Peer-to-Peer Networking and Applications (2021) 14:2511–2523
https://doi.org/10.1007/s12083-020-00994-3

Application of big data optimized clustering algorithm in cloud computing environment in traffic accident forecast

Zhun Tian[1,2] · Shengrui Zhang[1,3]

Received: 3 July 2020 / Accepted: 26 August 2020 / Published online: 30 September 2020

Abstract

As the usage rate of cars is getting higher and higher, the injuries and losses caused by traffic accidents are also getting bigger and bigger. If some traffic accidents can be predicted, then such losses can be greatly solved. Although there are abundant research results on intelligent transportation, there are not many research results on how to predict traffic accidents. For this issue, the main aim of this paper is to propose a continuous non-convex optimization of the K-means algorithm in order to solve the model problem in the traffic prediction process. First, this paper uses clustering algorithm for feature analysis and big data for the establishment of simulation model in cloud environment. Through this paper an equivalent model, using matrix optimization theory to analyze and process K-means problem, and design efficient and theoretically guaranteed algorithms for big data. By simulating the traffic situation in Shanghai city within three years, the outcomes display that the model endorsed in the given paper can predict traffic accidents at a rate of 93.88% and the accuracy rate of traffic accident processing time is 78%, which fully illustrates the effectiveness of the model established in this paper.

Keywords Cloud computing · Big data · Clustering algorithm · Traffic accident prediction · Algorithm optimization

Abbreviations

NAP	Normalized Traffic Accident Propensity
EEG	Electroencephalogram
R-CNN	Region Based Convolutional Neural Network
ANN	Artificial Neural Network
MEC	Mobile Edge Computing
MLP	Multiple layer Perceptron Neural Networks
SVM	Support Vector Machine
GDP	Gross Domestic Product
HDFS	Hadoop Distributed File System
EGS	Expanded Graphites
YOLO	You only look once
Fuzzy ARTMAP	Fuzzy Adaptive Resonance Theory
RF	Random Forest
KM-MBFO	K-means Modified Bacterial Foraging

1 Introduction

The rate and severity of traffic accidents in China have remained high for a long time.By analyzing the recorded causes of traffic accidents, it can be found that traffic accidents in China are mainly caused by two reasons: people (93%) and cars (4%). The reasons for road, traffic, weather, and management account for only 3%, which is far from the 16% of other developed countries. The main reason is that the main responsible person must be determined during the accident handling process, so it is not surprising that the proportion of human reasons is large [1]. In fact, the occurrence of a traffic accident is generally accidental, and it is usually caused by two or more reasons, among which there are no shortage of road, traffic, weather, and management reasons [2, 3]. For the purpose of the reduction of traffic accidents, it is necessary to make a

This article is part of the Topical Collection: *Special Issue on Network In Box, Architecture, Networking and Applications*
Guest Editor: Ching-Hsien Hsu

✉ Zhun Tian
tianzhunxauat@126.com

[1] College of Transportation Engineering, Chang'an University, Xi'an 710064, Shaanxi, China

[2] School of Civil Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, Shaanxi, China

[3] Key Laboratory of Transport Industry of Management, Control and Cycle Repair Technology for Traffic Network Facilities in Ecological Security Barrier Area (Chang'an University), Xi'an 710064, Shaanxi, China

prediction of traffic accidents in order to take effective preventive measures in a timely manner.

However, the current situation is that on the one hand, the data collected by many established traffic monitoring systems are not valued; on the other hand, many people think of solving traffic problems and immediately think of using manual traffic data sampling surveys. This method not only increases the cost, but also the survey data has large contingency, poor reliability, and poor adaptability, and some hidden laws and traffic characteristics cannot be found [4]. How to use the existing traffic data for analysis, find some traffic flow characteristics from the numerous data, formulate effective predictions for these characteristics, and be targeted, and these methods should be reliable, that is, when traffic flow characteristics change At present, the prediction and prevention of the occurrence of accidents with the purpose of prevent them in time has become a hot and difficult point of current research [5].

The main contribution of this research is to propose a continuous non-convex optimization of the K-means algorithm to solve the prediction issues. Aiming at the inefficiency caused by fuzzy clustering algorithm in high-dimensional data, clustering on feature space is considered in this paper. Unlike traditional methods, the algorithm in the article can make the fuzzy clustering accuracy on the feature space and the fuzzy clustering accuracy on the original data consistent, that is, the dimensionality reduction (feature space) process and clustering process of big data. In big data clustering, the clustering process is generally divided into two parts: data preprocessing and data clustering. The effect of clustering is directly related to the matching degree of these two steps. At the same time, when the clustering center and the distance function meet certain constraints, the fuzzy K-means algorithm in the feature space can be accelerated, aside from ensuring the degree of accuracy of the algorithm, it also efficaciously improves the efficiency of the algorithm, making the application on big data becoming.

2 Related work

Taamneh M proposed two analytical models and the “Artificial Neural Network” (ANN) model, by the utilization of historical data from 1986 to 2005 to determine the number of traffic accidents mortalities in three major cities in Turkey. The accuracy of this method is not high [6]. Lee S.L offered a classification model for road accidents to envision degree of severity of injury caused by accidents. First, the traffic accident characteristic information is obtained and stored in the accident characteristic database. The accident feature data is classified into two disjoint sets: the training set and the test set. Next, Lee S L uses the training data set to generate a classification model using machine learning techniques, and then verifies the performance of the model generated using the test data set. Lee SL uses known classification methods-decision trees, neural network models, and Bayesian networks to generate classification models. But his method is not very practical [7]. En ZM introduced the establishment of a normalized traffic accident propensity (NAP) prediction model after drinking, and measured the EGS and traffic accident propensity of eighteen drivers under different drunken states, taking into account the measurements from the left frontal lobe. The instantaneous complexity and long-term periodicity of EGS, proposed and calculated the power gain and fuzzy entropy of δ wave of EEG, introduced and studied the hybrid Sigma-Pi neural network, from the power gain of δ wave and fuzzy entropy of EEG A prediction model for NAP was established [8]. This method has high accuracy, but it is not universally applicable. Through the medium of Deep learning and wireless communication methods Deeksha Gour and Amit Kanskar devised YOLO (you only look once), regression based algorithm, which detects accidents using web cam. It alerts nearby vehicles about emergency. It uses faster R-CNN method but the problem is that it assumes that every vehicle is connected, which may be not, and the other thing such big data is very complex to monitor through web cam [9].

V.Priya and C.Priya formulated a deep model by scrutinizing “the spatial and temporal patterns” of traffic accident frequency, established on recurrent neural network. This model presents top notch results as compared to simple neural network. But its minus point is the limitation of client data and their lack of availability with respect to distance [10]. Lu Wenqi and Luo Dongyu designed TAP-CNN model by “the use of state matrix”, for input traffic condition, and “convolution neural network”, as traffic accident prediction model. Intrinsically this model developed to predict the traffic accidents on highways. As comparison to “traditional back propagation network”, TAP-CNN model exhibits accuracy 7.7% higher than the traditional model as it occupies the accuracy of 78.5%.

Though a very good approach is pocketed but it still has some shortfalls for example model doesn’t consider the structure of road deeply, which should be put in matrix as input, and it lacks the proper amount of data, which has to be trained [11]. MEC, Mobile Edge Computing, platform is used to detect accidents and inform the drivers and medical emergency nearby about the accident in real time situations by the means of low latency, high bandwidth and massive computing power. To deploy this idea MySQL, Java Servlet TomCat on WiCloud is utilized. A total accuracy of 72% of accident vehicles is detected. MEC make it sure that related objects can receive information instantaneously. That’s indeed a very good approach to make improvements in traffic flow management and emergency medical assistance [12]. By using more variables and larger dataset Bayesian network constructed to get more accurate results regarding road accidents. Limitations like asymmetry in data regarding injuries and the unevenness of data made it difficult to work efficiently. Data, which traffic police provided, created confusion as other causes are possible apart from the causes traffic police recorded [13]. A clustering algorithm is applied to train NN and decision tree, as the observations shows that clustering algorithm works better. “Levenberg Marquardt algorithm” and “Fuzzy ARTMAP” trained neural network gives the accuracy of 65.6% and 60.4% respectively as compared to only “Fuzzy ARTMAP” which provides the accuracy of 56.1%. It doesn’t provide sufficient information about the actual speed of vehicle, 67.68% vehicle’s speed was unknown [14]. Many suspected variables are used to build SEM to analyze road accidents. Results of the research simulate the relationship between accident size and the multiple factors which engage in causing the accident. Despite the lacking of many points, like the limitation of information, this research provided appreciable view regarding accidents [15]. Multiplelayer Perceptron Neural Networks (MLP) caters better result as compared to regression model. MLP and Fuzzy ARTMAP, two ANN models, are deployed for the research. Fuzzy ARTMAP presented the accuracy of 56.2% while MLP prone rather higher accuracy. Although it provided better results but still many factors were not discussed so improvements are needed [16]. Random Forest (RF) algorithm provides higher accuracy as compared to ANN and SVM algorithms. In the paper vehicle to vehicle communication for the exchange of information is used. It is conjectured that every single vehicle is able to telecast its microscopic data and in similar terms every connected vehicle is capable to receive and decode it. And this the shortfall of this research [17].

K-mean method is a galore method for the training of data. Hadoop cluster is based on this method. It is a cluster which includes deep process from master node to several slave nodes and it is based on parallel computational process. KM-MBFO shows better performance as compared to typical K-means. This approach that is hybrid in nature can be good to handle big data sets which are minus point in many previous researches. It provides better time span but still has some minus points [18]. The related work engaged ANN and Decision Tree analysis techniques. Id3 is constructed and used and best result obtained through it is 77.70% correct data that is reported and 22.29% incorrect data. Decision Tree shows better results as compared to RBF model which shows 54.73% correctly reported occurrences and MLP model which demonstrates 52.70% correctly reported instances [19]. Established on the concept of Internet of Things (IOT) accident detection system was programmed. A mobile application is developed to monitor the road accidents as well as a navigation system to report the accident to the adjacent available hospital. The proposed application can read data in real time and can transfer the information for the further validity and computation of information. This work reduces the chances of false reports as it works in real time environment but it still holds some limitations like simulated environment and the lack of privacy [20].

In overall the issues that identified from the existing work are practical implementation, power gain, complex monitoring, lack of availability, accuracy rate, time span, false reports, security and privacy and many more. So in order to tackle some of the issues the continuous non-convex optimization of the K-means algorithm is proposed to solve the model problem in traffic prediction process.

3 Proposed method

3.1 Technology of cloud computing

It has evolved with the development of PCs and the Internet. A PC is a combination of software and hardware. Because many software have high requirements on hardware, the hardware of the PC determines the overall function of the PC [21, 22]. The popularity of PCs and the power of software have shifted attention from the processing of data and resources to the sharing of data and resources. The Internet connects the entire world, people can share the information they have to the Internet, and they can find the information they need on the Internet. As Internet developed with the passage of time, the explosive growth as well as the accumulation of data resources spread at the speed of light. How to find the required information quickly and efficiently through calculations in so many fast-growing information resources is a problem.

(1) Cloud computing technology structure

Cloud computing technology is a comprehensive technology system that integrates multiple technologies such as network technology, parallel computing technology, virtualization technology, interface technology, interactive technology, database technology, sandbox technology, etc.; it integrates multiple remote and heterogeneous cloud resources After unified management, after virtualization and generalization, various resources will be provided in a more intuitive and faster way to users in different locations, and fees will be generated according to the actual usage of users; the technology system it uses is Formed to meet the goals of cloud computing services under commercial network conditions [22, 23].

(2) Key technologies of cloud computing

Cloud computing does not provide a single service, but a collection of services. Its technical framework is shown in Fig. 1. Provide IT services to users through three forms: infrastructure, platform, and software [24, 25]. Cloud computing play a significant role while handing accidents and it gives some support to accident handing. However, a cloud based system for traffic accident offers the functionality in order to process real-time average speed value of the vehicles for detecting accidents and it provide services over the network (SaaS, PaaS and IaaS). SaaS is a delivery model and on-demand software which is fully functional, based on the subscription the software is delivered to the customers over the web. In PaaS, a platform is delivered to the users if they be able to initialize, manage and develop

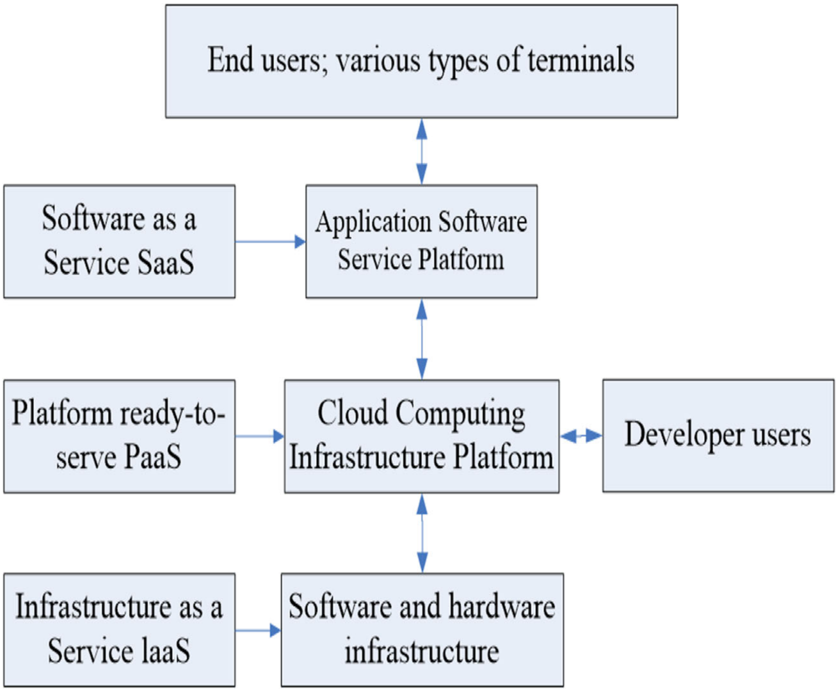

End users; various types of terminals
Software as a Service SaaS
Application Software Service Platform
Platform ready-to-serve PaaS
Cloud Computing Infrastructure Platform
Developer users
Infrastructure as a Service IaaS
Software and hardware infrastructure

Fig. 1 Cloud computing technology framework

applications. IaaS permits access to infrastructure and it offer additional storage for network bandwidth, data backups, etc.

Figure 1 is a highly generalized view of cloud computing service forms [26].

3.2 Big data

Big data helps to predict road accidents using analysis including data-sampling, data collection, data processing, classification, etc. As compared to traditional data management technology, the exploration scope of big data technology is wider. [27, 28]. It includes the structured data in the known range as well as the degree of correlation of the data [29]. There are three points:

- To store large amounts of structured and unstructured data [30, 31]
- To process large amounts of data in real time.
- To establish an algorithm model and perform continuous optimization based on real-time data [32].

Typical big data technology architecture is shown in Fig. 2:

3.3 Cluster analysis method

3.3.1 Data structure and similarity measurement of cluster analysis

A. Data structure for cluster analysis

Many memory-based clustering algorithms choose the following two representative data structures:

Use p variables (also called measures or attributes) to person represent n objects. This data structure is in the form of a relational table, which is regarded as a matrix of $n\ *\ p$:

$$X = \begin{bmatrix} x_{11} & x_{1f} & \cdots & x_{1p} \\ x_{21} & x_{2f} & \cdots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{pl} & x_{pf} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

B. Dissimilarity matrix (also called object-object structure): Stores the similarity between n object in pairs, and its representation is an $n*\ n$-dimensional matrix.

$$D = \begin{bmatrix} 0 & & \cdots \\ d(2,1) & 0 & \cdots \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ d_{(n,1)} & d_{(n,2)} & \cdots \end{bmatrix} \tag{2}$$

Here $\mathbf{d_{(i,j)}}$ is a quantified representation of the dissimilarity between objects i and j. The difference between

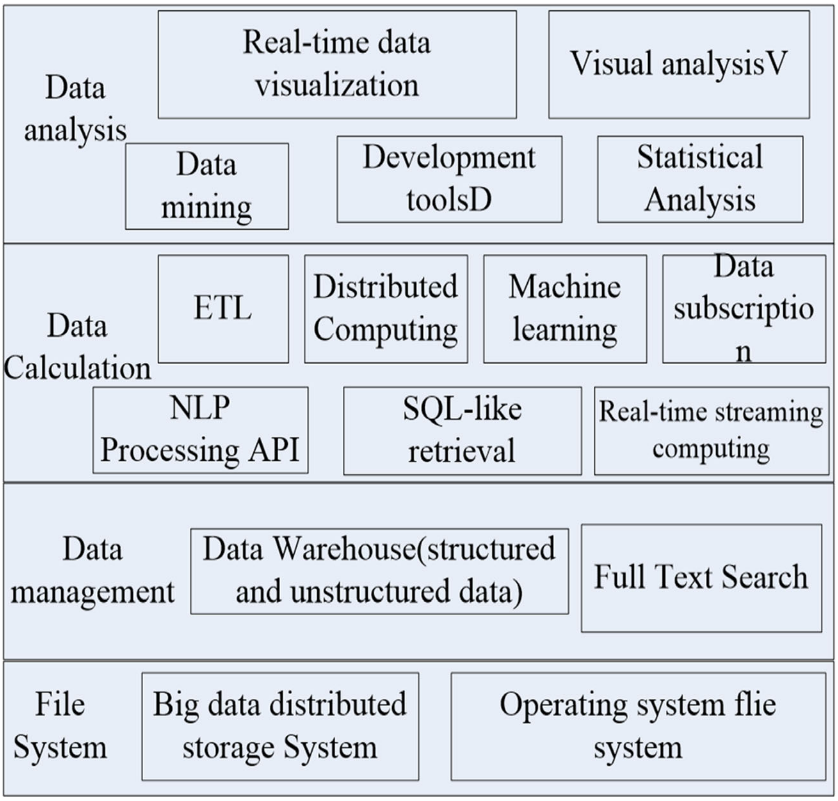


Fig. 2 Big data technology architecture

objects is generally calculated on the base of distance between objects. Assuming i = $(x_{i1}, x_{i2}, \ldots, x_{ip})$ and j = $(x_{i1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, usually $d_{(i,j)}$ is non-negative value. When the objects i and j are more "similar" or "close", the value is closer to 0, and $\mathbf{d_{(i,j)} = 0}$ that is phase of the object and its own. The degree of difference is 0: the difference between the two objects are, determines the largeness of value. Obviously, i = j in the element $\mathbf{d_{(i,j)}}$ located on the diagonal of the matrix, which is a measure of the dissimilarity between an object and itself, so the values on the diagonal are 0. Data matrix also called a "two-mode matrix", and the dissimilarity matrix is called "single-mode matrix".

3.3.2 Similarity measure of cluster analysis

During clustering, similar objects are put into the same cluster according to the similarity between the objects, and dissimilar objects are classified into different clusters. Therefore, the measure of similarity directly determines the result of clustering. Based on the similarity, the cluster analysis goal is to group a collection of patterns. Distance measures or similarity are the main modules used via distance based clustering in order to cluster the comparable data points into similar group although disparate data points into different groups. For numerical data and enumerated data, the method of measurement is also different.

For numerical attributes, there are two similarity measures: distance similarity and angular similarity. There are four types of distance similarity:

$$\text{Absolute distance}: \quad di_j(1) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{3}$$

$$\text{European distance}: \quad d_{ij}\ (2) = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^2 \right]^{1/2} \tag{4}$$

$$\text{Ming's distance}: \quad d_{ij}\ (q) = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^q \right]^{1/q} \tag{5}$$

$$\text{Chebyshev distance}: \quad d_{i\ j}(\infty) = \max_{1 \le k \le p} \left| x_{ik} - x_{jk} \right| \tag{6}$$

The first two measurement methods meet the following mathematical requirements for the distance function:

a.

$d_{(i,j)} \geq 0$: The distance is a non−negative value;

b.

$d_{(i,j)} = 0$

: the distance between an object and itself is 0;

c.

$d_{(i,j)} = 0$: the separation function is symmetrical;

d.

$d_{(i,j)} \le d_{(i,h)} + d_{(h,j)}$: The direct distance from object j will not be greater than the distance

(triangular inequality) through any other object h.

The triangular inequality is defined as the sum of any two-sides length of a triangle and it would be greater than the remaining side length. It follows the shortest path characterization between two points and if triangle has non-zero area then the inequality is said to be severe. The triangular inequality has no unavoidable metric space rule so it naturally forms metric spaces.

For enumerated data, the number of the same attribute value can be used as the similarity measure, or ratio of the number of the same attribute value as compared to the number of different attribute values can be used as the similarity measure.

3.3.3 Clustering of fuzzy K-means

It's an algorithm based on the objective function. Data set X is divided into K classes, and each object in X is an m-dimensional vector, that is, $x_i = \{x_{i1}, x_{i2}, \ldots x_{im}\}$. The j-th clustering center is also an m-dimensional vector, that is, $\mathbf{vj} = \{\mathbf{vj1}, \mathbf{vj2}, \ldots, \mathbf{vjm}\}(\mathbf{1} \le \mathbf{j} \le \mathbf{k}); \mathbf{V} = \{\mathbf{v1}, \mathbf{v2}, \ldots, \mathbf{vk}\}$ is a $\mathbf{k} \times \mathbf{m}$ matrix composed of k clustering center vectors. In fuzzy division, each sample cannot be strictly divided into a certain category, but belongs to a certain category with a certain degree of membership.

The Objective function of fuzzy k – means clustering is:

$$\min J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{k} (u_{ji})^m (d_{ji})^2 \tag{7}$$

The constraints are:

$$S = \left\{ U \epsilon\, R \middle| \forall u_{ji} \epsilon [0, 1], \sum_{j=1}^{k} u_{ji} = 1, 0 \le \sum_{i=1}^{n} u_{ji} \le n \right\} \tag{8}$$

Among them, u_{ji} represents the membership degree of the i-th sample x_i belongs to the j-th category, m *is a* weighted index, and $d_{ji} = \|x_i - v_j\|$ represents the distance of the j-th center of the i-th sequence (Fig. 3).

4 Experiments

4.1 Acquisition of data

The data needed for accident risk prediction research in this paper comes from the Shanghai Municipal Traffic Management Department, which is composed of historical traffic accident data and traffic flow data. The historical data of traffic accidents comes from Shanghai's 2017–2019 total accident handling records, a total of "39,282". These data include the time, place, latitude and longitude of the accident, and analysis of the accident. Shanghai Municipal Traffic provided the Traffic flow data by using various traffic flow detection methods and technologies. These data include the device number, road segment number, lane-number, collection-date, collection time, flow, speed, time occupation Rate etc.

The above two kinds of data play different roles in real-time traffic accident risk prediction. For the original traffic flow data detected by the microwave detector, without further analysis, it is impossible to intuitively see whether an accident occurred, and the time and place of the accident. The accident history data can provide various information about the accident more reliably, but the traffic flow parameters at the time of the accident cannot be known. Therefore, this paper analyzes the two data comprehensively.

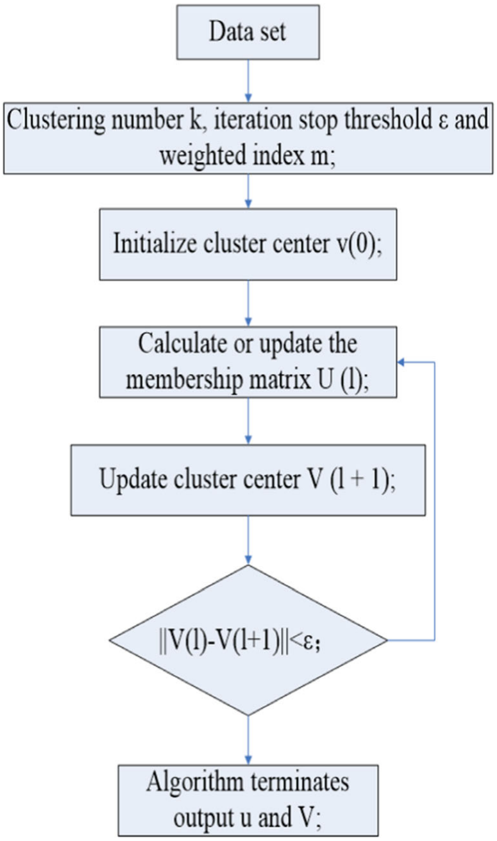

Data set
Clustering number k, iteration stop threshold ε and weighted index m;
Initialize cluster center v(0);
Calculate or update the membership matrix U (l);
Update cluster center V (l + 1);
||V(l)-V(l+1)||<ε;
Algorithm terminates output u and V;

Fig. 3 The flow of "fuzzy k-means clustering algorithm"

4.2 Analysis indicators of road traffic accidents

There are three main types of quantitative indicators used in this paper: absolute indicators, relative indicators, and combined indicators.

(1) Absolute index. Absolute indicators are absolute numbers used to reflect the overall scale and level of accidents. Four absolute statistical indicators usually evaluate road safety level for traffic, these are: accident deaths, injuries, accidents, and direct economic losses. The four indicators represent the overall situation of road traffic safety in a region from different perspectives, and all of them have good operability.

(2) Relative indicators. Absolute indicators can reflect the overall situation of the accident, but cannot reveal the regularity within the overall. Moreover, absolute indicators have no common basis and are difficult to compare directly. To this end, relative indicators need to be established. The analysis of influencing factors of traffic accidents on road shows that the level of road traffic safety in each region is related to the total population, motor vehicle ownership, highway-mileage, and regional GDP in the region. The relative indicators take into account the role of these influencing factors. Commonly used relative indicators are the death-rate of 100,000 people, the death-rate of 10,000 vehicles, the death-rate of 100 million vehicles per kilometer, and the accident rate per kilometer. The death rate of 100 million vehicle kilometers is poorly measurable and difficult to obtain.

(3) Combined indicators. Relative indicators often only consider a certain influencing factor, and combined indicators are more scientific. Comprehensively reflect the state of traffic safety, and consider the comprehensive impact of all elements of road traffic, including people, cars, and roads. Equivalent comprehensive mortality is a typical combined indicator of road traffic safety.

4.3 Determination of weights

Objective evaluation of things may include many evaluation factors. It is not possible to make a judgment based on the quality of a certain indicator, but a comprehensive evaluation should be based on multiple factors. There are usually subjective and objective assignment methods to opt the weight of each indicator. The subjective evaluation method, also known as the expert evaluation method, is to determine the weight coefficient of the evaluator according to the subjective emphasis on each attribute. The characteristic of 'subjective assignment method' is that it contains subjective color, and the result of assignment is related to the knowledge structure, work experience and preference of the evaluator. This method is difficult to avoid subjectivity and will vary from person to person. The evaluation criteria are based on multiple factors, when it is subjective to this research work the evaluation methods includes expert consultation, analytical hierarchy process and feature vector method. In order to enhance the objectivity of index selection, the determination of index weights should adopt the objective method. To calculate the weight of each index mathematical method based on the internal relationship between the indicators is used. This article uses the "entropy method" to determine the weight. The entropy method is a generally used weighting method which helps to determine the weight criteria, measure the uncertainty information and measure the amount of useful information.

Entropy is a term in thermodynamics. There are n clustering objects and m clustering indicators to form the original indicator data matrix X = (X_{ij}) n* m. The steps to determine the weight of an indicator using the entropy method am as follows:

(1) All indicators are quantified. and the weight of the j-th index value of the 1-th object is calculated:

$$p_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij} \tag{9}$$

Here, suppose $x_{ij} = \geq 0, and \ \sum_{i=1}^{n} x_{ij} > 0$

(2) Calculate the entropy of the j-th index:

$$e_j = k \sum_{i=1}^{n} p_{ij} 1\mathrm{n}\, p_{ij} \tag{10}$$

Where $k > 0$ and $e_j > 0$

(3) Calculate the coefficient of difference for the j-th index:

$$\mathrm{g}_i = 1 - e_j \tag{11}$$

When g_j is larger, the index is more important.

(4) Calculate the weight of the j-th index:

$$\mathcal{W}_j = \mathrm{g}_j / \sum_{j=1}^{m} \mathrm{g}_j \tag{12}$$

The larger the entropy value, the more chaotic the system is. and the more orderly it is. The greater the degree of variation in the value of an indicator. The smaller the information entropy. The greater the amount of information provided by the value, and the greater the weight of the indicator. Conversely, the smaller the degree of variation in the value of an indicator, the more information The larger the amount of information provided by the indicator. The smaller the weight of the indicator. The entropy method has strong self-adaptive function and greater credibility, but it lacks the horizontal comparison between various indicators.

4.4 Determine the optimal number of clusters

In Fuzzy clustering analysis, different clustering results can be obtained for different $\lambda \epsilon [0, 1]$.To determine a specific classification number k of the sample. This raises the question of how to determine the threshold λ many practical problems need to select a certain threshold λ. There are generally two methods:

(1) It is necessary to accurately estimate the samples in advance and divide them into several categories, and then adjust the values in dynamic clustering to obtain appropriate clustering results according to actual needs. Of course, the threshold can also be determined by experienced experts combined with their professional knowledge, so as to obtain the equivalent number k at the level.
(2) Determine the best value of λ with F statistic

The proportion of human reasons is also large in road traffic accidents so it is necessary to eliminate the subjective influencing factors of human. For the elimination the F statistic is used to determine the best classification to eliminate the influence of human subjective factors, and it is a better method for determining the number of clusters. The calculation principle is as follows:

Let the classified sample universe be $X = \{x_1, x_2, \ldots, x_n\}$ and each meta-x_i, consists of m.

attributes, that is, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$,$(i = 1,\ 2,\ \ldots,\ n)$. Let the number of classes corresponding to the λ value be r, n_j, be the number of samples in the j-th class, and the sample in the jth class be $\left\{x_1^j, x_2^j\right\}, \ldots, x_{n_j}^j$. The cluster center vector $\overline{x^j}$ $= \left\{\overline{x_1^j, \overline{x_2^j, \ldots x_m^j}}\right\}$ of class j,$\overline{x}_k^j A = \pi r^2$ is the center of the k-th attribute of the sample of this class:

$$\overline{x_k^j} = \sum_{i=1}^{\mathrm{n}_j} \frac{x_{ik}^j}{n_j}, (k = 1, 2, \ldots, m) \tag{13}$$

Let the classified sample universe be $X = \{x_1, x_2, \ldots, x_n\}$, and each meta- x_i consists of m attributes, that is, $x_i = \{x_{i1}, x_{i2}, \ldots x_{im}\}$, $(i = 1, 2, \ldots, n)$. Let the number of classes corresponding to the λ value be r, n_j be the number of samples in the j-th class, and the sample in the jth class be $\left\{x_1^j, x_2^j, \ldots, x_{n_j}^j\right\}$. The cluster center vector $\overline{x^j} = \left\{\overline{x_1^j}, \overline{x_2^j}, \ldots, \overline{x_m^j}\right\}$ of class j, $\overline{x_k^j}$ is the center of the k-th attribute of the sample of this class:

$$\overline{x_k^j} = \sum_{i=1}^{n_j} x_{ik}^j / n_j, (k = 1, 2, \ldots, m) \tag{14}$$

The population sample cluster center vector is: $\overline{x} = \{\overline{x_1}, \overline{x_2}, \ldots \overline{x_m}\}$, $\overline{x_k}$ is the center of the second attribute of the population sample:

$$\overline{x_k} = \sum_{i=1}^{n} x_{ik} / n$$

The distance from the i-th sample of the j-type center to the center of the sample is:

$$\left\|\overline{x_i^j} - \overline{x}\right\| = \sqrt{\sum_{k=1}^{m} \left(\overline{x_k^j} - \overline{x_k}\right)^2} \tag{15}$$

The distance from the j-type center to the overall center is:

$$\left\|\overline{x^j} - \overline{x}\right\| = \sqrt{\sum_{k=1}^{m} \left(\overline{x_k^j} - \overline{x_k}\right)^2} \tag{16}$$

The F statistic is defined as:

$$F = \frac{\sum_{j=1}^{r} n_j \frac{\left\|\overline{}\left\| x^j - \overline{x} \right\|\right\|}{r-1}}{\sum_{j=1}^{r} \sum_{i=1}^{n_j} \frac{\left\| x_i^j - \overline{x^j} \right\|}{r-1}} \tag{17}$$

The numerator of the F statistic indicates the distance between the various centers and the overall center; the denominator indicates the distance between the samples within the various types and the various centers: the larger the F value, the more reasonable the classification. Obviously $F \backsim F(r-1,$ $n1)$, the judgment criterion used in this article is: if $F_a > F_a(r-$ $1, n-r)$ .The corresponding classification is reasonable; if more than one Fr makes the above formula hold. The difference is calculated; if. $F_r > F_a(r-1, n-r)$. The classification of F_z can be determined to be appropriate.

4.5 Steps to improve fuzzy K-means clustering algorithm for big data

Step (1) **Data pre-processing:**

D, data sample set, is stored as a text file on an HDFS-based storage medium and then it is initialized as: {thing

Table 1 Accuracy prediction of accident casualties in different models

Model	Actual occurrences	Predicted occurrences	True times	Prediction success (%)
CNN	70	91	36	51.43
SVM	62	84	48	77.42
FK-means	88	101	60	68.18
TAP-CNN	78.5	86	65	60.15
Method in this paper	96	105	77	80.21

ID, attribute-value 1, attribute-value 2, ..., attribute-value n}.

Step (2) **Data block processing.**

D, which is initialized as data set, is evenly divided into n blocks, a serial number $\{D1, D2, \ldots, Dn\}$ is formulated for each block data set after the block while the dependence of the size of n is on the availability of storage space initially available.

Step (3) **Select initial *k* cluster centers of the data set D_i:**

First, suppose that there are *n* objects. $D_i = \{x_1, x_2, \ldots, x_n\}$ in given sample set, where x_i is initial candidate point and the distance between x_i and other sample points is calculated. Select the data point with a distance less than x_i. If it exists, calculate the "center of gravity" C_1 of all sample data points in range of x_i, as first cluster point. Otherwise, select point again. Then select the data *xj* furthest from C_1 as the second candidate point, and revise the process. "Center of gravity" of all points in the range of candidate point *xj* is repeatedly calculated as C_2, and so on until the *k* initial cluster centers are selected.

Step (4) **Cluster the sample set D_i:**

Read the initial cluster center file then use the MapReduce model to process the data in parallel. In the process of "Map function" design, set only single record per row to compare with the initial *k* selected, then compute the distance between every single sample data object and the cluster center in D_i, after that append the obtained object into distance. Cluster where the nearest cluster center is located, with the center-ID as the key and the records contained in the center as the value output. If$|J_c(I) - J_c(I+1)| < \varepsilon$, then explain the algorithm converged and the calculation stopped. Otherwise, iteration continues. During the iteration, the J_c value decreases gradually until it reach its minimum value.

Step (5) **Sample the data set D from this set:**

For each cluster obtained after clustering, use SOSS to select sample data from each cluster to form a sample set S. During this process the data in the sample set is presented as "{thing ID, attribute value 1, attribute value 2, ..., attribute value n}".

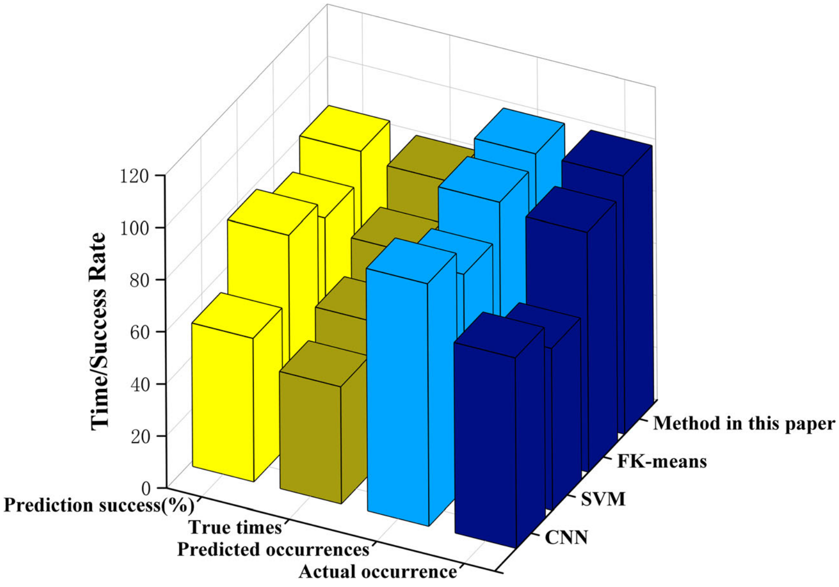


Fig. 4 Accuracy prediction of accident casualties in different models

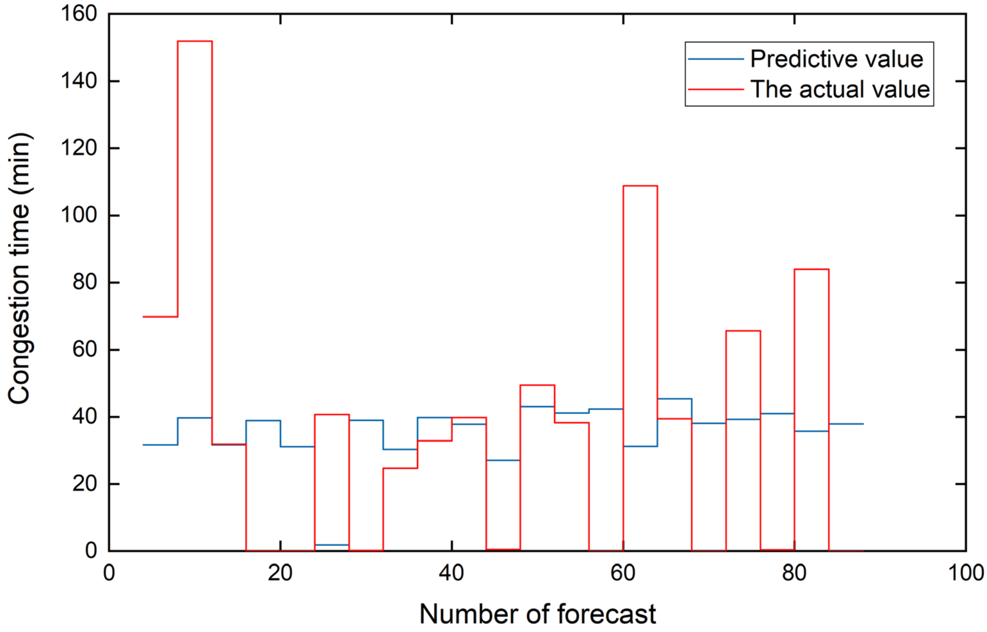


Fig. 5 Duration of accidents prediction results

5 Discussion

5.1 Traffic accident prediction and analysis

5.1.1 Forecast of traffic accident casualties

In the experiment, the accident-prone Jiangyang North Road to Tieshan Road was selected as the prediction object. The convolutional neural network, support vector machine, fuzzy K-means clustering and the improved model in this paper were used to predict the casualty situation. The prediction accuracy results are shown in Table 1 and Fig. 4. Due to the large sample size of the casualty data, the model performs very well during transaction and the accuracy is relatively high. Among them, the "convolutional neural network model" predicted a total of 36 casualty accidents. After combining with the actual 70 casualty accidents, the prediction success rate recorded only 51.43%. Although the accuracy of the support vector machine was low, the prediction success rate reached 77.42%; The success rate of fuzzy K-means algorithm is relatively inadequate, only 68.18%. The model's prediction success rate reached 80.21%.

Table 2 Two algorithms' results on dataset

Number of running	Optimized fuzzy k-means		Original fuzzy k-means	
	ID	Accuracy rate%	ID	Accuracy rate
20	18,225	93.88	6402	90.31
40	9365	85.26	33,151	81.64
60	25,663	88.34	18,227	59.86
80	36,074	72.61	14,000	64.23
100	447	77.92	15,369	80.59
120	10,091	70.49	27,022	77.61
140	12,717	72.43	337	72.50
160	7399	80.01	9008	69.99
180	21,051	89.39	11,089	78.80
200	17,668	93.21	29,756	71.34

5.1.2 Predicting the handling time of traffic accident

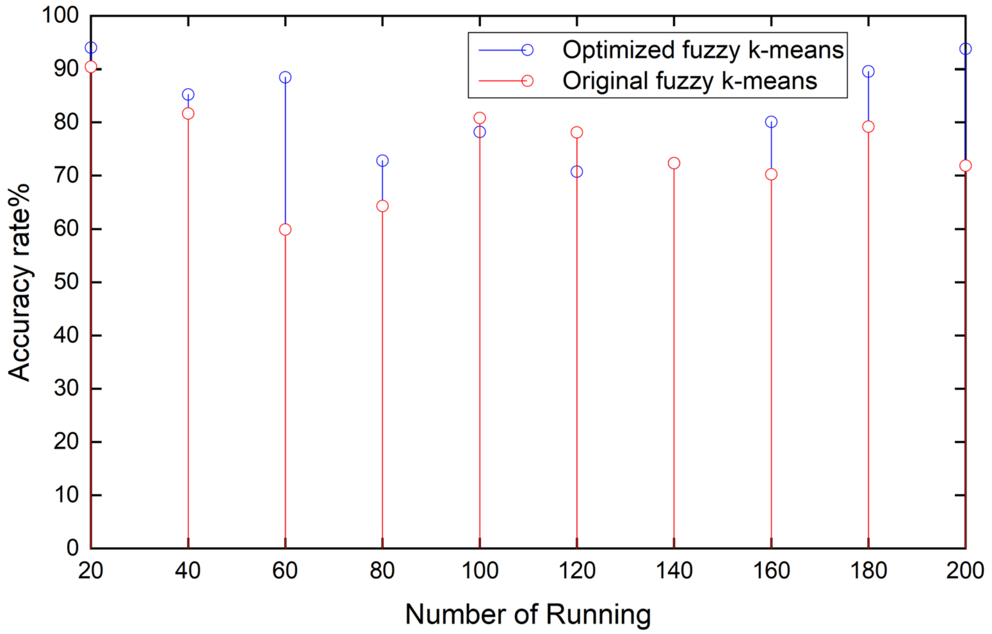
The duration of a traffic accident refers to the total time from the occurrence of traffic accident to the restoration of traffic. It can be divided into three phases: accident detection and response, accident clearing phase traffic recovery phase. Usually, the incident response processing time and the congestion dissipating time are two Indicators to describe the impact of congestion over time. The congestion duration is the sum of the incident response processing time and the congestion dissipation time. The prediction result of the traffic accident processing time is shown in Fig. 5.

5.2 "Improved fuzzy K-means algorithm "performance analysis

5.2.1 Analysis of traffic accident prediction accuracy

The original "Fuzzy K-means algorithm" and "optimized fuzzy K-means algorithm" were run 200 times each. The results are presented in Table 2 and Fig. 6. This article uses $\rho = \frac{C}{N} \times 100\%$ to calculate the accuracy of the two algorithms, where C represents the number of data objects that can be correctly assigned to the specified class while N represents total number of sample data objects. It is observed that the original fuzzy k-means

Fig. 6 Test results of two algorithms on dataset

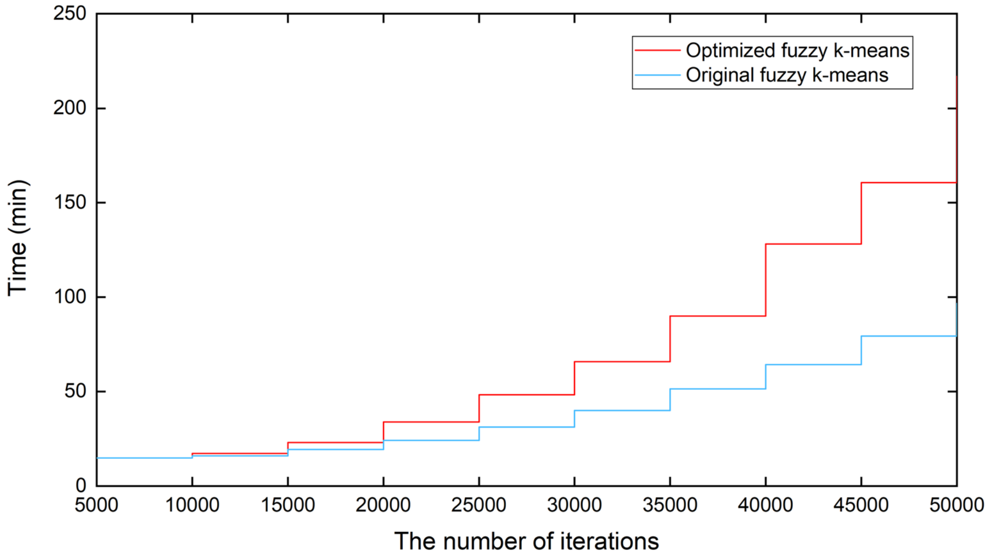


algorithm holds the accuracy of 90.31% at the highest, 59.86% at the lowest while the accuracy on average is 75.08%. Range of accuracy varies greatly and the cause is original fuzzy K-means algorithm. The random selection method is used to select the initial clustering center point, which does not guarantee that the clustering center selected each time is reasonable. Initial center point will easily lead to reduction in the accuracy of the clustering result If it is not proper. Highest accuracy of improved algorithm is 93.88%, the lowest is 70.49%, and the average is 82.19%. The overall accuracy of the algorithm is higher than the original "fuzzy K-means algorithm", which indicates that "initial center point" selected by the "improved algorithm" is more reasonable and the obtained clustering results are more accurate.

5.2.2 Analysis of the number of iterations

The comparison of the two algorithms in the number of iterations is shown in Fig. 7. Observation shows that if initial clustering centers selected are different, then the number of iterations of the algorithm is also different. Number of iterations of the original fuzzy K-means algorithm is relatively unstable, with large fluctuations. This phenomenon shows that if the initial cluster center selected is far from the actual center point of each cluster, it will lead to a slow convergence of the objective function and increase the number of algorithm iterations. The improved algorithm proposed in this paper is significantly less than

Fig. 7 Relationship between algorithm runtime and data volume



the original fuzzy K-means algorithm in terms of the number of iterations. It shows that the improved algorithm obtains the initial cluster center point which is close to the cluster center point of the actual data, which makes the algorithm converge faster, accelerates the clustering process, and obtains a more stable clustering result.

6 Conclusions

Fuzzy clustering analysis is an important means and a method for partitioning data or packet processing, pattern recognition belongs to the Central African supervised pattern recognition category. The application of fuzzy clustering technique to analyze causes of accident black spots, the road traffic is forecast to solve the problem of effective method. Presented to some work, the main study was the initialization problem of fuzzy clustering and fuzzy clustering algorithm based on information cooperation, with examples analyzed.

This paper considers clustering on feature space by aiming at the inefficiency caused by fuzzy clustering algorithm in high-dimensional data. Unlike traditional methods, the algorithm in the article can make the fuzzy clustering accuracy on the feature space and the fuzzy clustering accuracy on the original data consistent, that is, the dimensionality reduction (feature space) process and clustering process of big data Are matching. At the same time, when the clustering center and the distance function meet certain constraints, the fuzzy K-means algorithm in the feature space can be accelerated, while ensuring the efficiency of the algorithm; it also effectively improves the efficiency of the algorithm, making the application on big data become may.

The number of iterations of the original fuzzy K-means algorithm is more unstable, large fluctuations. This phenomenon indicates that, if the selected initial cluster centers from the actual center point of each cluster far will lead to slow convergence objective function, resulting in increased number of iterations of the algorithm. The improvements proposed fuzzy K-means algorithm in terms of the number of iterations significantly less than the original fuzzy K-means algorithm. This paper improved the description of the algorithm and made the cluster center point of the actual data closer to the initial cluster centers, so that the algorithm converges faster, speed up the process of clustering, has been relatively stable clustering results.

Acknowledgments This work was supported by Scientific Research Project of Education Department of Shaanxi Provincial Government (Project No. 18JK0450), and Natural Science Basic Research Plan of Shaanxi Province (Project No. 2020JQ-682).

Funding Scientific Research Project of Education Department of Shaanxi Provincial Government (Project No. 18JK0450) ; Natural Science Basic Research Plan of Shaanxi Province (Project No. 2020JQ-682).

References

1. Twala B (2014) Extracting grey relational systems from incomplete road traffic accidents data: the case of Gauteng Province in South Africa. Expert Syst 31(3):117
2. Zhao Y-F, Zhang S-R, Ma Z-L (2018) Analysis of traffic accident severity on highway tunnels using the partial proportion odds model. China Journal of Highway and Transport 31(9): 159–166
3. Lei Y, Yuan L (2014) Traffic incident duration prediction based on adaptive neural-fuzzy inference system. Comput Eng 40(2):189–192,198
4. Yousefzadeh-Chabok S, Hosseinpour M, Kouchakinejad-Eramsadati L et al (2016) Comparison of revised trauma score, injury severity score and trauma and injury severity score for mortality prediction in elderly trauma patients. Ulusal travma ve acil cerrahi dergisi = Turkish Journal of Trauma & Emergency Surgery: TJTES 22(6):536–540
5. Gokulakrishnan P, Ganeshkumar P (2015) Road accident prevention with instant emergency warning message dissemination in vehicular ad-hoc network. PLoS One 10(12):e0143383
6. Taamneh M, Taamneh S, Alkheder S (2016) Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. Int J Inj Control Saf Promot 24(3):1–8
7. Lee SL (2015) Predicting traffic accident severity using classification techniques. Adv Sci Lett 21(10):3128–3131
8. En ZM (2015) A prediction model for traffic accident proneness caused by drunk driving via numerical characteristics of drivers' EEGs. Journal of Shanghai Jiaotong University(Science) 49(2): 287–292
9. Gour D, Kanskar A (2019) Automated AI based road traffic accident alert system: YOLO Algorithm. IJSTR, August 2019, ISSN 2277–8616
10. Priya V, Priya C (2019) A cognitive contemplation of road accident prediction through deep learning. IJITEE, December 2019, ISSN: 2278–3075
11. Wenqi L, Dongyu L, Menghua Y (2017) A model of traffic accident prediction based on convolutional neural network[C]//2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE): 198–202
12. Liao C, Shou G, Liu Y, Hu Y, Guo Z (2017) Intelligent traffic accident detection system based on mobile edge computing[C]// 2017 3rd IEEE InternationalConference on Computer and Communications (ICCC): 2110–2115
13. De Ona J, Mujalli RO, Calvo FJ (2011) Analysis of traffic accident injury severity on Spanish rural highways using Bayesian Network. Accidents Analysis and Prevention 43(2011):402–411
14. Chong MM, Abraham A, Paprzycki M (2014) Traffic accidents analysis using decision trees and neural networks. IJITCS 6(2):22–28
15. Lee J-Y, Chung J-H, Son B (1955–1963) Analysis of traffic accident size for Korean highway using structural equation models. Accid Anal Prev 40(2008):2008
16. Abdelwahab HT, Abdel-Aty MA (2017) Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation research record 1746 paper no. 01–2234

17. Dogru N, Subasi A (2018) Traffic accident detection using random forest classifier.15th L&T Conference, 2018
18. Nair SC, Elayidom MS, Gopalan S (2019) KM-MBFO: a hybrid hadoop map reduce access for clustering big data by adopting modified bacterial foraging optimization algorithm. IJRTE, ISN:2277–3878, September 2019
19. Olutayo VA, Eludire AA (2014) Traffic accident analysis using decision trees and neural networks. International Journal of Information Technology and Computer Science 2: 22–28
20. Bhatti F, Shah MA, Maple C, Islam SU (2019) A novel internet of things-enabled accident detection and reporting system for smart city environments. Sensors 19(9):2071
21. Al-Balushi IA, Yousif JH, Al-Shezawi MO (2017) Car accident notification based on Mobile cloud computing. Social Science Electronic Publishing 2(2):46–50
22. Bera S, Misra S, Rodrigues JJPC (2015) Cloud computing applications for smart grid: a survey. Parallel & Distributed Systems IEEE Transactions on 26(5):1477–1494
23. Toosi AN (2014) Interconnected cloud computing environments: challenges, taxonomy, and survey. ACM Comput Surv 47(1):1–47
24. Oguchi M, Hara R (2016) A speculative control mechanism of cloud computing systems based on emergency disaster information using SDN. Procedia Comput Sci 98:515–521
25. Tamura Y, Yamada S (2016) Reliability computing and management considering the network traffic for a cloud computing. Ann Oper Res 244(1):163–176
26. Bowen D, Huang R, Xie Z (2018) KID model-driven things-edge-cloud computing paradigm for traffic data as a service. IEEE Netw 32(1):34–41
27. Abbass H, Tang J, Amin R, Ellejmi M, Kirby S (2014) The computational air traffic control brain: computational red teaming and big data for real-time seamless brain-traffic integration. J Air Traffic Control 56(2):10–17
28. Huang Z, Cao F, Jin C, Yu Z, & Huang R (2017) Carbon emission flow from self-driving tours and its spatial relationship with scenic spots–A traffic-related big data method. J Clean Prod 142:946–955
29. Balaji Ganesh R (2015) An intelligent video surveillance framework with big data management for Indian road traffic system. Int J Comput Appl 123(10):12–19
30. Bortyakov DE, Mescheryakov SV, Shchemelinin DA (2014) Integrated management of big data traffic systems in distributed production environments. Spbspu Journal Computer Science Telecommunication & Control Systems, pp 105–113
31. Weihua C, Jun Z, Peng W (2017) Network traffic detection and analysis based on big data flow. J Nanjing Univ Sci Technol 41(3):294–300
32. Wang L-L, Ngan HYT, Yung NHC (2015) Automatic incident classification for big traffic data by adaptive boosting SVM. Inf Sci:467

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zhun Tian is a lecturer at the School of Transportation Engineering, Chang'an University. Her research fields are traffic safety, intelligent traffic system, and traffic management.

Shengrui Zhang is a professor at the School of Transportation Engineering, Chang'an University. His research interests include traffic planning, road safety, and intelligent transportation systems.