Research Article

Highway Event Detection Algorithm Based on Improved Fast Peak Clustering

Lili Pei ¹, Zhaoyun Sun ¹, Yuxi Han, ¹ Wei Li ¹ and Huaixin Zhao²

¹School of Information Engineering, Chang'an University, Xi'an 710064, Shaanxi, China

²Shaanxi Provincial Department of Transportation, Xi'an 710000, Shaanxi, China

Correspondence should be addressed to Zhaoyun Sun; zhaoyunsun@126.com

Received 10 July 2020; Revised 5 January 2021; Accepted 13 February 2021; Published 20 February 2021

Academic Editor: Frederic Kratz

Copyright © 2021 Lili Pei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the mining of traffic events based on large amounts of highway data, this paper proposes an improved fast peak clustering algorithm to process highway toll data. The highway toll data are first analyzed, and a data cleaning method based on the sum of similar coefficients is proposed to process the original data. Next, to avoid the shortcomings of the excessive subjectivity of the original algorithm, an improved fast peak clustering algorithm is proposed. Finally, the improved algorithm is applied to highway traffic condition analysis and abnormal event mining to obtain more accurate and intuitive clustering results. Compared with two classical algorithms, namely, the k -means and density-based spatial clustering of applications with noise (DBSCAN) algorithms, as well as the unimproved original fast peak clustering algorithm, the proposed algorithm is faster and more accurate and can reveal the complex relationships among massive data more efficiently. During the process of reforming the toll system, the algorithm can automatically and more efficiently analyze massive toll data and detect abnormal events, thereby providing a theoretical basis and data support for the operation monitoring and maintenance of highways.

1. Introduction

With the gradual improvement in the highway network and the arrival of the information era, the data generated by intelligent toll systems [1], intelligent road detection systems, and other facilities have formed a certain scale [2]. The highway network has become increasingly complex, and the probability of abnormal events has also increased [3]. Additionally, their occurrence is inevitable, and related data may also contain some unique information [4]. The accurate and efficient identification of abnormal events conveyed by toll data is thus of great significance to the statistics of toll-evading vehicles and the upgrade of intelligent detection equipment [5, 6].

Abnormal highway traffic events include traffic accidents and traffic incident [5, 7]. A traffic accident refers to an abnormal traffic condition in which a vehicle crashed [8], while a traffic incident refers to abnormal conditions such as vehicle breakdown, expired parking, equipment failure, and toll evasion [9]. The detection of this type of abnormal event

has always been the key to highway electromechanical systems but hard to find. Before the emergence of data mining analysis methods, traffic administrative departments mainly relied on simple sampling and statistical methods to detect events and analyze highway traffic conditions, which resulted in substantial investments and poor application results.

In the course of past decade, many abnormal event detection algorithms have been developed based on computational intelligence [10]. Jin et al. [11] proposed a new technique for the detection of abnormal highway events using a constructive probabilistic neural network (CPNN), which is tested in the situation of constantly changing traffic environment of stations. Wu [12] and Sun et al. [13], respectively, carried out the detection of abnormal highway events and abnormal states caused by highway events based on a support vector machine (SVM) classifier. Xiao [14] proposed an ensemble learning method that first trains individual SVM and k -nearest neighbor (KNN) models and implements a strategy to combine them to achieve better

final outputs. Ye et al. [15] build an accident detection algorithm based on data mining and parameter sensitivity analysis methods in a rural road condition. However, the network parameter combinations of these methods are too complicated to obtain and have time-consuming steps to process data. Li et al. [16] constructed a discrete DBN network structure by selecting the factors that have great influence on highway operation and established a real-time highway risk evaluation model. Nevertheless, this method is able to detect only small abnormal events. Li et al. [17] proposed a method that takes into account the traffic ratio at the entrances and crossway in the network. They established a simulation algorithm to describe the movement of the vehicles, and the proposed method was found to be capable of effectively detecting abnormal highway events. This method may have a high accuracy, but only compatible for the low-level real-life implementation.

As typical unsupervised algorithms, clustering algorithms [18–20] can effectively mine outliers in data. They have been widely used in the medical, military, construction, and other fields [21–23], and previous work has provided a good foundation for the research of the detection of abnormal highway events. Huang et al. [24] proposed an average clustering algorithm for detecting highway toll fraud. The produced algorithm can meet the demands of toll collection. Abualigah [25, 26] used an improved krill herd algorithm (MMKHA) and the feature selection method with the particle swarm optimization (PSO) algorithm (FSPSOTC) [27] to cluster text documents. This type of algorithm is a relatively new biological heuristic algorithm that seeks potential solutions by simulating the behavior of a group of animals to avoid falling into a local optimum [28].

To more specifically detect abnormal events via data mining, this paper proposes an algorithm for the detection of abnormal highway events based on improved fast peak clustering. The algorithm uses toll data collected from highway toll station. Compared to many methods in the past using data from different detectors, the proposed algorithm requires less hardware facilities support and has better economic benefits. Meanwhile, the proposed algorithm automatically generates the parameters. The improved algorithm is used to analyze and verify traffic conditions, detect abnormal events, and identify problems such as vehicle overload, equipment damage, and network failure. It has high recognition accuracy of abnormal events and provides data support for highway operation and management.

2. Methodology

The primary methodology undertaken in this study is presented in Figure 1. The process included descriptive data analysis, data cleaning, the improvement of clustering algorithms to detect traffic events, and comparison with the results of three other clustering methods. Optimized results were obtained by improving the fast peak clustering algorithm.

2.1. Fast Peak Clustering Algorithm. The fast peak clustering algorithm is a new clustering algorithm proposed in *Science* in June 2014 [29]. This algorithm overcomes the deficiencies

of the data requirements of general clustering algorithms and can process data of any shape. While expanding the scope of applicable data, it also avoids the need for a large amount of calculation. Moreover, the performance results on various standard data sets have also verified the effectiveness of the algorithm.

The density-based fast peak algorithm is based on the assumption that the cluster center feature has a high local density and a large distance from high-density points; this assumption is the basis of the clustering process. In this process, the number of classes is intuitively visible, and outliers can be visually presented to facilitate accurate analysis. The algorithm uses the ρ - δ decision graph to achieve the selection of cluster centers. To be specific, ρ_i , the number of points whose distance from a given point is less than the cut-off distance, is the local density of certain data. The definition of ρ_i is given by the following equation:

$$\rho_i = \sum_j x(d_{ij} - d_c), \quad (1)$$

where i and j are two different data points. When i is less than 0, $\chi(i) = 1$; otherwise, $\chi(i) = 0$. Moreover, d_{ij} is the distance between point i and point j , d_c is the cut-off distance, which is set by the user, and δ_i is the distance between high-density points of certain data, and its definition is as follows:

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}). \quad (2)$$

For the point with the highest density, δ_i is defined as the maximum distance between this point and all other points, given as follows:

$$\delta_i = \max_j (d_{ij}). \quad (3)$$

After calculating the two quantities of each point, all points are visualized and output with ρ and δ as two dimensions, and the output graph is called a decision graph. Generally, points with high ρ and δ values are selected as cluster centers, points with low ρ and high δ values are regarded as noise points, and points with relatively high ρ and low δ values are selected as cluster center points within the cluster.

After finding the cluster center, the number of classes is determined. It is then necessary to reasonably allocate the remaining points so that they can be allocated to the most suitable cluster. The distribution principle of these points is that each remaining point will be divided into clusters with the closest points with higher densities. This operation is performed in a single step until all points are assigned to a corresponding class.

2.2. Improved Fast Peak Clustering Algorithm. During the process of cluster center selection, the fast peak clustering algorithm generates a ρ - δ decision graph for the selection of cluster centers based on the calculated local density ρ and δ . The point at which the two values are larger is the cluster center. However, for large-scale, high-speed toll data,

FIGURE 1: General implementation of the traffic event detection method undertaken in this study.

manual selection is characterized by high subjectivity and instability, and the most accurate cluster center cannot be selected by users who do not understand the principle of the fast peak clustering algorithm. Moreover, it is difficult to select accurate cluster centers for those data sets whose decision graphs are complicated and whose cluster center distributions are unclear. The clustering algorithm has a high dependency on the center; once the cluster center has a deviation, the subsequent allocation and optimization of noncentral points and the discovery of noise points will be seriously affected, which will affect the analysis of toll data. In view of the shortcomings of fast peak clustering, which requires the manual selection of cluster centers based on the decision graph, this section proposes an improved fast peak clustering algorithm that can automatically determine the centers.

It can be seen from Section 2.1 that a point is selected as the cluster center only when the two values of ρ_i and δ_i are large enough. Therefore, γ_i is introduced as the judgment standard, and its definition is as follows:

$$\gamma_i = \rho_i \times \delta_i. \quad (4)$$

The larger the value of γ_i , the more likely that it is the cluster center; in other words, the value of γ_i of the cluster center must be large. Therefore, the improved steps are as follows. First, the point with the larger value of γ_i is selected, and the real cluster center is then selected from it. The values of γ_i are then sorted in descending order to obtain a descending graph of γ_i in preparation for the subsequent steps.

Let the critical point P be the point at which $\gamma_{[1-P]}$ and $\gamma_{[P-n]}$ change the most. Using the slope of the value of γ_i in descending order to represent the degree of change, the definition of P is given as follows:

$$P = \max\{i \mid |k_i| - |k_{i+1}| \geq \beta, i = 1, 2, \dots, n-2\}, \quad (5)$$

$$\beta = \frac{\alpha(j)}{(n-2)}, \quad (6)$$

$$\alpha(j) = \sum_{j=1}^{n-2} ||k_j| - |k_{j+1}||, \quad (7)$$

where k_i represents the slope of the line segment between the i -th point and the $i+1$ -th point, β represents the average value of the slope difference, as given by equation (6), $\alpha(j)$ represents the phase in the descending order of γ_i values, the sum of the slope difference between two adjacent points is given by equation (7), and i is the point that is either the critical point or the point whose slope difference is greater than or equal to the mean value β and has the largest sequence number.

The cluster center may then exist in the range expressed by equation (8), and the points in this range are called pseudo-centers:

$$sp_q = \{q \mid \gamma_q \geq \gamma_p, q = 1, 2, \dots, p\}. \quad (8)$$

The first pseudo-center in the same area is considered as the cluster center, and the distances from other pseudo-centers to the cluster center are judged. If the distance is less than the cut-off distance d_c , this pseudo-center will be removed; if the distance is greater than d_c , it will be used as another cluster center.

After the cluster center is determined, each remaining point is divided into clusters with the closest higher density points until all the points are assigned to the corresponding clusters. A boundary area is then determined for each class. The definition of the boundary area is that the distance of a point assigned to a certain type of cluster from the point of another type of cluster is less than the cut-off distance d_c . The point with the highest density in the boundary area of each cluster is then determined, and its density is ρ_b . Traversing each point in the cluster, the points whose densities are greater than ρ_b are categorized as cluster points; otherwise, they are categorized as noise points.

3. Materials

3.1. Raw Data. The data used in this study are the toll data of a provincial highway in China from 2016 to 2017. The data include 27 items of vehicle information, and each piece of data has a unique ID number. The detailed information is presented in Table 1.

Every vehicle driving on the highway is issued an IC card that contains detailed information about the inbound and outbound stations, and the tolls are calculated automatically.

TABLE 1: Original data features.

Feature	Comment	Example
ID	Serial number	57 261 916
CardNo	Card number	52 011 328 220 200 196 796
ICCardNo	IC card number	52 011 328 220 200 196 796
LastBalance	Balance after consumption	2887.00
OutTime	Outbound time	2017/1/31 00:00:00
OutLoad	Total weight of outbound vehicle	35
OutStationName	Name of the outbound station	3
InTime	Entry time	2017/1/30 23:42:48
InLoad	Total weight of entering vehicle	35
InStationName	Name of the entering station	2
VehiclePlate	Vehicle plate number	** 4820
VerificationCode	Verification code	
CreateTime	Creation time of the record	2017/1/31 14:58:47
SettlementTime	Settlement time of the record	
ShiftTime	Time when the balance updates	2017/1/31 14:58:47
State	State	0
HandleTime	Constant time verification	2017/1/30 00:00:00
CardType	Type of card	22
OutBusiNo	External transaction serial number	
ExitType	Turnover type	
BatchNo	Batch number	
VehicleType	Type of vehicle	1
VehicleTypeColor	Color of vehicle plate	0
TransferType	Transaction type	0
TransferState	Transaction state	0
TransferMoney	Transaction expenditure	9.50
TransferTime	Time of transaction	

Instead of recording the lane type, vehicle type information was collected to support the analysis.

There were many characteristics in the original data table that were not utilized. After communicating with experts in the transportation department, the focus of this research was placed on the following features: LastBalance, Credit (yuan), OutTime, OutLoad (100 kg), OutStationName, InTime, InLoad (100 kg), and InStationName. The data were integrated on this basis, and a new data set was created for subsequent analysis. Some data are exhibited in Table 2.

The problems with the data can be divided into the following four categories:

- ① Wrong data. This includes mutated data that do not conform to common sense, which are usually caused by equipment failure, line failure or transmission error, or failure to comply with driving specifications.
- ② Missing data. In this case, there are no corresponding data from the moment at which data should have been collected. Usually, some information is missed due to dense vehicles, personnel operation errors, equipment failure, and other reasons.
- ③ Redundant data. This includes duplicated toll records and is usually due to network and system failures or software bugs; the lane machine will upload data that have already been uploaded to the provincial (sub) center.
- ④ Abnormal data. It should be noted that abnormal data are not all erroneous data, but data points

randomly appear in traffic data. This also coincides with the randomness of traffic data. Abnormal data are usually caused by abnormal events, such as traffic accidents and equipment failures on the highway during a certain period of time. The traffic flow in this case reflects the real traffic flow under special traffic events and is of great significance for the analysis and excavation of traffic operation conditions.

3.2. Data Cleaning. As noted in Section 3.1, the collected toll data often generate abnormal data due to information entry, operation records, transmission errors, equipment damage, system failures, and clock asynchronization. In this case, the data do not fully reflect the actual traffic conditions, so it is necessary to clean the high-speed toll data [30]. Due to the multidimensional characteristics of the toll data, if the traditional distance-based outlier detection method is used to clean the toll data separately, the following problems may arise: the prolonging of the calculation time, the cleaning of correct data, and the disregard of abnormal data. These problems are a result of the overlooking of the correlations between data dimensions. To solve them, by combining the characteristics of the original toll data and the reasons for abnormal data, an outlier detection algorithm based on the sum of similar coefficients is proposed.

The data set $X = \{x_1, x_2, x_3, \dots, x_n\}$ is set as the object to be detected, where each object has m attributes, given as follows:

TABLE 2: Toll data samples.

ID	Last balance (yuan)	Credit (yuan)	OutTime	OutLoad (100 kg)	OutStation name	InTime	InLoad (100 kg)	InStation name
50968476	2345.5	1092.5	06:25.	40	16	15:32	30	5
50968477	2302.75	42.75	06:28.	40	15	2016/10/26 09:46: 50.00000:0	40	16
50968478	3645.04	14.25	57:14.	150	5	49:32	100	4
50968479	3635.54	9.5	10:58.	82	3	2016/10/26 15:59: 38.00000:0	83	5
50968480	3079.79	555.75	45:55.	12	1	12:48	82	20

$$x_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}\}. \quad (9)$$

The process of the outlier detection algorithm based on the sum of similar coefficients is as follows.

① Normalize the original data.

Because the dimensions of these data sets are different and the data distribution is uneven, if different dimensions are used, the distance calculation results will be different. Common data normalization methods are min-max normalization, log function conversion, Atan function conversion, z-score normalization (zero-mean normalization), and the fuzzy quantization method.

After comparing the advantages and disadvantages of various standardization methods [31], it can be determined that z-score standardization exhibits better performance when the distance is used to measure similarity. Therefore, the z-score method is used in this study, given as follows:

$$x^* = \frac{x - \mu}{\sigma}, \quad (10)$$

where μ is the mean and σ is the standard deviation. The original toll data set is normalized in the interval $[-1, 1]$. The normalized X is denoted as X' , and its matrix representation result is as follows:

$$X' = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1m} \\ x'_{21} & x'_{22} & \cdots & x'_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x'_{n1} & x'_{n2} & \cdots & x'_{nm} \end{bmatrix}. \quad (11)$$

② Calculate the similarity coefficient.

By calculating the similarity coefficient between two factors, their degree of dispersion can be determined. The similarity coefficient matrix is as follows:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}, \quad (12)$$

where $r_{ij} = \sqrt{(1/n) \sum_{k=1}^m (x'_{ik} - x'_{jk})^2}$.

③ Calculate the sum of each row in the similarity coefficient matrix.

A larger value indicates that the object is farther away from other objects, i.e., it is more likely to be an outlier. The calculation formula is as follows:

$$P_i = \sum_{j=1}^m r_{ij}. \quad (13)$$

④ Determine whether the object i is an abnormal value.

$\lambda_k = (p_{\max} - p_k / p_{\max}) \times 100\%$ is set, where λ is the threshold set artificially according to experiments. If $\lambda_i > \lambda$, the object i is considered to be an outlier.

Via these algorithm steps, it is possible to find outliers in the toll data set relatively accurately.

The process of the outlier detection algorithm based on the sum of similar coefficients is presented in Figure 2.

To more intuitively describe the cleaning effect, 2000 data points of the “Out Station Name,” “Out Load,” and “Credit” features of the toll data before and after the cleaning process were selected for clustering processing. The original sample data distribution is presented in Figure 3(a), and the sample data after cleaning based on the sum of similar coefficients anomaly detection are presented in Figure 3(b).

It can be intuitively seen from Figure 3 that it is reasonable to use the outlier detection algorithm based on the sum of similar coefficients to process the data. Although some data are actually correct when considered from multiple dimensions, when each column is processed separately, these data will be mistakenly recognized as abnormal. The red and green dots in the figure represent normal values, and the blue dots are abnormal values or noise values that do not seem to conform to the distribution rules of the original data.

4. Highway Event Detection Algorithm Based on Improved Fast Peak Clustering

4.1. The Feasibility and Accuracy Validation of the Improved Algorithm. First, part of the toll data collected during the Spring Festival and the fourth week of February were, respectively, clustered by the improved algorithm to verify its feasibility, and the results are exhibited in Figure 4.

Figure 4 illustrates the clustering results in two different traffic levels. In Figure 4(a), as one of the biggest festivals in China, the traffic level reaches a massive state during the

FIGURE 2: The process of the outlier detection algorithm based on the sum of similar coefficients.

FIGURE 3: Toll data (a) before and (b) after data cleaning.

FIGURE 4: The clustering results of the improved algorithm: (a) the Spring Festival period; (b) the fourth week of February.

Spring Festival period. As presented in Figure 4(a), the improved algorithm divided the toll data during the Spring Festival into four categories. Regarding green points and blue points, although there was little difference in their transit time, the difference in their total vehicle weight led the algorithm to classify these points into different categories. Therefore, it can be concluded that, during the Spring Festival, the transit time of some passenger cars increased to

a level similar to that of trucks due to the influence of vehicle congestion and other factors. In Figure 4(b), the fourth week of February has a traffic level in the normal circumstance, and the effective data with outliers removed are classified into two categories.

To verify the outlier detection results, the black outlier points in Figure 4 are checked according to the raw data. It shows that all vehicles corresponding to outliers are

confirmed to have experienced abnormal events. Take Figure 4(a) as an example, and the vehicles, corresponding to the outlier points having a comparatively light total vehicle weight and a comparatively long transit time, have experienced breakdown or crash, while outliers have a normal transit time and a comparatively heavy total vehicle weight, and the corresponding vehicles have been overloaded.

In order to verify the clustering results, the toll data during the Spring Festival and the fourth week of February were classified according to the vehicle type, and the real classification results were obtained. The results are presented in Figure 5.

Take Figure 5(a) as an example, the vehicle type of red points refers to passenger car, vehicle type of yellow points refers to bus, while vehicle type of green points refers to trucks. Comparing with clustering result of Figure 4(a), it is obvious that the improved algorithm can distinguish the vehicle type correctly.

Next, the accuracy of the clustering results was compared to verify the accuracy of the algorithm. Two classical clustering algorithms, namely, the k-means and density-based spatial clustering of applications with noise (DBSCAN) algorithms, the original fast peak clustering algorithm, and the improved algorithm were, respectively, used to cluster the same toll data. The classifications results of the four clustering algorithms in both Spring Festival period and the fourth week of February period were compared with the classification of the original vehicle type. The results are depicted in Figure 6.

As shown in Figure 6, the accuracy of the improved fast peak clustering algorithm is notably higher than those of the original fast peak clustering algorithm and the classical k-means [32, 33] and DBSCAN [34] algorithms for data from both the Spring Festival and the fourth week of February, which demonstrates that the improved fast peak clustering algorithm has a higher validity and accuracy than the others. It also indicates that the algorithm results are highly consistent with the actual traffic situations.

The improved algorithm does not increase the time complexity, which is $O(n^2)$. It only slightly extends the calculation time than the original algorithm. The time taken by the improved algorithm and others to process all the data is presented in Figure 7. Although the processing time of the improved algorithm was similar to that of DBSCAN and longer than the original algorithm, it is obviously shorter than the k-means algorithm, and efficiency was not sacrificed due to the increase in accuracy.

4.2. Analysis of Vehicle Traffic Characteristics. To conduct an overall analysis of vehicle traffic in the time domain, the following experiment was designed. First, 700,000 pieces of data were randomly sampled by the hour, and 4800 pieces of data were obtained. Then, the travel time, travel mileage, and load were visualized according to time, as shown in Figure 8.

It can be seen from Figure 8 that, from 7:00 to 20:00, the traffic volume was relatively high, and the highest peak in each figure appears around 18:00. As shown in Figures 8(a) and 8(b), the number of small cars was densely distributed.

However, the number of large trucks with a load capacity of more than 20,000 kg was basically unchanged with time, and the distribution was relatively even within 24 hours. Moreover, the mileage and transit time exhibited overall increasing trends over time, respectively, as is shown in Figures 8(c) and 8(d). According to information such as the latitude and longitude, it is inferred that the toll station is located near the technology industrial estate. This provides ideas for analyzing the rationality of data distribution more profoundly. For example, most of vehicles entering and leaving the estate are passenger cars with a comparatively light weight, which are probably used to carry staff members. In addition, trucks with a heavy weight have a high mileage and transit time, which means that there could be a long distance between estate and its raw material production site or commodity storage site. Moreover, trucks that leave the estate during the day time (7:00–19:00) usually have a comparatively low mileage, while trucks that leave the estate during the night time (20:00–6:00) usually have a comparatively long transit time. Analysis above lays the foundation for further analyzing drivers' driving habits and exploring the conditions for the occurrence of abnormal events.

4.3. Analysis of Improved Algorithm in the Detection of Abnormal Events. Abnormal highway events were detected based on the improved fast peak clustering algorithm, which was used to calculate the distance between each data point on the filtered toll data, and the distance matrix was obtained as an input. The local density ρ and δ for each data point were calculated as described in Section 2.2, and the γ_i values of each point were calculated according to equation (4) and then arranged in descending order. The results are exhibited in Figure 9. To find the pseudo-center, the distribution of pseudo-centers is presented in Figure 10. Figure 11 presents the final cluster center distribution determined after correcting the cluster centers. The final clustering result is shown in Figure 12.

The red and green points in the figures are the valid data points of clustering, and the black points are abnormal data points. According to the 2799 pieces of sample data selected at the designated entry and exit stations, the abnormal events identified by fast peak clustering are mainly divided into the following four types:

- ① The transit time is too long: the transit time of most vehicles is about 1–2 hours, and the transit time of abnormal data is mostly more than 5 hours. The long transit time between two toll stations that are close to each other may be caused by accidents, parking, clock asynchronization, recording errors, or suspected fee evasion.
- ② The transit time is too short: the minimum transit time can be calculated from the distance between the two stations and the maximum transit speed of the road section. Data lower than this value are considered to be abnormal data, which may be caused by vehicle speeding, network failure, clock asynchronization, recording errors, or suspected fee evasion.

FIGURE 5: Results of toll data classified by vehicle type: (a) the Spring Festival period; (b) the fourth week of February.

FIGURE 6: Comparison of the accuracy of four clustering algorithms.

FIGURE 7: Comparison of the computation time of four clustering algorithms.

- ③ The total weight of the vehicle is too high: this problem mainly affects trucks and may be caused by overloading, failure to weigh equipment, recording errors, or suspected fee evasion.
- ④ The total weight of the vehicle is too low: this problem mainly affects trucks and may be caused by the failure to weigh equipment, recording errors, or suspected fee evasion.

Anomalies in toll data can be used to accurately track the basic information of the vehicle, station, lane, and personnel associated with an event. Moreover, the possible causes of an incident can be analyzed, and the scope of the incident investigation can be greatly reduced. For example, during the period of January 9–10, a large amount of abnormal data regarding the duration of traffic appeared at the same entrance or exit. This demonstrates that there may be a

FIGURE 8: Distributions of vehicle traffic characteristics in the time domain: (a) InLoad; (b) OutLoad; (c) mileage; (d) transit time.

FIGURE 9: Toll data γ_i values arranged in descending order.

problem with the station's toll system software, communication network, or lane computer clock, which require timely inspection and maintenance. Another example is presented in Table 3, in which the transit times of multiple data of the same vehicle were significantly lower than the normal value (the average is 1–2 h). This demonstrates that the vehicle was likely to be speeding or attempting to evade fees, or that software or network failure had occurred, and special verification of the license plate is required.

According to the online center's duty records and manual confirmation, a total of 72 pieces of traffic jam data caused by various reasons were recorded, as well as 31 pieces of event data such as speeding, system failures, and suspected fee evasion. Two types of data were found by using the fast peak algorithm and outlier detection algorithm, respectively. By comparing the abnormal events detected by the fast peak algorithm with the real abnormal events, it can be found that 70% of the abnormal data detected by the

FIGURE 10: Toll data pseudo-centers.

FIGURE 11: Toll data cluster centers.

FIGURE 12: Toll data clustering results.

TABLE 3: Examples of abnormal data with a short transit time.

ID	Transit time (h)	Weight (kg)
** 31417	0.53 611 111	1500
** 31417	0.75 555 556	1400
** 31417	0.74 444 444	1400
** 31417	0.80 533 333	1500
** 31417	0.65 000 000	1500

FIGURE 13: Time distribution of abnormal events.

clustering algorithm corresponded to real traffic jams, accidents, system failures, and other abnormal events. It demonstrates that the algorithm can quickly and accurately identify abnormal events such as road congestion, system failures, and suspected fee evasion hidden in the toll data.

The times of abnormal events were statistically analyzed to examine the distribution of abnormal events in the province. Abnormal event detection was carried out on 70,000 pieces of data, and a total of 1,506 abnormal events were obtained. These events were visualized in the time domain, and the results are presented in Figure 13.

It can be seen from Figure 13 that there were two obvious peaks in the occurrence of abnormal events. The first peak time was between 10:00 and 13:00, which accounted for 53.78% of the abnormal events over the entire period. The second peak was from 15:00 to 18:00 accounted for 27.76% of the total. The reason for this phenomenon may be that the flow of vehicles passing through the toll gate during these two time periods was greatly increased, leading to an increase in abnormal events. In order to determine whether the relationship between the out-time and the occurrence of abnormal events is significant, a statistical test analysis with SPSS software was performed on the data distribution in Figure 13. The chi-square test results show that σ is 0.038, which is less than 0.05, so it is reasonable to accept that abnormal events have significant differences at different outbound times. This shows that the abnormal event detection result is in line with the facts. The results indicate that, to quickly detect abnormal events in massive toll data,

the traffic control department can increase its investigation efforts during these two time periods.

5. Conclusions

This paper focused on a highway event detection method based on the fast peak clustering algorithm. The main conclusions of this research are as follows:

- (1) An outlier detection and data-filling algorithm for multidimensional data based on the sum of similar coefficients is proposed. The proposed algorithm improves the accuracy of the outlier algorithm by 10%.
- (2) The advanced fast peak clustering algorithm is improved. Compared with the original fast peak clustering algorithm, the accuracy of the proposed algorithm is increased by 20%.

A case analysis of highway events based on the proposed fast peak clustering algorithm can be conducted to accurately locate the vehicles, stations, and other related information. The scope of the investigation of abnormal events can be narrowed to a great extent. Abnormal events such as long-term stay and vehicle overload hidden in the toll data can be easily identified.

However, this research focused on the analysis of historical data and did not include integration with the toll system to complete real-time data analysis. In future research, due to the complicated research directions and

problems involved in highway operation management, the proposed algorithm must be further improved and optimized. Moreover, the algorithm can be combined with other data sources, such as the correlation analysis of operational indicators, to more accurately determine specific reasons for the occurrence of events.

Data Availability

The data utilized in this research were obtained from the Shaanxi Provincial Department of Transportation of China. They contain sensitive information about the owners and therefore cannot be shared publicly.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

L.P., Z.S., and W.L. conceptualized the study and obtained the resources. L.P. and Z.S. did data curation and administrated the project. L.P., Z.S., and Y.H. performed formal analysis. L.P., Y.H., and H.Z. prepared methodology, investigated the study, and wrote the original draft preparation. W.L. was responsible for funding acquisition. L.P., W.L., H.Z., and Y.H. validated the study. L.P., Z.S., Y.H., W.L., and H.Z. reviewed and edited the manuscript.

Acknowledgments

This work was supported by the National Key Research and Development Program for the Comprehensive Transportation and Intelligent Transportation Project (Grant no. 2018YFB1600202), the Shaanxi Province Transportation Technology Research Project (Grant no. 18-22R), and the Fundamental Research Funds for the Central Universities (Grant no. 300102240201).

References

- [1] A. F. Oklilas, A. Fitriyani, and A. Zarkasi, "Automatic toll road payment system security using radio frequency identification (RFID) with the blowfish method," *Journal of Physics Conference Series*, vol. 1196, no. 1, pp. 1742–6596, 2019.
- [2] D. K. Liang, L. Y. Fan, and Z. Hao, "Primary study on intelligent traffic system (ITS)," *Technology of Highway and Transport*, vol. 20, no. 6, pp. 1009–6477, 2004.
- [3] R. Zhou, L. Zhong, N. Zhao et al., "The development and practice of China highway capacity research," *Transportation Research Procedia*, vol. 15, no. 1, pp. 14–25, 2016.
- [4] K. M. Zhang, Y. W. Cai, and Y. Ren, "Space anomaly events detection approach based on generative adversarial nets," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 45, no. 7, pp. 1329–1336, 2019.
- [5] J. Weng, L. Liu, and B. Du, "ETC data based traffic information mining techniques," *Journal of Transportation Systems Engineering and Information Technology*, vol. 10, no. 2, pp. 57–63, 2010.
- [6] R. C. Williams, J. C. Hildreth, and M. C. Vorster, "Highway construction data collection and treatment in preparation for statistical regression analysis," *Journal of Construction Engineering and Management*, vol. 135, no. 12, pp. 1299–1306, 2009.
- [7] J. L. Li and W. H. Sun, "Analysis of causes of expressway traffic accidents and counter measures," *Chinese Journal of Safety Science*, vol. 15, no. 1, pp. 62–65, 2005.
- [8] H. P. Lu, S. X. Luo, and R. M. Li, "Spatial distribution characteristics of road traffic accidents in shenzhen based on GIS analysis," *Chinese Journal of Highways*, vol. 32, no. 8, pp. 156–164, 2019.
- [9] F. Meguellati, A. N'Guessan, and T. Hermitte, "Analyzing the maximum abbreviated injury scale in vehicle crashes by using a logistic normal model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2432, no. 1, pp. 74–81, 2014.
- [10] S. K. Barai, "Data mining applications in transportation engineering," *Transport*, vol. 18, no. 5, pp. 216–223, 2003.
- [11] X. Jin, D. Srinivasan, and R. L. Cheu, "Classification of freeway traffic patterns for event detection using constructive probabilistic neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 12, no. 5, pp. 1173–1187, 2001.
- [12] X. P. Wu, "Detection method of abnormal events on expressway based on multi-svm classifier fusion," *Modern Traffic Technology*, vol. 11, no. 4, pp. 63–67, 2014.
- [13] J. Y. Sun, R. J. Mou, and X. B. Su, "Detection of abnormal state of expressway based on Support Vector Machine," *Journal of Chongqing Jiaotong University (Natural Science Edition)*, vol. 37, no. 9, pp. 1–7, 2016.
- [14] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Physica A: Statistical Mechanics and Its Applications*, vol. 517, pp. 29–35, 2019.
- [15] Z. J. Ye, X. H. Zhang, H. K. Xu, and Q. Wang, "Research on expressway toll audit technique based on data mining," *Advanced Materials Research*, vol. 186, pp. 363–367, 2011.
- [16] S. L. Li, X. D. Zhang, and H. W. Li, "Risk assessment model of expressway under adverse weather based on dynamic bayes network," in *Proceedings of the 19th COTA International Conference of Transportation Professionals*, Nanjing, China, July 2019.
- [17] S.-B. Li, T. Sun, D.-N. Cao, and L. Zhang, "Incident detection method of expressway based on traffic flow simulation model," *Communications in Theoretical Physics*, vol. 71, no. 4, p. 468, 2019.
- [18] L. M. Abualigah, A. T. Khader, E. S. Hanandeh, and A. H. Gandomi, "A novel hybridization strategy for krill herd algorithm applied to clustering techniques," *Applied Soft Computing*, vol. 60, pp. 423–435, 2017.
- [19] L. M. Abualigah and A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," *The Journal of Supercomputing*, vol. 73, no. 7, pp. 4773–4795, 2017.
- [20] G. Nathiya, S. C. Punitha, and M. Punithavalli, "An analytical study on behavior of clusters using K means, EM and K* means algorithm," *International Journal of Computer Science & Information Security*, vol. 7, no. 3, pp. 21–30, 2010.
- [21] L. M. Abualigah, "Multi-verse optimizer algorithm: a comprehensive survey of its results, variants, and applications," *Neural Computing And Applications*, vol. 30, pp. 1–21, 2020.
- [22] L. M. Abualigah and E. S. Hanandeh, "Applying genetic algorithms to information retrieval using vector space model," *International Journal of Computer Science, Engineering and Applications*, vol. 5, no. 1, p. 19, 2015.

- [23] Y. A. Shah, H. A. Habib, F. Aadil, M. F. Khan, M. Maqsood, and T. Nawaz, "CAMONET: moth-flame optimization (MFO) based clustering algorithm for VANETs," *IEEE Access*, vol. 6, pp. 48611–48624, 2018.
- [24] Z. J. Huang and C. W. Xia, "Average clustering of discrete data based on probability and its application to expressway toll fraud detection," in *Proceedings of the 2009 WRI Global Congress on Intelligent Systems*, Xiamen, China, May 2009.
- [25] L. M. Abualigah, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*, Springer, Berlin, Germany, 2019.
- [26] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "Hybrid clustering analysis using improved krill herd algorithm," *Applied Intelligence*, vol. 48, no. 11, pp. 4047–4071, 2018.
- [27] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A new feature selection method to improve the document clustering using particle swarm optimization algorithm," *Journal of Computational Science*, vol. 25, pp. 456–466, 2018.
- [28] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis," *Engineering Applications of Artificial Intelligence*, vol. 73, pp. 111–125, 2018.
- [29] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
- [30] H. X. Zhao, R. R. Deng, and Y. J. Zhang, "A Toll data Mining method for expressway traffic Flow analysis," *China Highway Journal*, vol. 31, no. 8, pp. 155–164, 2015.
- [31] L. Pei, Z. Sun, T. Yu et al., "Pavement aggregate shape classification based on extreme gradient boosting," *Construction and Building Materials*, vol. 256, Article ID 119356, 2020.
- [32] I. B. Mohamad and D. Usman, "Standardization and its effects on K-means clustering algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299–3303, 2013.
- [33] W. Rao, J. Xia, W. Lyu, and Z. Lu, "Interval data based k means clustering method for traffic state identification at urban intersections," *IET Intelligent Transport Systems*, vol. 13, no. 7, pp. 1106–1115, 2019.
- [34] M. Ester, H. P. Kriegel, and J. Sander, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, August 1996.