Algorithm of Vehicle’s Data Cleaning and Monitoring

Hao Li¹, Di Ma¹, Zhijiang Yan¹, Jie Fu¹, Mengjie Zeng²,* and Wuzhida Bao²

¹ State Grid Jiangxi Maintenance Company, Nanchang, China
² School of Information Engineering, Nanchang Hangkong University, Nanchang, China

*Email: 547155474@qq.com

Abstract. The aim of this paper is to screen out abnormal data which is caused by human factors and natural factors. We proposed an algorithm of vehicle’s data cleaning and monitoring. First, the valid data is filtered out by an improved DBSCAN method, through data analysis, and then get the threshold range. After that, it screens out the abnormal data in the valid data through the threshold range. Finally, the abnormal data is classified and counted according to the factors which was stipulated by the enterprise. The results show that the proposed algorithm can simpler and faster to process the abnormal data than the other similar algorithm.

1. Introduction

Nowadays, employees use vehicles more frequently, which makes the abnormal data that car produced become more numerous. Usually, these factors that cause vehicle anomalies include fuel consumption, mileage, trajectory, expense and so on. The abnormal data cannot intuitively reflect the developments prospects of enterprises vehicles, and might even influences the data processing of other parts of the enterprises (such as finance department, human resources department and marketing department) making substantial variation even mistakes to later results.

Fuel, consumption, travel trajectory and maintenance are the research objects of this algorithm. This algorithm is designed by following steps: data cleaning, getting the threshold range by data analysis and artificial experience, classifying and counting abnormal data.

Compared with other algorithms, most common situations are that filtrate out the abnormal data and delete them directly, which may conduct some loss. But this algorithm keeps and classifies those abnormal data. Using the vehicle’s real data set, systematic experiments and analysis were performed to verify the effectiveness of the algorithm.

2. Method

The algorithm includes data cleaning algorithm [1] and data monitoring algorithm. According to the characteristics of small amount of abnormal data set and discrete distribution, we can find the range of the data set, which is used to troubleshoot abnormal data. This algorithm not only can filter the data of vehicle’s information but also can identify and classify anomalies in the data.

2.1. Data Cleaning Algorithm

The data cleaning algorithm realises the preliminary examination and correction of the data. It aims to delete repeating data, correct wrong data, and fill in missing data. This algorithm includes data
deduplication algorithm, data error value processing algorithm and data missing value processing
algorithm.

2.1.1. Data Deduplication Algorithm
For exception problems in original trajectory data, the algorithm sorts the data by timestamp, the
repeat data is deleted by the same timestamp, and the other data is deleted by different longitude and
latitude in the same timestamp.

When the other data is deleted by different longitude and latitude in the same timestamp, this
algorithm will calculate the distance from multiple point \( B_i(x_i, y_i), B_i \in B, B = \{ B_1, B_2, \cdots, B_n \}, i \in (1, 2, \cdots, n) \) to \( A \) at the same time by using \( A(X_1, Y_1) \) which is found in the first \( n \) seconds of the same
time point. The distance formula is shown in (1):

\[
dist(A, B_i) = \sqrt{(X_1 - x_i)^2 + (Y_1 - y_i)^2}
\]

When \( dist(A, B_i) \) is the minimum in \( B \), the algorithm will take the data at point \( B_i \) as the correct
time at that point in time, then it will discard other \( B_i \) directly.

For other raw data, this algorithm uses a field matching algorithm to deduplicate the original data,
and reasonably removes duplicate data according to the uniqueness of the license plate number and
timestamp.

2.1.2. Data Error Value Processing Algorithm
After data deduplication, inevitably, some erroneous data which because of nonstandard data or
abnormal collectors. Thus, we need to undergo the second data cleaning.

There are many irregularities in the data of vehicle’s information, such as the nonstandard letters in
the license plate number.

In the vehicle’s trajectory data, there are many problems of longitude and latitude deviation. The
algorithm proposed improved DBSCAN method [2] and local mean correction method. The following
are related definitions for designing the algorithm:

- Vehicle’s trajectory data set: Set the vehicle’s trajectory set of a vehicle on one day as \( D = \{x_1, x_2, \cdots, x_i\}, i \in (1, n) \), \( n \) is the number of objects in the \( D \), \( x_i \) contains longitude and
  latitude data which is indicated as \( x_i(\text{Lon}_i, \text{Lat}_i) \).
- Partial trajectory data set: we find \( \delta \) objects on both sides of \( D \) with \( x_i \) as the center, then we
  mark this set as \( T(x_i) = \{t_1, t_2, \cdots, t_j\}, j \in (1, 2, \delta) \). \( |T(x_i)| \) represents the number of data
  objects in \( T(x_i) \).
- Eps neighbour data set: we make \( x_i \) as the center, find the set of objects in the \( T(x_i) \) whose
distance between \( t_j \) and \( x_i \) is less than \( Eps \), the set of objects as \( L_{Eps}(x_i) \), \( |L_{Eps}(x_i)| \)
  represents the number of data objects in \( L_{Eps}(x_i) \). The description can be expressed as the
  formula (2):

\[
L_{Eps}(x_i) = \left( \left| \right\{ x_i \in D \mid dist(x_i, t_j) \right\} \leq Eps \right)
\]

\( dist(x_i, t_j) \) represents the distance formula of two data objects, and this algorithm uses
formula (1) for calculation. We set the minimum number allowed in the \( Eps \) neighbour data
set as \( Minpts \).
- Core object set: If \( |L_{Eps}(x_i)| \) is not less than \( Minpts \), we will cluster \( x_i \) into Core object set.
The core object set is denoted as \( C \).
- Outlier object set: If \( |L_{Eps}(x_i)| \) is less than \( Minpts \), we will cluster \( x_i \) into Outlier object set.
The outlier object set is denoted as \( O \). The description can be expressed as the formula (3):

\[
\begin{align*}
x_i = \begin{cases} 
\in C, & |L_{Eps}(x_i)| \geq Minpts \\
\in O, & |L_{Eps}(x_i)| < Minpts
\end{cases}
\end{align*}
\]
• Local mean: For \( x_i \) in \( O \), find the partial trajectory data set \( T(x_i) \). We set the sum of \( T(x_i) \) as \( \sum(T(x_i)) \). Let \( \gamma \) as the average of \( T(x_i) \). The description can be expressed as the formula (4):

\[
\gamma = \sum(T(x_i)) * (|T(x_i)|)^{-1}
\]  

The specific algorithm process is:

• Input \( D = \{x_1, x_2, \cdots, x_n\} \).
• Step 1. Traverse the data set \( D = \{x_1, x_2, \cdots, x_n\} \), and find every \( T(x_i) \)
• Step 2. Set the value of \( Eps \) and \( Minpts \) by data analysis.
• Step 3. Traverse the set again, find every \( L_{Eps}(x_i) \). For \( \forall x_i \in D \), if \( |L_{Eps}(x_i)| \geq Minpts \), we add \( x_i \) to the core object set \( C = C \cup \{x_i\} \). If \( |L_{Eps}(x_i)| < Minpts \), we add \( x_i \) to the outlier object set \( O = O \cup \{x_i\} \).
• Step 4. Traverse the data set \( O \), find every \( T(x_i) \), replace the \( x_i \) data with the \( \gamma \) data.
• Output. Result set is \( R = \{r_1, r_2, \cdots, r_n\} \). The outliers which has corrected by local average method are in this set \( R \).

The improved DBSCAN method [3] and local mean correction method can screen and correct the wrong data. They can reduce the impact of the subsequent vehicle’s trajectory monitoring algorithm.

2.1.3. Data Missing Value Processing Algorithm
When vehicles were dispatched, we will get a set of current location information which GPS collects every 15s. However, because of factors such as geographic location or signal strength, GPS may not collect some data. Such as Table 1.

| Plate number | Time               | Longitude  | Latitude  |
|--------------|--------------------|------------|-----------|
| GANA0M593    | 2020-03-13 12:29:55| 115.902325 | 28.594455 |
| GANA0M593    | 2020-03-13 12:29:55| 115.90237  | 28.594428 |
| GANA0M593    | 2020-03-13 12:29:55| 115.902411 | 28.59439  |
| GANA0M593    | 2020-03-13 12:30:55| 115.9025718| 28.594436 |

The vehicle is missing a set of data at 2020-03-13 12:30:25 which sets as \( x \). In order to avoid errors in the following mileage statistics, this algorithm takes the first \( n \) and last \( n \) data of the missing point \( x \) as \( M = \{m_1, m_2, \cdots, m_{2n}\} \). After that, we set the average value of \( M \) as \( \gamma \). Then we use \( \gamma \) to fill in the missing data of \( x \). We can calculate averaging by the formula (5):

\[
\gamma = \left(\sum_{i=1}^{2n} m_i\right) * (2n)^{-1}
\]  

(5)

2.2. Data Monitoring Algorithm
Algorithm for processing abnormal data is mainly divided into threshold range definition and abnormal data processing.

2.2.1. Threshold Range Definition
The threshold definition algorithm got a reasonable threshold range which is based on artificial floating value and annual data. For example, the process of calculating the upper threshold of average fuel consumption of \( n \) km is shown below:

We set \( X = \{x_1, x_2, \cdots, x_n\} \) as the data set of average fuel consumption of \( n \) km, \( x_i (i = 1, 2, \cdots, m) \) indicates the fuel consumption data for the \( i \) quarter. Then we define the cost of the fuel...
consumption in \( x_i \) is \( \text{expense}_i \) the mileage in \( x_i \) is \( \text{mileage}_i \). We can calculate the average by the formula (6):

\[
x_i = (\text{expense}_i \times (\text{mileage}_i)^{-1}) \times n \tag{6}
\]

We set the average fuel consumption fee is capped as \( thr \), the artificial floating value is \( \delta \). We can calculate \( thr \) by the formula (7):

\[
\text{thr} = \left( \sum_{i=1}^{n} x_i \right) 
\times \frac{1}{n} + \delta \tag{7}
\]

Thus, we can calculate the threshold range is \((0, \text{thr})\).

2.2.2. Abnormal Data Processing

We divide the algorithm of abnormal data processing into fuel consumption monitoring [4], mileage monitoring, and maintenance supervision.

(1) Fuel consumption monitoring

This algorithm includes monitor whether the fuel consumption data of the fuel consumption data during a certain period is within the budget. Such as judging whether the quarterly average fuel consumption of n km is normal is shown below:

The set of the quarterly average fuel consumption of n km is \( O = \{o_1, o_2, \cdots, o_m\} \). \( o_i(i = 1, 2, \cdots, m) \) indicates the fuel consumption data for the \( i \) quarter. We define the cost of the fuel consumption in \( o_i \) is \( \text{expense}_i \), the mileage in \( o_i \) is \( \text{mileage}_i \). We can calculate \( o_i \) by the formula (8):

\[
o_i = (\text{expense}_i \times (\text{mileage}_i)^{-1}) \times n \tag{8}
\]

We set the average fuel consumption fee is capped as \( \text{thr} \), \( N \) is the data set for normal data, \( A \) is the data set for abnormal data. We can categorize \( O \) by the formula (9):

\[
O \in \{N, 0 \leq o_i \leq \text{thr} \cup A, o_i > \text{thr}\} \tag{9}
\]

When \( x_i > \text{thr} \), \( o_i \) is the abnormal data, classify \( o_i \) into \( A \). Otherwise, classify \( o_i \) into \( N \).

(2) Mileage monitoring

This algorithm includes monitoring whether the actual mileage of a vehicle is reasonable and the vehicle refuels within a certain area.

To calculate whether the actual mileage of the vehicle is reasonable, this algorithm refers to the Shortest path algorithm which was based on improved Dijkstra algorithm [5] and Earth formula [6]. calculating it is shown below:

- First, we define a car to drive from the starting point A to the end point B. Then, the set of key points between A and B is \( X = \{(\text{Lon}_{i1}, \text{Lat}_{i1}), (\text{Lon}_{i2}, \text{Lat}_{i2}), \cdots (\text{Lon}_{in}, \text{Lat}_{in})\} \), where \( n \) is the number of key points and \( \text{Lon}_{i1}, \text{Lat}_{i1} \) represent longitude and latitude respectively.

\[
C = \sin(\text{Lat}_{i1}) \sin(\text{Lat}_{i+1}) \cos(\text{Lon}_{i} - \text{Lon}_{i+1}) + \cos(\text{Lat}_{i}) \cos(\text{Lat}_{i+1}) \tag{10}
\]

\[
\text{Distance}(\text{Lon}_{i}, \text{Lat}_{i}, \text{Lon}_{i+1}, \text{Lat}_{i+1}) = \arccos(C) \tag{11}
\]

- Where the longitude of point \( i \) is \( \text{Lon}_{i} \) and the latitude is \( \text{Lat}_{i} \); longitude of point \( i + 1 \) is \( \text{Lon}_{i+1} \) and the latitude is \( \text{Lat}_{i+1} \). Use formula (10) to calculate the radian \( C \) from point \( i \) to point \( i + 1 \), and get the true distance \( \text{Distance} \) between two points by formula (11). Therefore, the calculation formula (12) of cumulative true distance \( \beta \) is:

\[
\beta = \sum_{i=1}^{n-1} \text{Distance}(\text{Lon}_{i}, \text{Lat}_{i}, \text{Lon}_{i+1}, \text{Lat}_{i+1}) \tag{12}
\]

- Next, we define the optimal mileage which is calculated by GPS as \( \alpha \), and upper limit as \( \delta \) which is the difference between the real mileage and \( \alpha \). Then we define \( N \) as the data set for normal data. Therefore, we set \( A \) as the data set for abnormal data. According to the formula (13):
Finally, we take this formula to judge whether the true mileage data in the period is abnormal. If $\beta - \alpha \leq \delta$, $X$ is the abnormal data, $X$ is classified into $A$, otherwise, $X$ is classified into $N$.

To calculate whether the vehicle refuels within a certain area, this algorithm uses the Earth formula. 

calculating it is shown below:

- First, we define the upper limit of distance as $\varepsilon$, $M(Lon_2, Lat_2)$ is the longitude and latitude given by vehicle’s trajectory information, $N(Lon_2, Lat_2)$ is the longitude and latitude given by GPS.
- Therefore, we set the distance between $M$ and $N$ as $\text{dist}(M, N)$. We can categorize $M$ by the formula (14):

$$
M \in \begin{cases} 
N, & \text{dist}(M, N) \leq \delta \\
A, & \text{dist}(M, N) > \delta 
\end{cases}
$$

Finally, we take this formula to judge whether the true mileage data in the period is abnormal. If $\text{dist}(M, N) \leq \delta$, $M$ is the abnormal data, classify $M$ into $A$, otherwise, classify $M$ into $N$.

3. Experiment

A vehicle with a license plate number of GANa0m593 is collected. The maintenance cost on March 13, 2020 is 451 yuan, and 308 vehicle’s trajectory data were generated. After the vehicle’s data cleaning and monitoring algorithm, the following data can be clearly observed:

| Plate number | Original data | Cleaned data |
|--------------|---------------|--------------|
| GANA0m593    | 451           | 451.00       |
| Longitude    | NAN           | 115.902425   |
| Latitude     | NAN           | 28.594425    |

We can see that the cleaned data is more normalised, the missing value is reasonably filled with the average value, which reduces the error for the subsequent threshold range judgment.

Comparison table of fuelling points fed back by vehicles and gas station locations fed back by GPS:

| sample | GPS location | Vehicle’s location | Plate number | distance |
|--------|--------------|--------------------|--------------|----------|
| 1      | (114.916163, 27.033216) | (114.916163, 27.034816) | GANA0M593 | 187m     |
| 2      | (114.916163, 27.033216) | (114.916163, 27.038724) | GANA0M593 | 607m     |
| 3      | (114.916163, 27.033216) | (114.916190, 27.033289) | GANA0M593 | 96m      |

We can see that the distance of the license plate number of GANA0M593 in sample 2 is greater than 300m, which is recorded as abnormal data.

Comparison of the original data of vehicle’s trajectory latitude and longitude and the data after cleaning:
Figure 1. Original scatter plot of longitude and time.

Figure 2. Modified scatter plot of longitude and time.

Figure 3. Original scatter plot of latitude and time.

Figure 4. Modified scatter plot of latitude and time.

Figure 1 reflects the situation of original longitude and time, and Figure 3 reflects the situation of original latitude and time. We can see that Figure 1 and Figure 3 do not conform to the kinematic transformation. While after data cleaning, Figure 2 reflects the situation of longitude and time after data cleaning, and Figure 4 reflects the situation of original latitude and time after data cleaning. Obviously, Figure 2 and Figure 4 are more comfortable with the actual vehicle latitude and longitude transformation process.

4. Conclusion
The paper proposed the algorithm of vehicle’s data cleaning and monitoring in order to deal with the impact of abnormal vehicle’s data on enterprises. The algorithm makes two principal contributions: First, it optimizes the business management model, enhances the company's ability to macro-control vehicles, and reduces the need for labour. Second, through accurate calculation formulas, the average value of massive data is got, allowing companies to view abnormal data more intuitively and analyse the causes of abnormal data which can eliminate errors caused by human error. Using the DBSCAN method and data analysis to define various threshold ranges, to filter out abnormal data generated by vehicles in a certain period.

The experimental results show that our algorithm can accurately classify and filter abnormal data, while normalizing normal data. In the future, we can use abnormal data and daily corporate data for data mining, such as correlation analysis algorithms and clustering algorithms. In order to reduce the cost of manual analysis and improve the efficiency of enterprise management.
References

[1] Tian D et al. 2019 An Effective Fuel-Level Data Cleaning and Repairing Method for Vehicle Monitor Platform 2019 IEEE Transactions on Industrial Informatics 15 410-422

[2] Zhang Y, Feng D, Zhang R and Geng N 2017 Multi-stage optimization of taxi service stations location using GPS data 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, pp. 316-322.

[3] Angkhawey U and Muangsin V 2018 Detecting Points of Interest in a City from Taxi GPS with Adaptive DBSCAN 2018 Seventh ICT International Student Project Conference (ICT-ISPC), Nakhonpathom, pp. 1-6.

[4] Moreno-Gutiérrez J, Pájaro-Velázquez E, Amado-Sánchez Y, et al. 2019 Comparative analysis between different methods for calculating on-board ship’s emissions and energy consumption based on operational data Science of the Total Environment 650 575-584

[5] Qing G, Zheng Z and Yue X 2017 Path-planning of automated guided vehicle based on improved Dijkstra algorithm 2017 29th Chinese Control and Decision Conference (CCDC), Chongqing, pp. 7138-7143.

[6] Zhu J 1994 Conversion of Earth-centered Earth-fixed coordinates to geodetic coordinates 1994 IEEE Transactions on Aerospace and Electronic Systems 30 957-961