neural network. It uses the improved k-means clustering method to select typical sample points from the characteristic parameters extracted from the vehicle GPS positioning platform. The Backpropagation neural network algorithm is designed to learn the clustering results, and the online classification evaluation of bad driving behavior is realized. It provides a new direction for the transportation vehicle management department to carry out safety management of the road transportation process.

1. INTRODUCTION

As people's demand for automobiles continues to increase, China's car ownership has been increasing year by year. Automobiles bring convenience to people, but at the same time, environmental pollution and traffic safety problems are becoming more and more prominent. In recent years, due to the government's increasing emphasis on information
technology in the transportation industry, more and more units and enterprises have begun to build satellite positioning systems for road transport vehicles. Satellite positioning systems, which collects driving behavior data of the automobile during driving to regulate the driver's driving behavior and ensure road traffic safety. Due to the above reasons, researching on traffic safety with bad driving behavior has attracted much attention. For example, using the vehicle trajectory data collected by the vehicle GPS module, a method for realizing the safety analysis of vehicle driving behavior is proposed, which provides a scientific basis for the bus traffic management department to evaluate and manage driving behavior [1]. In addition, for the satellite positioning data, a method for identifying the illegal driving behavior with anti-noise is proposed, which identifies the driver's overspeed, unruly driving and other illegal driving behaviors[2]. And an improved PWARX identification model based on two clusters is proposed to identify the driver's driving behavior[3].

By comparison, it is found that the main method for studying the bad driving behavior of the car is to evaluate the driver's driving behaviors by extracting the original driving data in the vehicle satellite positioning system and extracting the extracted feature data. However, this evaluation method tends to be subjective therefore not scientific. Thus, this paper proposes a method based on improved k-means clustering and neural network to evaluate poor driving behaviors. This method extracts the characteristics of bad driving behaviors from the original vehicle GPS positioning platform. Firstly, the improved k-means clustering algorithm is used to screen out the outliers, and the typical sample points are obtained as the training set. The neural network is used for learning. The Backpropagation neural network is used to construct the driving behavior evaluation grading model, and the driver's driving behavior is dynamically evaluated to improve the accuracy of the identification of poor driving behaviors and to provide a new method for evaluating bad driving behaviors.

**DRIVING BAHAVIORS EVALUATION PARAMETERS EXTRACTION**

The data in this paper comes from a GPS vehicle positioning platform, a total of 450 vehicles, with each vehicle containing 80,000 information. The vehicle GPS platform can record the satellite positioning data and the driver's daily driving information during the working of the vehicle in real time, including the basic information such as the latitude, longitude, engine state, steering angle, GPS speed, and mileage of the vehicle. The following pictures show the original driving data of the platform:

- (a) speed variation chart
- (b) mileage variation chart
1.1 Definition of Vehicle Samples Information

A sample of a driving vehicle $M_i$ is defined as follows:

$$M_i = (i, j, w_i, v_i, S_j, r_i)$$

(1)

Where $i$ is vehicle number, $j$ is a certain time the vehicle is traveling, $w_i$ is the longitude of the current vehicle, and $v_i$ is the latitude of the current vehicle. Where $v_i$ is the driving speed of the vehicle, and $r_i$ is the current direction angle of the vehicle, that is, the horizontal angle from the clockwise direction to the traveling direction from the north direction of the positioning point.

1.2 Construction of Driving Behaviors Evaluation Index System

While referring to the relevant literature on bad driving behavior evaluation, this paper considers the driver's driving behavior from three aspects: driving safety, energy saving and efficiency.[6] It establishes an evaluation index system for bad driving behaviors, as shown in Figure 2. A total of 11 first-level evaluation indicators for bad driving behaviors were extracted, and corresponding secondary evaluation indicators were extracted from these first-level indicators, as shown in TABLE I.

The travel time utilization and mileage percentage formulas are as follows:

$$\delta = \frac{\delta_{end} - \delta_{start}}{T}$$

(2)

$$\eta = \frac{T_{move}}{T_{move} + T_{stop}}$$

(3)

The travel time utilization rate refers to the proportion of the vehicle's travel time in the total travel time. See formula (2), where $\delta$ is the daily average mileage, $\delta_{start}$ is the mileage of the transport vehicle, and $\delta_{end}$ indicates the mileage at the end of the transport vehicle. The number $T$ is the total number of days.
TABLE I. BAD DRIVING BEHAVIOR EVALUATION INDEX.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>first-level evaluation indicator</th>
<th>secondary evaluation indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rapid acceleration</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>2</td>
<td>rapid deceleration</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>3</td>
<td>speed exceeding</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>4</td>
<td>fatigue driving</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>5</td>
<td>flameout</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>6</td>
<td>sudden change</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>7</td>
<td>extra long idle speed</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>8</td>
<td>idle preheating</td>
<td>cumulative duration , times</td>
</tr>
<tr>
<td>9</td>
<td>vehicle speed stability</td>
<td>standard speed difference</td>
</tr>
<tr>
<td>10</td>
<td>driving time utilization</td>
<td>travel time utilization</td>
</tr>
<tr>
<td>11</td>
<td>daily average mileage</td>
<td>mileage percentage</td>
</tr>
</tbody>
</table>

Figure 2. Evaluation index system diagram.

The mileage percentage refers to the percentage of the mileage traveled by the vehicle on the day of the total mileage. See formula (3), where \( \eta \) is the utilization rate of the vehicle, \( T_{move} \) is the length of travel of the transport vehicle, and \( T_{stop} \) is the length of rest of the transport vehicle.

1.3 Driving Behaviors Evaluation Parameters Extraction

Then, according to the relevant traffic road behavior norms promulgated by the state, combined with the actual operation of the driver on the vehicle platform, the second-level evaluation indicators are scored. There are three common scoring models:

\[
\begin{align*}
y_1 & = \begin{cases} 
100; & t \leq 480 \\
0; & t > 480 
\end{cases} \\
y_2 & = 100 - k_t n \\
y & = 0.5y_1 + 0.5y_2 
\end{align*}
\]

\[
\begin{align*}
y_1 & = \begin{cases} 
100; & 0 < \sigma < 20 \\
80; & 20 < \sigma \leq 40 \\
60; & \sigma > 40 
\end{cases} \\
y_2 & = 100 - k_\sigma n \\
y & = 0.5y_1 + 0.5y_2 
\end{align*}
\]
Where $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are coefficients, and the specific values of the coefficients are determined according to the actual operation of the driver of the vehicle platform. $t$ is the cumulative time and $n$ is the cumulative times.

According to the existing research direction [7], this paper adopts formula 4 for the evaluation of rapid acceleration, rapid deceleration, sharp change lane and speed exceeding behaviors, and adopts formula 5 for fatigue driving, flameout, super long idle speed and idle speed preheating driving behaviors. Formula 6 is adopted for the vehicle speed stability, the utilization of travel time and the average daily driving situation.

After obtaining the evaluation results of each driving behavior, this paper refers to Liu Yingji's method [4], and uses the analytic hierarchy process to obtain driving behavior evaluation parameters from three aspects: safety, energy saving and efficiency. The evaluation parameters are shown in Figure 3:

![Figure 3. Evaluation parameters.](image)

2. DRIVING BEHAVIORS EVALUATION ALGORITHM BASED ON IMPROVED K-MEANS CLUSTERING AND NEURAL NETWORK

2.1 Traditional K-means Clustering Algorithm

The traditional k-means clustering algorithm selects $K$ samples randomly from a given data set as input parameters. The n sample data $X = \{x_1, x_2, ..., x_n\}$ is taken as the data set, where $x_j = (x_{j1}, x_{j2}, ..., x_{jd})^T$ is the d-dimensional vector. The goal of the $k$-means algorithm is to find the set $C = \{c_1, c_2, ..., c_k\}^T$ of each cluster center. The sum of squared errors in the class is often used as the criterion function. The objective function is:
\[ E = \sum_{i=1}^{k} \sum_{x_j \in S_i} d(x_j, c_i) \]  

(7)

Where \( S_i \) is the \( i \)th sample set, \( d(x_j, c_i) \) is the distance between the sample data and the cluster center.

\[ d(x_j, c_i) = \| x_j - c_i \|^2 = \left( \sum_{i=1}^{n} | x_j - c_i |^2 \right)^{\frac{1}{2}} \]  

(8)

Where:

\[ c_i = \frac{1}{n_i} \sum_{x_j \in c_i} x_j \]  

(9)

\( c_i \) is the center position of the \( i \)th class, \( i = 1, 2, \ldots, k \). \( n_j \) is the number of sample data in class \( c_i \), and \( x_j \) is the sample data in table \( c_i \).

2.2 Typical Sample Selection

However, the traditional algorithm has certain shortcomings. It relies on the central point. The initial cluster center point is completely randomly selected and therefore can easily falls into the local optimal solution. At the same time, according to the average value of all samples in the cluster class, the cluster center is calculated. If the cluster contains more noise points and more isolated points, it will have a great influence on the mean value, and the clustering result will have a large deviation. In order to eliminate the influence of noise and isolated points on the clustering results during clustering, and to cluster in the data with noise space, this paper proposes an improved clustering algorithm based on \( k \)-means clustering and combines DBSCAN algorithm [5]. It filters out typical sample points in the sample set and removes points farther from the center point. The specific steps to improve the \( k \)-means clustering algorithm are as follows:

Step 1: For \( k \) sample data points, each sample point is used as an initial center point, and the distance of each other sample data to each center point is calculated by formula (8).

Step 2: According to the distance, the sample is assigned to the class closest to it. Recalculate the value of the cluster center according to formula (8).

Step 3: If the cluster center does not change or satisfies the objective function formula (7), the class set \( (C_1, C_2, \ldots, C_i) \) will be obtained, where \( C_i = (p_1, p_2, \ldots, p_n) \), \( n_i \) is the number of samples in \( C_i \).

Step 4: Calculate the distances of the two samples in class \( C_i \), respectively, to obtain the distance set \( L_i \), and find the minimum value \( \min(L_i) \) and the maximum value \( \max(L_i) \) in \( L_i \).

Step 5: Divide the interval \([\min(L_i), \max(L_i)]\) into \( N \) intervals, and the number of samples in each interval is counted to obtain a matrix \( \delta \).

\[ \delta = \begin{bmatrix} \alpha_1 & 2 & \ldots & N \end{bmatrix} \]  

(10)
Where \( N \) is the interval number of \( \min(L_i) - \max(L_i) \) and \( \alpha_i \) is the sample logarithm of the \( N \)th interval.

Step 6: Label the interval number with the largest number of samples as \( m \), and the corresponding neighborhood radius parameter \( \epsilon_i \) in \( C_i \) is calculated, that is, the center value corresponding to the distance of the \( m \)-th interval sample:

\[
Eps_i = \min(D_i) + m\Delta d_i - \frac{\Delta d_i}{2}
\]

\[
\Delta d_i = (\max(L_i) - \min(L_i))/m
\]

(11)

Step 7: Find all the sample sizes of the samples in class \( C_i \) in the neighborhood \( Eps_i \), and set the minimum value to \( \text{MinPts}_i \).

Step 8: Iteratively performs the above steps to obtain \( Eps \) and \( \text{MinPts} \) in each class, and the largest \( \text{MinPts} \) is used as the a parameter of all classes, and the corresponding parameters are \( (Eps_1, \text{MinPts}_1), (Eps_2, \text{MinPts}_2), \ldots, (Eps_i, \text{MinPts}_i) \).

Step 9: Randomly select any sample point to find the class to which it belongs. Under the condition of \( (Eps, \text{MinPts}) \), if the point is a core point, treat it as a typical sample point. If the point is a non-core point, mark it as noise point while removing it.

2.3 Backpropagation Neural Network Design

Backpropagation neural network belongs to a branch of artificial neural network. Since its introduction, it has been widely used in image processing, pattern recognition, machine learning and other fields. Backpropagation neural network is a multi-layer feedforward network with threshold and weight adjustment through backpropagation algorithm. The learning algorithm is the steepest descent method. Through the back dissemination of the output error, the network connection weight coefficient and the threshold information are continuously adjusted to minimize the square error of the neural network, thereby achieving the desired requirement.

This paper designs a three-layer Backpropagation neural network model, which consists of an input layer, an implicit layer, and an output layer. It takes the typical sample points obtained by the clustering algorithm as input. The hidden layer in the neural network is composed of 10 neurons, and the output layer is the corresponding driving behavior evaluation result. The number of output layer nodes in this paper is 1. In this paper, the
typical sample points selected are used as training samples, and a driving behaviors evaluation model[8] is obtained by Backpropagation neural network training. The driving behaviors are classified and evaluated by the model.

4. ANALYSIS OF EXPERIMENTAL RESULTS

Based on the actual vehicle GPS platform data, this paper sets the clustering category to 4, and uses the improved $k$-means algorithm for clustering. The clustering results are shown in Figure7. The black points in the figure are off-group points, and the remaining points are typical sample points selected. It can be seen that the outliers screened by the improved $k$-means clustering is better, and the category 5 is distributed at the outermost periphery, and has no typical features. Safety evaluation, rapid acceleration, rapid deceleration, and speed exceeding have a greater impact on driving safety. While energy conservation evaluation, the occurrence of idle behavior such as ultra-long idle speed and idle speed preheating will lead to a decrease in energy-saving evaluation indicators. Also, efficiency evaluation, driving conditions and time utilization have a great impact on driving efficiency.

![Figure 6. Cluster distance distribution.](image)

Category 1 belongs to the stable driving behavior in which the safety, energy saving and efficiency indicators are relatively stable. Category 2 is a more dangerous driving behavior with lower safety indicators. Category 3 is an inefficient driving behavior with low efficiency indicators. Category 4 is a non-energy-saving driving behavior with low energy-saving indicators[9].

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Safety score</th>
<th>Energy-saving score</th>
<th>Efficiency score</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79.25</td>
<td>78.1</td>
<td>73.2</td>
<td>stable</td>
</tr>
<tr>
<td>2</td>
<td>46.8</td>
<td>65.09</td>
<td>63.58</td>
<td>dangerous</td>
</tr>
<tr>
<td>3</td>
<td>86.65</td>
<td>84.68</td>
<td>42.02</td>
<td>inefficient</td>
</tr>
<tr>
<td>4</td>
<td>67.23</td>
<td>44.7</td>
<td>58.13</td>
<td>no energy-saving</td>
</tr>
</tbody>
</table>
According to the improved clustering results of $k$-means algorithm, this paper uses the typical samples selected as input variables to design a three-layer Backpropagation neural network, using 70% of the data set as the training set, 15% of the data as the verification set, and 15% of the data as the test set. The number of hidden neurons is 10, the number of output layer nodes is 1, and the output variable is the category of vehicle driving behavior. Neural network training effect Figure 8, uses the trained neural network model to achieve online classification and evaluation of driving behavior.

5. CONCLUSIONS

According to the driving data of a GPS vehicle platform, this paper constructs the corresponding evaluation index system and model, and extracts the corresponding bad driving behaviors evaluation parameters. The improved $k$-means clustering algorithm is used to screen the typical samples of driving behaviors, and the discrete points are removed. According to the clustering results, the representative samples are put into the Backpropagation neural network model for training, and the model trained by Backpropagation neural network is used to realize the online classification evaluation of bad driving behaviors. Effectively identify bad driving behaviors, and classify bad driving behaviors into four types: stable, relatively dangerous, inefficient, and non-energy-saving. From the perspective of data mining, it provides a scientific and feasible new direction for the evaluation of bad driving behaviors.

ACKNOWLEDGEMENT

Shengte Cup Research Project of Southwest Petroleum University, 2019-09-007.

REFERENCES