Data Fault Identification and Repair Method of Traffic Detector

LI Xiao-lu[1], CHEN Jia-xu[1], YU Xin-ming[1], ZHANG Xi[2], LEI Fang-shu[2], ZHANG Peng[3], and ZHU Guang-yu[1]

[1]. MOE Key Laboratory for Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China;
[2]. Beijing Transport Institute，Beijing Key Laboratory of Urban Traffic Operation Simulation and Decision Support ,Beijing,100073
[3]. Transport planning and research institute, Ministry of transport,China, Beijing, 100028
{ZHU Guang-yu} gyzhu@bjtu.edu.cn com

Abstract. The quality control and evaluation of traffic detector data are a prerequisite for subsequent applications. Considering that the PCA method is not ideal when detecting fault information with time-varying and multi-scale features, an improved MSPCA model is proposed in this paper. In combination with wavelet packet energy analysis and principal component analysis, data fault identification for traffic detectors is realized. On the basis of traditional multi-scale principal component analysis, detailed information is obtained by wavelet packet multi-scale decomposition, and a principal component analysis model is established in different scale matrices; fault data is separated by wavelet packet energy difference; according to the time characteristics and space of the detector data Correlation fixes fault data. Through case analysis, the feasibility of the method was verified.

Keywords: Fault Data Recognition, Fault Data Repair, Wavelet Packet Energy Analysis, Principal Component Analysis.

1 Introduction

The development of information technology and the wide application of various traffic detectors provide a large amount of traffic data for the intelligent transportation system. However, traffic detector inherent defects, disrepair, communication failures and environmental impact and other factors can produce traffic flow fault data, and reduce the credibility of data, thus affecting the reliability of traffic system[1]. Therefore, it is of great significance to identify the traffic flow fault data and to repair it reasonably and improve the quality of traffic detection data.

At present, the research on traffic detector fault data identification is mainly divided into data fault recognition based on traffic flow three-parameter law, data fault recognition based on statistical analysis and data fault recognition based on artificial intelligence[2]. Xu Cheng[3] designed the method of data quality control by

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

analyzing the influence of sampling interval of detector and intrinsic law of three parameters of traffic flow. Xiao Q[4] uses wavelet analysis and least square method to study the detection of traffic flow anomaly data, and effectively reduces the misjudgment rate and false rate. Ngan H Y T[5] proposes a Dirichlet process hybrid model to identify the traffic detector abnormal data, which has good robustness. Wong C H M[6] and Dang T T[7] first identify potential outliers by clustering, and use principal component analysis (PCA) to transform ST (spatial temporal) signals into two-dimensional coordinate planes to reduce the size. Furthermore many scholars have proposed a multi-scale principal component analysis (MSPCA) model, which combines principal component analysis to remove correlation among variables, extract the decisive characteristics of wavelet analysis, and remove the advantages of measurement auto-correlation, and calculate the PCA model of wavelet coefficients at all scales. Traffic flow data restoration is one of the important measures to ensure the quality of data, and its research mainly focuses on time correlation, spatial correlation and historical correlation[8].

To summarize, principal component analysis is limited to the establishment of a fixed and single scale model. When detecting the time varying and the multi-scale characteristics of fault information, the method is not ideal. Therefore, we combine wavelet analysis with PCA, and use PCA to do multivariate statistical analysis of off-line data. Aiming at data fault recognition problem of traffic detector, a fault data identification and repair model based on improved multi-scale principal component analysis is proposed in this paper. First, the wavelet packet is used to decompose the original data with multi-scale, and the corresponding principal component analysis model is established. Then the real value of the fault data is estimated based on the temporal characteristics and spatial correlation.

2 Fault data recognition model based on improved MSPCA

In the actual traffic data monitoring process, the distribution of noise is random. Its intensity is also time-variable. However, when dealing with wavelet coefficients beyond the statistical control limit, MSPCA uses a uniform threshold at this scale to reconstruct the wavelet without considering the time variability of the noise, so part of the noise is mistakenly identified as a fault to be separated and partially covered by the noise of the fault will be expanded, leading to false alarm phenomenon.

At the same time, in order to solve the problems of MSPCA modeling fixed, principal component subspace and SPE(Squared prediction error), and single parameter, this paper draws on the idea of adaptive PCA's principal component recursion, and modifies the following three points for traditional MSPCA:

(1) Subsection processing of traffic flow data.

(2) The wavelet decomposition is changed to wavelet packet decomposition to improve the resolution of the model

(3) Detection of fault information by using wavelet packet energy difference method.

Figure 1 shows the flow chart of the improved method .

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

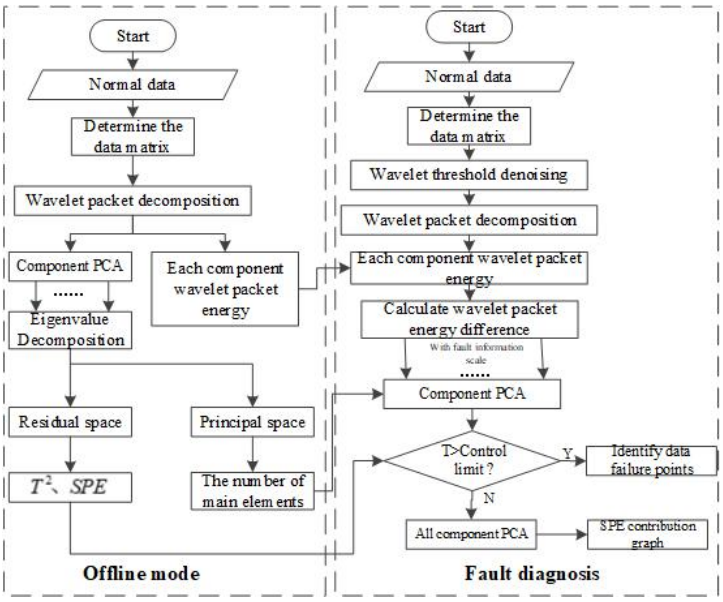


Fig. 1. The flow chart of the improved method .

Step 1: Sampling the detector to get the original data.

A sample data containing m sensors: $x \in R^m$, in the sample data, each sensor has n independent sample data, which is constructed into a data matrix X of $m \times n$ size, where each column of X represents a measurement variable, and each row represents a sample.

Step 2: By $S = \mathrm{cov}(x) \approx \dfrac{X^T X}{n-1}$ calculating the co-variance matrix of X

Step 3: Calculate the eigenvalues and eigenvectors of the related data matrix Λ .

The co-variance of the main element is Λ represents the larger eigenvalues(v) of the former A of the S.

Step 4: Calculate the main element $T = XP$.

Where $P \in R^{m \times A}$ is the load matrix, $T \in R^{n \times A}$ is the scoring matrix, and the columns of T are called principal variables and A is the number of principal components.

The principal component T represents the projection of the data matrix x in the direction of the load vector corresponding to this principal component. The larger its length, the greater the degree of coverage or variation of x in the p direction.

If $\|t_1\| > \|t_2\| > \ldots > \|t_m\|$, then P_1 represents the maximum direction of the data X change, and P_m represents the smallest direction of the data change.

Step 5: Calculate the principal component cumulative variance and contribution rate.

The cumulative variance contribution rate indicates that the amount of data that the former A principal can explain accounts for 6 of the total data.

According to CPV to calculate, normally when CPV reaches 85% or more, the previous principal can be assumed to explain most of the data changes.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

3 Fault data repair model

3.1 Data restoration based on time characteristics

The trend of traffic flow time series is positively correlated with the trend of historical time series, but when the traffic density is small and crowded, the interaction between vehicles becomes very small. At this time, traffic flow time series shows short-range correlation.

Using the ARIMA model to establish a model of the stationary sequence, and through the inverse change to the original sequence.

The general form of the ARIMA process can be expressed as follows: $\varphi(\mathrm{B})\,z_t = \phi(\mathrm{B})\nabla^d z_t = \theta(\mathrm{B})\,\mathrm{a}_t$.

z_t : original sequence

a_t : white noise sequence

$$\phi(\mathrm{B}) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \tag{1}$$

$$\theta(\mathrm{B}) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{2}$$

3.2 Data restoration based on space characteristics

In the traffic network, the traffic flow in the road section is affected by the upstream and downstream sections in the spatial characteristics, and the traffic flow sequences in the upstream and downstream sections show some correlations in time characteristics.

By statistics, correlation reflects the linear correlation between different sets of data, the greater the correlation can be linearly expressed with each other. Therefore, based on the adjacent sections of traffic flow data as an independent variable, using multiple linear regression model for traffic flow repair.

$$x_1(t) = f(x_1(t-1), \cdots; x_1(t-\tau_{\max}), x_2(t), \cdots; x_2(t-\tau_{\max}), \cdots; x_8(t), \cdots; x_8(t-\tau_{\max})) + e \tag{3}$$

$x_1(t)$ 、 $x_1(t-1)$ means the traffic flow parameter value at the moment t and t-1 when the detector No.1 is at the cross-section, and so on， $f(\cdot)$ is a function to be estimated， $\tau_{\max}$ is the maximum time lag value, e is relative error.

4 Case study

The data of the coil detector in the intersection of Chongqing is selected as the experimental data source during the week（2016-08-15 to 2016-08-20）.There is a coil in front of the stop line in each lane of the intersection. There are 8 coils in total. Each detector generates about 555 data a day. The data amount every day for 555×8. In coil No. 3, two faults occur between data points 300-400. Using the correlation of the data in time and space to repair fault data.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

4.1 Data fault recognition and analysis.

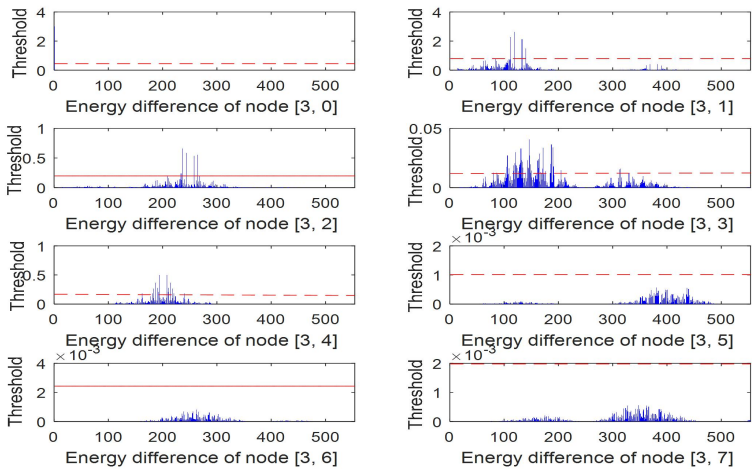
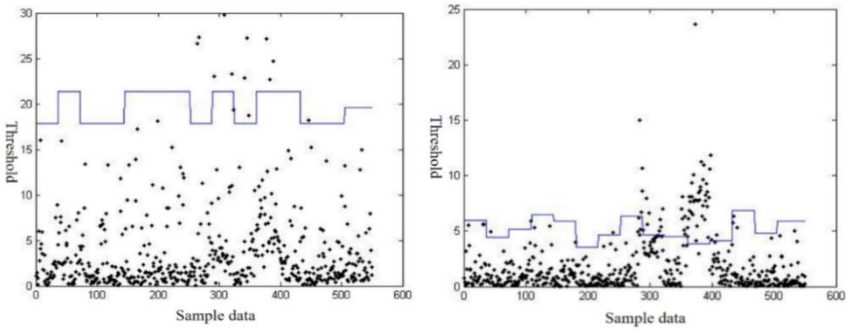


Fig. 2. Energy difference result of third layer decomposition

Using No.1 detector as a typical example, figure 2 shows the energy difference of eight vectors decomposed by layer 3 wavelet packet of the detector data. The dashed line represents the energy difference threshold of each component. It can be seen that the node [3,0] and [3,7] found fault information with different degree of data failure near 150-200.



(a) Improved MSPCA-T^2 (b) Improved MSPCA-SPE

Fig. 3. Control chart of improved MSPCA fault diagnosis model

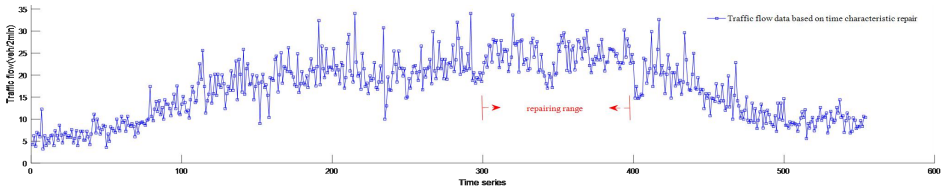
To validate the advantages of the proposed fault diagnosis data model, this paper will first signal to be detected in scale after wavelet packet decomposition for 3, every dimension are calculated respectively under the corresponding normal data with wavelet packet energy scale energy difference between the same below nodes. After discovering the node with obvious abnormality, that is, when the fault information is found, the node data matrix of the signal is modeled by MSPCA modeling and improved MSPCA.

Inspected after wavelet threshold in addition to the noise signal, the reconstructed signal dimension is 3 after wavelet packet decomposition, received

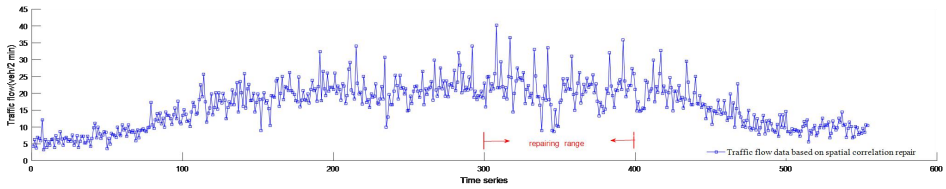
ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

the first scale 2 nodes energy, the second dimension for 4, the third dimension won eight node energy. Every dimension are calculated respectively under the corresponding normal data with wavelet packet energy scale with the node, the energy difference between the to find more apparent anomaly node, which found that after the location of fault information, the node data matrix of this signal PCA modeling, the result is shown in figure 3.

4.2 Data recovery



(a) Traffic flow data based on time characteristic repair



(b) Traffic flow data based on spatial correlation repair

Fig. 5. Traffic flow data of No.3 detector

For each model training, firstly, the original data sequence stability was verified by the ADF unit root, and the difference number d was determined and the time sequence was smooth and steady. Secondly, the model order number p and q are determined by AIC criterion, and the model parameters are estimated. Finally, the obtained model is used to repair and restore the difference. The fix results are shown in figure 5(a).

First, select the collect data of the fifth day to analyze. Calculate the rest of the detector flow data of correlation coefficients of data collected from the detector No.3, when the time lag values are 0, 1, 2, 3, respectively. Secondly, set the correlation coefficient threshold as (0.8,1). The sequence of conforming data is used as a preselected independent variable, corresponding detector No.3 is a dependent variable .Spatial estimation model of traffic flow by stepwise linear regression.

$$Q_{3\_t} = 3.017 + 0.276Q_{1\_(t-2)} + 0.511Q_{2\_(t-1)} + 0.203Q_{4\_(t-1)} \quad (4)$$

Among them, $Q_{1\_(t-2)}$, $Q_{2\_(t-1)}$ and $Q_{4\_(t-1)}$, respectively represents the traffic flow of detectors No.1, 2, 4 during t, t-1 and t-2 periods.

The traffic flow data collected by the detector 3 at 2016-08-08 is repaired by the formula (4), and the result is shown in figure 5(b).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52

5 Conclusion

In this paper, fault diagnosis and data restoration of traffic flow data are studied. Considering that the time-varying and multiscale features of PCA fault information are not ideal, this paper proposes an improved MSPCA model. Based on the traditional MSPCA model, using wavelet packet decomposition, and then wavelet packet energy difference method is used. Detect fault information and separate fault data to improve detection accuracy. Then use the repair model to calculate the true value according to the temporal and spatial correlation of the traffic flow data respectively.Case studies have found that the improved MSPCA fault data diagnosis model and data repair model can effectively identify abnormal data and repair it.

Acknowledgment

This work is supported by the National key research and development plan of Ministry of science and technology (2016YFB1200203-02, 2016YFC0802206-2), the National Science Foundation of China (Nos. 61572069,61503022), the Fundamental Research Funds for the Central Universities (Nos. 2017YJS308, 2017JBM301), Beijing Municipal Science and Technology Project (Z161100005116006, Z171100004417024); Shenzhen public traffic facilities construction projects(BYTD-KT-002-2).

References

1. Wang X Y, Zhang J L, Yang X Y.:Key theory and method of traffic flow data cleaning and state identification and optimization control [M]. Science Press, Beijing (2011).
2. Wen C L, Lv F Y, Bao Z J, etc.: A Review of Data Driven-based Incipient Fault Diagnosis [J]. Acta Automatica Sinica, 42(9):1285-1299.(2016).
3. Xu C, Qu Z W, Tao P F, etc.: Methods of real-time screening and reconstruction for dynamic traffic abnormal data[J]. Journal of Harbin Engineering University, 37(2):211-217(2016).
4. Xiao Q, Wang D J, Liu D.: Abnormal traffic flow data detection based on wavelet analysis[C].Matec-Conferences 01090(2016).
5. Ngan H Y T, Yung N H C, Yeh A G O.: Outlier detection in traffic data based on the Dirichlet process mixture model [J]. Intelligent Transport Systems Iet, 9(7):773-781(2015).
6. Wong C H M, Ngan H Y T, Yung N H C.: Modulo-k Clustering based Outlier Detection for Large-scale Traffic Data[C] Proc. Int'l Conf. IEEE Information Technology and Application. IEEE (2016).
7. Dang T T, Ngan H Y T, Liu W.: Distance-based k-nearest neighbors outlier detection method in large-scale traffic data[C] IEEE International Conference on Digital Signal Processing. IEEE, 507-510 (2015).
8. Lu H P, Qu W C, Sun Z Y.: Detection and repair algorithm of traffic erroneous data based on S-G filtering [J]. Civil Engineering Journal, (5):123-128(2015).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_52