International Journal of Hybrid Information Technology
Vol.9, No.3 (2016), pp. 301-310
http://dx.doi.org/10.14257/ijhit.2016.9.3.28

Anomaly Driving Speed Detection and Correction Algorithm based on Quantiles and KNN

Guo Yanling[1], Liu Lichen[1,2], Gao Meng[3] and Gao Lewen[4]

[1]College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, China
[2]Harbin Kejia Universal Electric Mechanical Corporation Co., Ltd., Harbin, China
[3]China Mobile Communications Corporation Co., Ltd. of Heilongjiang Branch, Harbin, China
[4]China Mobile Communications Corporation Design Institute Co., Ltd. of Heilongjiang Branch, Harbin, China
liulc1983@163.com

Abstract

Driving speed is a key parameter for building the traffic state identification model, its precision directly affects the model reliability and the traffic state identification accuracy. Aiming at the standard normal deviation method's defects in dealing with the extreme noise data, an anomaly driving speed detection algorithm based on quantiles is proposed, use historical data to establish the exception borders which are used to detect whether an unknown data is abnormal; on the basis of the abnormal data detection, a driving speed prediction algorithm based on improved KNN is proposed, use K-means algorithm to clustering the historical data, and predict the next moment's speed according to the distance between the data to be predicted and the clusters, the predicted speed can be used to correct the abnormal speed. Experimental results show that the detection rate of the proposed anomaly detection algorithm has improved about 4.25% compared with the standard normal deviation method, and the false detection rate has reduced about 25.51%; the mean relative error of the proposed speed prediction algorithm is 13.69%, it can predict the driving speed well, namely, the anomaly driving speed detection and correction algorithm based on quantiles and KNN is feasible and effective.

Keywords: *Driving speed, Anomaly detection, Anomaly correction, Quantiles, KNN*

1. Introduction

The rapid increase of vehicles has brought huge pressure for urban road traffic, with the fast development of modern information technologies, intelligent transportation system (ITS) is becoming an effective way to solve the traffic congestion [1], and the collected data can provide transportation decision support for different level's management. Traffic state identification is a key link to achieve the intelligent transportation, many scholars have established different traffic state identification models, in which driving speed is a critical parameter, and it can directly influence the model reliability. At present, the speed data is usually collected by GPS-based floating car, however, GPS has the signal blind area defect [2], and data accuracy is easily affected by different factors such as weather and road conditions, therefore, the collected data always has noises. In order to improve the data accuracy, it needs to detect and correct the data noises.

Transportation Institute of California Berkeley University compares the commonly used traffic data anomaly detection algorithms, they think that the effect

ISSN: 1738-9968 IJHIT

of the standard normal deviation (SND) is the best [3], but SND has effect in dealing with samples existed the extreme noises. Lichao Wang [4], Corrado de Fabritiis [5], Zhu Yan [6] *etc.*, have established different prediction models related with driving speed, these models have achieved good accuracy. However, the combination research of anomaly detection and correction is less, therefore, this paper puts forward a kind of anomaly driving speed detection and correction algorithm based on quantiles and KNN, its algorithm process is shown in Figure 1, use quantile detection method to find abnormal speed, use KNN algorithm to predict driving speed, and use the predicted speed to correct the abnormal data, this simulation results show that the proposed algorithm can effectively improve the accuracy of the collected speed data.

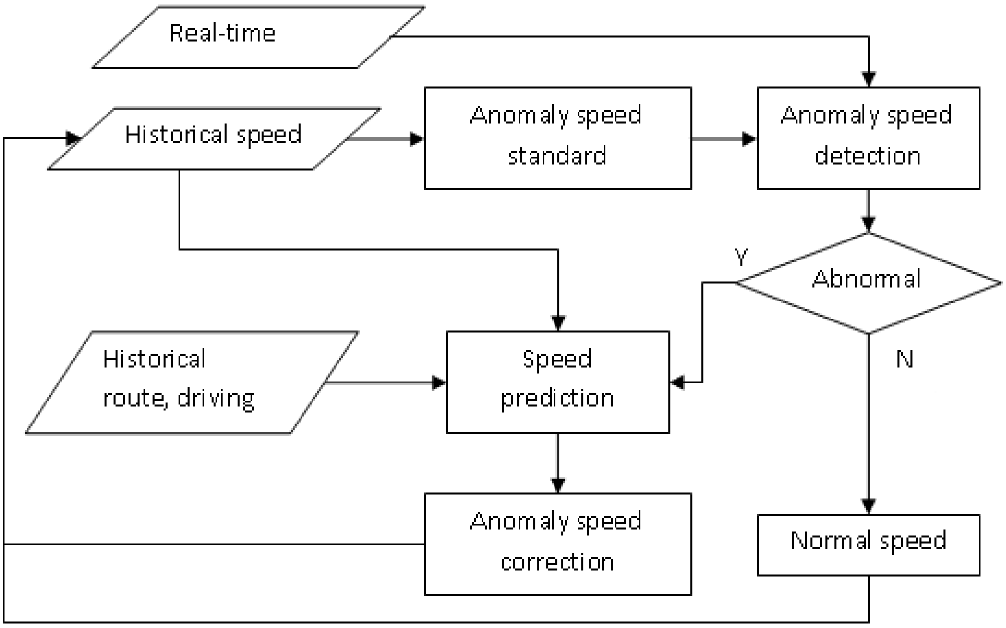

Figure 1. Process of the Proposed Algorithm

2. Anomaly DRIVING speed Detection Algorithm based on Quantiles

When the collected speed data existing extreme noises, the mean value and variance of the data center will be affected, therefore, SND method can not well reflect the difference between the center and the abnormal value, it can affect the judging of abnormal value when the new data is coming. Aiming at this problem, use quantiles [7, 8] of historical data to build the exception borders, and use quantile range to judge whether a new data is abnormal, this method can achieve a certain degree of anti-interference.

Some studies have shown that the driving speed of urban road traffic approximately obeys the normal distribution, assume that the normal distribution obeyed speed samples is X , its distribution function is:

$$F(X) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} d_t, -\infty < x < +\infty \tag{1}$$

Assume that x_1, x_2 are respectively the 1/4-quantile and 3/4-quantile, set $x' = x_2 - x_1$, and anomaly driving speed borders are:

$$\begin{cases} y_{\min} = x_1 - 1.5x' \\ Y_{\min} = x_1 - 3x' \\ y_{\max} = x_2 + 1.5x' \\ Y_{\max} = x_2 + 3x' \end{cases} \quad (2)$$

, among which $y_{\min}$ is the minimum anomaly border; $Y_{\min}$ is the extremely minimum anomaly border; $y_{\max}$ is the maximum anomaly border; $Y_{\max}$ is the extremely maximum anomaly border.

Assume that v_{dt} is the speed to be detected which is collected at moment t on date d, $v_{it}(i<d)$ is the historical driving speed at moment t on the current driving road, $\overline{v_{(i)t}}$ is the mean historical driving speed at moment t on the current driving road, the anomaly driving speed detection algorithm based on quantiles can be described as follows.

(1) Calculate the amount of historical driving speed data v_{it} corresponding with v_{dt}, and the amount is expressed by k;

(2) According to $\Delta v_{i(i-1)} = v_{it} - \overline{v_{(i-1)t}}$ and $\overline{v_{(i-1)t}} = \frac{\sum_{j<i-1} v_{jt}}{k-1}$, calculate the difference of historical driving speed and average historical driving speed;

(3) Do K-S normality test on the $k-1$ $\Delta v_{i(i-1)}$, and find its 1/4-quantile x_1 and 3/4-quantile x_2;

(4) Calculate the $y_{\min}$, $Y_{\min}$, $y_{\max}$ and $Y_{\max}$ of $\Delta v_{i(i-1)}$ set;

(5) Build the anomaly discrimination standard, value located in $[y_{\min}, Y_{\min}] \cup [y_{\max}, Y_{\max}]$ is the abnormal data, value located outside of $[y_{\min}, y_{\max}]$ is the extremely abnormal data;

(6) According to $\Delta v_{d(i)} = v_{dt} - \overline{v_{(i)t}}$ and $\overline{v_{(i)t}} = \frac{\sum_{j<i} v_{jt}}{k}$, calculate the driving speed difference to be discriminated;

(7) If $\Delta v_{d(i)} \in [y_{\min}, Y_{\min}] \cup [y_{\max}, Y_{\max}]$, then v_{dt} is the abnormal data; if $\Delta v_{d(i)} \notin [Y_{\min}, Y_{\max}]$, then v_{dt} is the extremely abnormal data; if $\Delta v_{d(i)} \in [y_{\min}, y_{\max}]$, then v_{dt} is the normal data.

3. Anomaly Driving Speed Correction Algorithm based on Improved KNN

Driving speed is affected by historical road traffic and the passed roads' traffic, the traffic has large uncertainty, nonlinearity and complexity, therefore, it is difficult to accurately reflect and predict the traffic situation by certain mathematical models, these models' timeliness and accuracy can not be guaranteed in practice. KNN is a kind of nonparametric regression prediction method without using mathematical models [9], it is based on historical data without any prior knowledge, and it is easy to add new data to the model, therefore, KNN is suitable for driving speed prediction. This paper proposes an improved KNN driving speed prediction algorithm based on clustering analysis, and uses the predicted speed to correct the detected abnormal speed, which can improve the accuracy of the underlying speed data.

3.1 Neighbors Selection Method based on K-means Algorithm

According to some similarity indexes, clustering analysis can automatically divide a non-identified data into a similar class under the condition of lacking prior knowledge. Aiming at the driving speed data, clustering the historical data related with driving speed, when predicting the next moment's speed, the obtained clusters can be used to finish the prediction.

3.1.1 Determination of the Target Clustering Number: In traditional K-means algorithm, the target clustering number k needs to be given in advance, and it is difficult to determine its value when lacking enough prior knowledge. This paper adopts a kind of BWP clustering effectiveness index to quantitatively evaluate the effect of the clustering results [10], the optimal clustering number is determined by comparing the BWP value under different k.

Assume that $K=\{X,R\}$ is the clustering space, $X=\{x_1,x_2,...,x_n\}$, n samples are divided into c classes, then,

$$b(j,i)=\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\left\|x_p^{(k)}-x_i^{(j)}\right\|^2\right) \tag{3}$$

$$w(j,i)=\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\left\|x_q^{(j)}-x_i^{(j)}\right\|^2 \tag{4}$$

$$baw(j,i)=b(j,i)+w(j,i) \tag{5}$$

$$bsw(j,i)=b(j,i)-w(j,i) \tag{6}$$

$$BWP(j,i)=\frac{bsw(j,i)}{baw(j,i)} \tag{7}$$

$$avgBWP(k)=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}BWP(j,i) \tag{8}$$

$$k_{opt}=\arg\max_{2\le k<n}\{avgBWP(k)\} \tag{9}$$

Among which k and j is the cluster label, $x_i^{(j)}$ is the j^{th} class's i^{th} sample, $x_p^{(k)}$ is the k^{th} class's p^{th} sample, n_k is the total samples' amount of the k^{th} class, $\|\cdot\|^2$ is the squared Euclidean distance, $x_q^{(j)}$ is the j^{th} class's q^{th} sample, and $q\ne i$, n_j is the total samples' amount of the j^{th} class, $b(j,i)$ is the minimum between-class distance of $x_i^{(j)}$, $w(j,i)$ is the inner-class distance of $x_i^{(j)}$, $baw(j,i)$ is the cluster distance of $x_i^{(j)}$, $bsw(j,i)$ is the cluster deviation distance of $x_i^{(j)}$, $BWP(j,i)$ is the between-class and inner-class division of $x_i^{(j)}$, $avgBWP(k)$ is the mean BWP value of the k^{th} class, k_{opt} is the optimal clustering number.

It can be seen from formula (3)-(9) that $w(j,i)$ reflects the inner-class tight degree, and $b(j,i)$ reflects the between-class distant degree. For the sake of a better clustering effect, it needs to make $w(j,i)$ to be as lower as possible and make $b(j,i)$ to be as higher as possible. In order to simultaneously consider the two aspects, design $bsw(j,i)$ to make the two functions consistent, a better clustering effect is corresponding with a higher $bsw(j,i)$. In order to eliminate the dimensions' influence on $bsw(j,i)$, introduce

$baw(j,i)$ to compress $bsw(j,i)$, and get $BWP(j,i)$ whose value is distributed between $[-1,1]$, it is convenient for effectiveness comparison and analysis.

For a dataset, its clustering effect needs to be evaluated by the mean value $avgBWP(k)$ of each single sample's BWP value, and a better clustering effect is corresponding with a higher $avgBWP(k)$, clustering number with the maximum $avgBWP(k)$ is the optimal k value.

3.1.2 Neighbors Selection based on Improved K-means Algorithm: Assume that historical dataset is $X=\{x_1,x_2,...,x_n\}$, and samples' amount is n; $[k_1,k_2]$ is the searching scope of the optimal clustering number k, according to empirical formula, set $k_1=2$ and $k_2=Int\sqrt{n}$; K is the neighbors' amount, according to empirical formula, set $K=Int\sqrt{n}$. The neighbors' selection process based on improved K-means algorithm is shown as follows.

(1) Set k and K, at initial time, $k=2$, if $k \le k_2$, turn to step (2); else, turn to step (7);

(2) Randomly select k samples from $X=\{x_1,x_2,...,x_n\}$ as the initial clustering centers $(c_1,c_2,...,c_k)$, each center is corresponding with a class C_j;

(3) Calculate the Euclidean distance d_{ij} between each x_i and $(c_1,c_2,...,c_k)$;

(4) According to $\min(d_{ij})$, divide x_i into the class C_j corresponding with c_j;

(5) Recalculate the mean value of all data objects in C_j, and use it as the new center c_j;

(6) Repeat step (3)-(5), re-divide all the objects until the generated classes do not change anymore, the classes set is C;

(7) Calculate the distance d_j between sample x' to be marked and center c_j of each class;

(8) Calculate $\min(d_j)$, and add the samples of class C_j corresponding with $\min(d_j)$ to a new set P;

(9) Calculate the distance d_m between x' and each sample x_m in P;

(10) Find sample x_m which meets $d_m < \min(d_j)$ and transfer it to a new set N until the samples number of N reaches K;

(11) Calculate the classes of the K neighbors, and determine which class the unknown sample belongs to.

3.2 Measurement of Distance and Determination of Weight Function

3.2.1 Measurement of Distance: After finding out the neighbors of the sample to be predicted, it needs to calculate the distance between them, it can be reflected by Euclidean distance which is shown as follows.

$$d_i = \sqrt{\sum_j (X_j - X_{ji})^2} \tag{10}$$

In which d_i is the distance between sample to be predicted and the i^{th} group data of neighbors; X_j is the value of the j^{th} sub item of the sample to be predicted; X_{ji} is value of the i^{th} sample's j^{th} sub item.

3.2.2 Determination of Weight Function: As formula (10) involves square root operation, in order to guarantee the accuracy, according to distance between neighbors and sample to be predicted, respectively give different weights to the neighbors, and the weight calculation method [11] is shown as follows.

$$w_i(x') = \frac{1/d_i^2}{\sum_{i=1}^{K} 1/d_i^2} \tag{11}$$

In which d_i is the distance between sample to be predicted and the i^{th} group data of neighbors; K is the neighbors' amount.

3.3 The Improved Anomaly Driving Speed Correction Algorithm based on KNN

Based on anomaly driving speed, use K-means algorithm to clustering the historical driving speed, select K historical samples which are nearest to the anomaly sample, calculate the weighted distance between anomaly sample and its neighbors, use the weighted values of the K neighbors to predict the anomaly sample, and use the predicted value to correct the anomaly value, which can achieve the goal of improving the driving speed data accuracy.

Assume that $v(t)$ is the driving speed at moment t; $v(t-1)$ is the driving speed at the prior moment $t-1$; $\overline{v(t)}$ is the mean road speed between moment t and $t-1$; $d_1(t)$ is the distance between vehicle and the front intersection at moment t; $d_2(t)$ is the distance between vehicle and the rear intersection at moment t; $T(t)$ is the driving time. The improved anomaly driving speed correction algorithm based on KNN is shown as follows.

(1) Construct state vector $X = [v(t), v(t-1), \overline{v(t)}, d_1(t), d_2(t), T(t)]$;

(2) According to algorithm discussed in 3.1.2, find K neighbors of X, the state vector is $X_i = [v_i(t), v_i(t-1), \overline{v_i(t)}, d_{1i}(t), d_{2_i}(t), T_i(t)], i = 1,2,...,K$;

(3) According to formula (10), separately calculate the distance d_i between K neighbors and X;

(4) According to formula (11), separately calculate the weight w_i of neighbor K_i;

(5) According to $v(t+1) = \sum_{i=1}^{K} w_i v_i(t)$, calculate the next moment's driving speed;

(6) Use $v(t+1)$ to correct the detected anomaly speed.

4. Simulation Experiment and Analysis

4.1 Anomaly Driving Speed Detection Experiment and Analysis

Choose the road between Zhengyi Road and Haping Road crossing with Hexing Road of Harbin as the study area, each vehicle's speed is collected by the on-board sensor nodes, time interval of data collection is 1min, use data collected from March 1 to April 25 of 2015 to do simulation experiments.

Based on the historical data, calculate the exception borders of real-time driving speed difference at each moment, when a new data is coming, according to the borders, detect and judge whether the new data is abnormal.

Regard judging whether the driving speed 5.32m/s collected at 10:52:30 on April 20, 2015 is the abnormal data as the example, find the historical speed data collected at 10:52:30, according to the algorithm based on quantiles to calculate the 1/4-quantile and

3/4-quantile of the historical driving speed, 1/4-quantile $x_1 = -0.784$, 3/4-quantile $x_2 = 1.662$, then, $x' = 2.446$, $y_{\min} = -4.453$, $Y_{\min} = -8.122$, $y_{\max} = 5.331$, $Y_{\max} = 8.94$, calculate the real-time driving speed difference, its value is -5.042 which is in $[y_{\min}, Y_{\min}] \cup [y_{\max}, Y_{\max}]$, namely, the speed 5.32m/s is an abnormal speed.

According to the anomaly detection results, comparing with the SND method in detection rate and false detection rate, the comparison results are shown in Figure 2 and Figure 3.

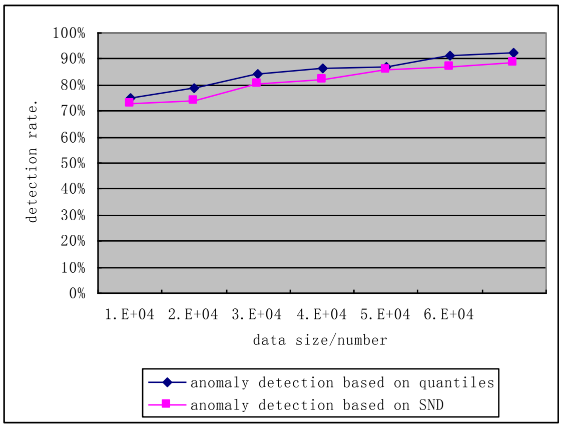

Figure 2. Detection Rate Comparison

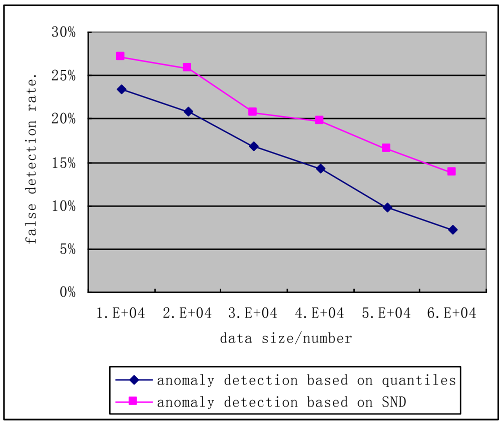

Figure 3. False Detection Rate Comparison

It can be seen from Figure 2 and Figure 3 that the improved algorithm based on quantiles is better than the SND method on detection rate and false detection rate. The mean detection rate of the improved algorithm is 85.0457% while SND method is 81.5793%, it has improved about 4.25%; the mean false detection rate of the improved algorithm is 13.1871% while SND method is 17.7043%, it has reduced about 25.51%. In addition, with the increasing of historical data amount, detection rate is increasing while false detection rate is reducing, which means enough historical data is benefit for detecting anomaly data.

4.2 Driving Speed Predicting Experiment and Analysis

In order to validate the performance of the improved driving speed prediction algorithm based on KNN, choose the road between Linxing Road and Zhengyi Road crossing with Hexing Road, Yanxing Road and Hexing Road crossing with Zhengyi

Road, Wenchang Road/Wen Zheng Road intersection and Hexing Road/Zhengyi Road intersection, Zhengyi Road and Heping Road crossing with Hexing Road as the study area, use data collected from 14:50:30 to 15:35:30 on April 20, 2015 to do simulation experiments.

Based on the historical data, use K-means algorithm to clustering, the historical data can be divided into 5 classes; sort the driving data of vehicle to be predicted by time, find the first record, and use the improved KNN algorithm to predict the driving speed at next moment.

Regard predicting the speed of 10 vehicles which are driving on the road between Hexing Road/Zhengyi Road intersection and Hexing Road/Heping Road intersection at 15:30:30 on April 20, 2015 as the example, their real driving speeds at current moment, next moment and the predicted speed at next moment are shown in Table 1.

Table 1. Prediction Results of the Improved KNN Algorithm

Vehicle No.	Speed at current moment （km/h ）	Real speed at next moment （km/h ）	Predicted speed at next moment （km/h）	Vehicle No.	Speed at current moment （km/h ）	Real speed at next moment （km/h ）	Predicted speed at next moment （km/h）
1	10.26	20.13	25.46	6	0	29.54	36.44
2	20.24	19.66	23.17	7	7.67	28.79	31.43
3	25.72	16.75	20.83	8	40.32	2.32	2.69
4	9.65	9.75	9.31	9	40.34	44.66	50.63
5	34.76	25.17	24.74	10	30.21	33.43	33.37

The comparison of the predicted speed and the real speed of 10 vehicles is shown in Figure 4, relative error of the predicted speed and the real speed is shown in Figure 5.

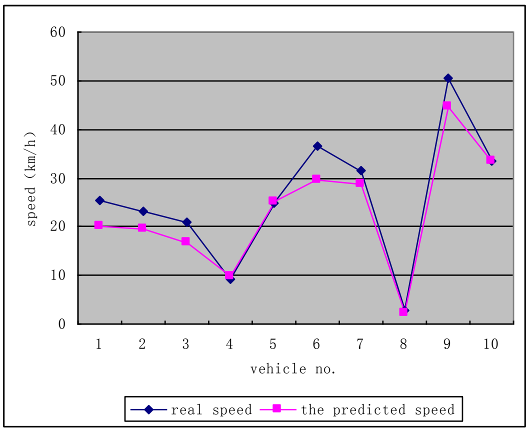

Figure 4. Speed Comparison Result

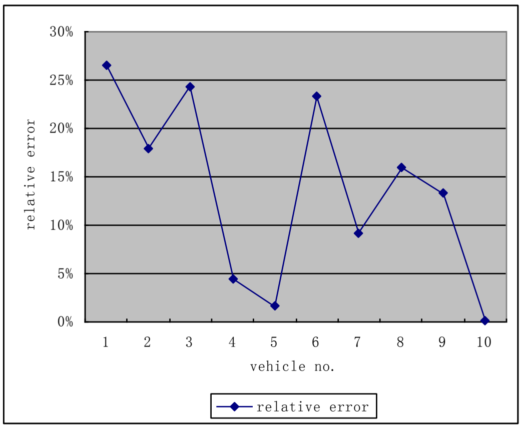

Figure 5. Relative Error Result

It can be seen from Figure 4 that the algorithm has a better performance, and the prediction results can fit the driving speed trend well. It can be seen from Figure 5 that the relative error of the real speed and predicted speed is under 26.47%, the minimum relative error is 0.1795%, and the mean relative error is 13.69%. The experimental results show that the improved driving speed prediction algorithm based on KNN has a good prediction accuracy, which can be used to correct the detected abnormal speed data.

5. Conclusions

SND method has defects in dealing with data with extreme noises, it can not effectively distinguish the sample center and abnormal value, this paper proposes an anomaly driving speed detection algorithm based on quantiles, and on the basis of the anomaly speed detection, this paper further puts forward an improved driving speed prediction algorithm based on KNN, which is used to correct the abnormal speed. This paper has got several conclusions as follows:

(1) The anomaly driving speed detection algorithm based on quantiles can effectively detect the anomaly speed, comparing with SND method, the detection rate has improved about 4.25% and false detection rate has reduced about 25.51%.

(2) The improved driving speed prediction algorithm based on KNN can well predict the driving speed, and the mean relative error is 13.69%, which means that the algorithm can be used to correct the abnormal speed.

This paper uses K-means algorithm to improve the KNN algorithm and gets an satisfied prediction accuracy, however, the running efficiency of the algorithm needs to be improved, because the determination of neighbors amount is according to empirical formula, the subsequent research should focus on how to determine this value, through shrink the searching scope to improve the efficiency of the algorithm.

Acknowledgement

This work is supported by The National 948 Project (2011-4-11) and The Fundamental Research Funds for the Central Universities (2572014AB23).

References

[1] K. Malecki, "Stanislaw lawn and Kinga Kijewska, Influence of Intelligent Transportation Systems on Reduction of the Environment Negative impact of Urban Freight Transport Based on Szczecin Example", Procedia-Social and Behavioral Sciences, vol. 10, **(2014)**, pp. 215-229.

[2] U. Dogan, M. Uluday and D.O. Demir, "Investigation of GPS positioning accuracy during the seasonal variation", Measurement, vol 7, **(2014)**, pp. 91-100.

[3] L. Xingquan, O. Yangjun and L. Feng, "Exploratory analysis of floating vehicle's driving speed", Science of Surveying and Mapping, vol. 09, **(2012)**, pp. 160-163.
[4] L. Wang, Q. Lu and X. Chen, "Application and Analysis of Time Domain cross Correlation for Traffic Flow Speed Measurement", 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, vol. 7, **(2007)**, pp. 274-279.
[5] C. De Fabritiis and G. V. R. Ragona, "Traffic Estimation and Prediction Based on Real Time Floating Car Data", Intelligent Transportation Systems, vol. 10, **(2008)**, pp. 197-203.
[6] Z. Li and Y. Dongyuan, "Dynamic Travel Speed Collection Technology based on Low Frequence FCD", Journal of Transportation Systems Engineering and Information Technology, vol. 8, **(2008)**, pp. 42-48.
[7] C. Yuzhi, "Extreme value prediction via a quantile function model", Coastal Engineering, vol. 7, **(2013)**, 91-98.
[8] H. Wu, L. Gao and Z. Zhang, "Analysis of crash data using quantile regression for counts", Journal of Transportation Engineering, vol. 4, **(2014)**, pp. 140-146.
[9] P. Nanbo, Z. Yanxia and Z. Yongheng, "A SVM-KNN method for quasar-star classification. Science China: Physics", Mechanics and Astronomy, vol. 6, **(2013)**, pp. 1227-1234.
[10] M. K. Pakhira, "A fast k-means algorithm using cluster shifting to produce compact and separate clusters", International Journal of Engineering, Transaction A: Basics, vol. 1, **(2015)**, pp. 36-45.
[11] J. Liangxiao, C. Zhihua and W. Dianhong, "Bayesian Citation-KNN with distance weighting", International Journal of Machine Learning and Cybernetics, vol. 2, **(2014)**, pp. 193-199.

Authors

Guo Yan-ling, Female, born in 1962, supervisor, working in College of Mechanical and Electrical Engineering of Northeast Forestry University, mainly engaged in electromechanical integration technology, Agriculture and forestry picking machine and robot technology.

Liu Li-chen, Male, born in 1983, Ph.D., studying in College of Mechanical and Electrical Engineering of Northeast Forestry University, working in Harbin Kejia Universal Electric Mechanical Corporation Co., Ltd., mainly engaged in electromechanical integration technology , vehicle control.

Gao Meng, Female, born in 1989, Ph.D., studied in Information and Computer Engineering College of Northeast Forestry University, working in China Mobile Communications Corporation Co., Ltd. of Heilongjiang Branch, mainly engaged in data mining and data analysis.

Gao Lewen, Male, born in 1981, Master's Degree, studied in Information and Computer Engineering College of Northeast Forestry University, working in China Mobile Communications Corporation Co., Ltd. of Heilongjiang Branch, mainly engaged in Engineering consulting and design.

