International Journal of Science and Engineering Applications
Volume 14-Issue 12, 35 – 38, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1412.1007

A Study on Efficient Traffic Flow Data Cleaning Approaches

Yiqing Song
School of Automotive
Engineering
Zibo Polytechnic University
Zibo, Shandong, China

Abstract: Enough and accurate traffic flow data was essential guarantee to realize Intelligent Transportation Systems. Many quality problems were existed inevitably in detected data, including inefficacy, redundancy, error, missing, time dot excursion etc. On the basis of sufficient study and analysis for the influence reasons of data quality, the definition of data cleaning was proposed, and the cleaning rules and cleaning steps of “dirty data” were studied at the same time. Then the proposed cleaning rules were calibrated with the detected data of loop vehicle detector. It is pointed that, the recognition rates of “dirty data” is up to 90%. The results show that, “dirty data” can be effectively detected to help to increase the validity and veracity of the following data mining according to cleaning rules and cleaning steps.

Keywords: traffic flow; ITS; data quality; data cleaning; rules

1. INTRODUCTION

High-quality traffic flow data are the basic guarantee of right decisions for Intelligent Transportation Systems（namely ITS）. Loop detector is relatively inexpensive, which is just adapted to the situation of China. However, because of large detected data quantity and short detected period, the detected data are often accumulated to mass data in traffic control centers. The detected data, which can timely, roundly, and reliably indicate the traffic state, should be analyzed fast, effectively and deeply to provide a basis for traffic control, traffic management, transport planning etc. by making the mass dynamic traffic data into the main body of ITS data. As the loop detector is out of work, or the faults caused by the detectors and transmission equipment etc. happened, various quality problem are existed inevitably in the detected data. There are a number of quality problems in the data collection inevitably, as well as the search process of data mining would be misled by problem data. Therefore, it is necessary to carry out data cleaning in order to improve the quality of the data set. And then the validity and veracity of followed data mining are improved.[1-9]. In view of this, the data pre-processing is studied both at home and abroad. Traffic problem data are divided into three types: missing data, distortion data and abnormal data, and correspondingly, the cause, recognization and modification of problem data are proposed in literature [10].The method of judging abnormal data of expressway traffic flow is made in the literature [11]. The ITS data quality control algorithms has been put forward in literature [12]. A theory of traffic flow data quality testing has been expressed in the literature [13]. And cleaning rules of section traffic data are proposed in the literature [14], which is suitable for online applications. However, only a few data quality issues are studied in the above-mentioned documents, also various quality problem are often existed in the detected data simultaneously, and there is no complete set of rules for cleaning data quality management. Combined with above-mentioned research results, the definition and rules of data cleaning in traffic flow are studied, in addition, the validity of data cleaning rules is validated in this paper.

2. CONCEPT of DATA CLEANING

From character of problem data, it can be divided into two types: normal data quality problem and abnormal one. Quality problems of normal data, including noise data, are inevitable quality problem, which are caused by large, puny and uncontrollable random factors, or accidental factors (short-time traffic flow fluctuation etc.) in detection process. Quality problems of abnormal data, including invalid, redundancy, errors, time-point shift, missing, and so on, are easy to recognize and reject, which are caused by small but remarkable, controllable systematic factors in detection process. If the problem data are not modified or smoothed, these are used in data mining directly, the added quality problem will be caused in the following application. To improve the data quality, the definition of cleaning rules in the field of traffic flow is studied as followed.

Due to environmental factors, equipment failures, communication failures, etc., the phenomenon, including invalid, redundancy, errors, time-point shift, missing, etc. is existed in detected data, which are called “dirty data”. In order to avoid such “dirty data” estimating, forecasting or evaluating the traffic status directly and becoming a bottle-neck in following models simultaneity, dirty data are needed to be eliminated, such as eliminating noise, modifying incorrect data, reducing redundant data, filling missing data and so on. Thereby, data quality would be enhanced for ITS, this course is known as data cleaning. That is, dirty data are modified or rejected by series of algorithms from mass raw data.

3. GENERAL STEPS OF DATA CLEANING

The process of data cleaning is divided into 5 phases: analysis data attributes, determined cleaning rules, calibration cleaning methods, execution cleaning components and data updating. Cleaning rules is the key of cleaning process of “dirty data”.

Whereas the incorrect, redundant or loss data are a frequent occurrence, the cleaning rules about above-mentioned types of dirty data are given. The smoothing method of noise data has been widely used, including fixed time average method, moving average method, exponential smoothing method, Kalman filter method etc., which would not be introduced in detail in this paper.

3.1 Attribute analysis of“dirty data”

Error data. When the traffic detectors are out of work, the detected data are usually wrong. These data are not expected or not satisfied within the framework of existing rules and principles. For example, when the flux of traffic flow is less, but the higher lane- occupancy, the data is obviously wrong.

Redundant data. For a single detector, the similar duplication data which are collected are defined redundant ones; for multi-detector, because of too much density in the same road or adjacent sections, the detected data are impacted directly with vehicles overlapping and redundancy, so the redundant data are defined.

Loss data. As the scanning frequency of detector is not fastness, transmission or storage of equipment is failure, the operation is error, and the detector can not be detected the correct vehicles because an over-density of vehicle, and so on , the dynamic traffic data can not be strictly uploaded with the specified interval time, then the data is lost.

3.2 Cleaning rules

In order to clean the ITS data, the incorrect, redundant, and missing data should be firstly identified.

As the error data are often expressed in outlier, therefore, the purpose of clean data can be reached by the detection and removal of outlier in source point. Then the quality of the data in data sources can be enhanced. In the area of traffic flow, the tested data is high-dimensional data, which have a number of attributes. Taking the occupation, speed ,traffic flow for example, if a traditional outlier detection algorithm is used, for the same goal of multi-attribute data sets, each attribute can be only detected one by one, then the time complexity is increased, as well as the inter-related of three attributes is separated. So the algorithm based on similar coefficient sum is proposed in this paper [16, 17].

The loss and redundant data[9, 10] can be distinguished by the following rule. The rule is noted to Rule 3: The data of a certain period is defined as a period of data, then the time of data is scanned and judged. If they are not got in a certain period, the data are considered as loss; if they are got more than one set, the data are considered as redundancy.

3.3 Cleaning steps of "dirty data"

The cleaning of the raw data which are obtained on the detector can be divided into two major steps, that is, the raw data are distinguished in accordance the above-mentioned rules, and problem data are modified according to appropriate algorithm.

4. APPLICATION

4.1Data sources

In order to achieve the quality management of the wrong data, this article is focused on verifying the validity of the cleaning rules. The data are collected at a section of highway in the area Zhangdian of city Zibo in Shandong Province, the road capacity is designed for the 1000, the ring coil detector is detected. The data include 3 parameters, which are traffic flow, speed, occupation, but they are not pointed by vehicle Models (large, medium and small). Acquisition time is from the October 15, 2007 to October 19, 2007 during 00:00:00-24:00:00, the collecting interval is 5min. The data on October 15, 2007 to October 18, 2007 are considered as training samples, the data of Oct. 19, 2007 are considered as test samples. At the same time, the rules ere checked by other data sets, this article is no longer listed for the length.

4.2 Model application

The detected data are cleaned in terms of cleaning steps, which is shown in Figure 1. $n=4, m=3$ and $\lambda=5\%$ are determined according to the data of the training sample. The ratios of the correct data and "dirty data" in tested samples are 95.83% and 4.17%. The speed-time charts of the fore-and-aft cleaning data are shown in Fig.1 (a) and Fig.1 (b); the flow - time charts of the fore-and-aft cleaning data are shown in Fig.2 (a) and Fig.2 (b); the share - time chart[18-19] of the fore-and-aft cleaning data are shown in Fig.3 (a) and Fig3 (b), in the figures , the error data are shown by the ellipse, the loss data are shown by the rectangle, the redundant data are shown by the hexagon.

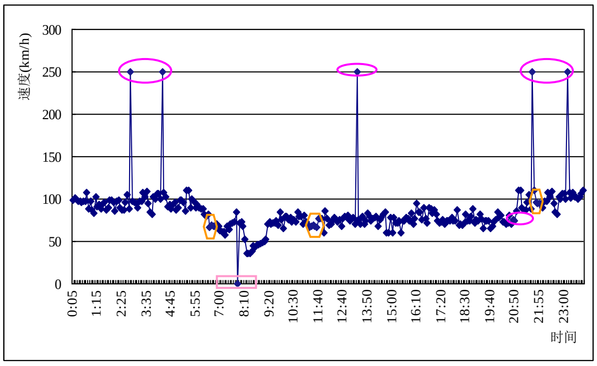

Fig.1 (a) the speed - time original data graph

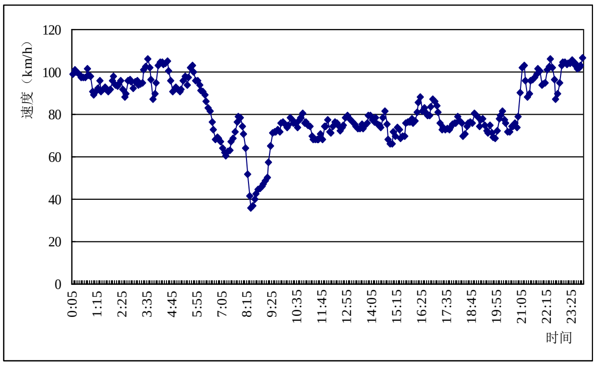

Fig.1 (b) after cleaning data of speed-time graph

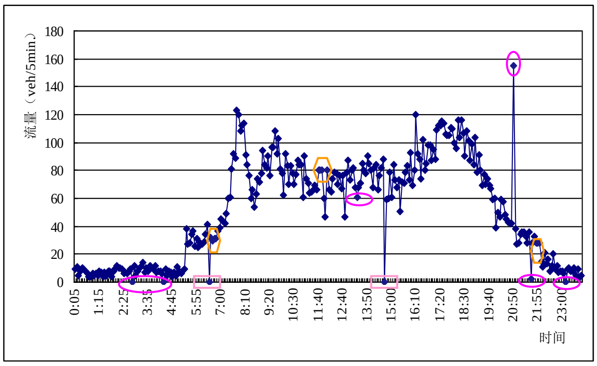

Fig.2 (a) the traffic flow – time original data graph

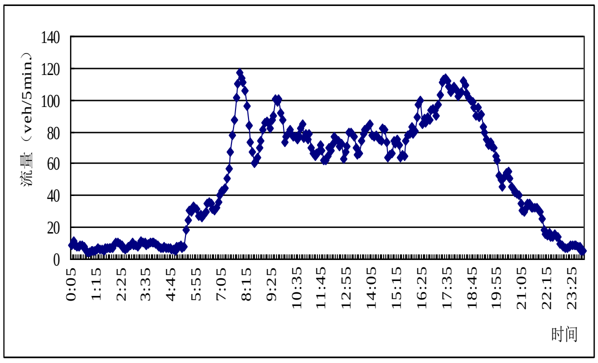

Fig.2 (b) after cleaning data of traffic flow-time graph

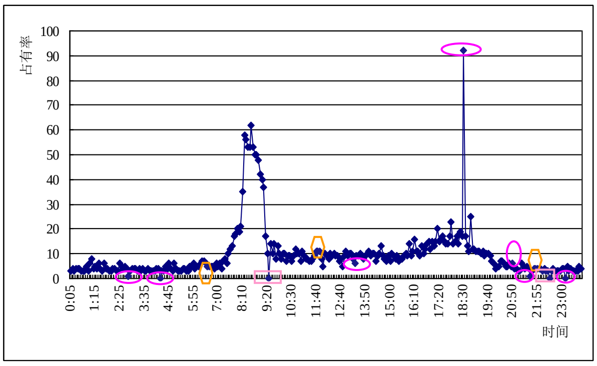

Fig.3(a) the share–time original data graph

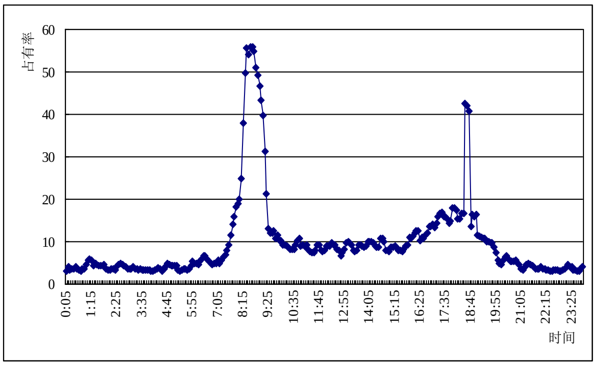

Fig.3(b) after cleaning data of share-time graph

4.3 Analysis

According to the cleaning rules, the recognition rates of three “dirty data” are shown in Tab.1.

Tab.1 "Dirty data" in recognition rate

	error data	loss data	redundant data
recognition rate	92.1%	90.3%	91.7%

Fig.1 (a), Fig.2 (a) and Fig.3 (a) are noted that wrong, loss and redundancy data are existed in the original traffic flow data of the different elements. It can be seen from Table1 that the identification rate of the error, loss, and redundant data is more than 90%.And dirty data can be detected by using cleaning rules proposed in this paper. From Fig.1 (b), Fig.2 (b) and Fig.3 (b), it can be seen that, error data are corrected, loss data are filled, redundant data are reduced after cleaning out the raw data, the macro characteristics of data is displayed clearly.

5. CONCLUSIONS

The error, loss, redundant data are cleaned effectively in this paper, and the quality of data is improved significantly, thus the effectiveness and accuracy of data mining are enhanced. However, the three mentioned "dirty data" can be only carried out by the cleaning rules, the time-drift data can not be cleaned, so there are still some limitations in the rules.

Error data can be regarded as the isolated points, but all outliers are not incorrect data. Whether the outlier is incorrect data or not is judged combined with threshold theory and the traffic flow theory. However, an algorithm for outlier detection based on similar coefficient sum is proposed in this paper, the time complexity is bigger, so the algorithm with smaller time complexity is need to be studied. With difference of the road in the level, the nature, control and the type of traffic-related parameters, the requirement of threshold is different, so threshold selection will be needed for judging whether the outlier is error data or not. It is best that threshold is determined based on the statistical distribution of each traffic parameters.

6. REFERENCES

[1] Daniel L.Gerlough, Matthew J.Huber. Traffic flow theory [M]. Beijing: China Communications Press, 1983.

[2] Yang Zhaosheng. Technology and application for basic traffic information [M].Beijing: China railway publishing house, 2005.

[3] WANG Xiao yuan ; JUAN Zhi cai ; JIA Hong fei ; PIAO Ji nan. Study of a statistical method of change-point to analyze traffic flow breakdown [J]. , China Journal of Highway and Transport, 2002,15(4):69-74.

[4] Zhang Jinglei ,Wang Xiaoyuan. Research .Progress of Traffic Incident Automatic Detection Algorithms[J]. Research Progress of Traffic Incident Automatic Detection Algorithms, 2005, 29(2):215-218.

[5] WANG Xiao yuan; JUAN Zhi cai; JIA Hong fei. Micro-simulation models of traffic flow of developing and evaluating ITS[J]. Journal of Traffic and Transportation Engineering, 2002,2(1):64-66.

[6] WANG Xiao yuan; JUAN Zhi cai; Piao Ji nan; JIA Hong fei. A Statistical Theory of Change point with Local Comparison and its Application in Studing Traffic Flow Breakdown[J]. Journal of Highway and Transportation Reseach andk Development, 2002,19(6):112-115.

[7] WANG Xiao-yuan ZHANG Jing-lei ZHANG Kai-wang WU Lei2. Study on Traffic Flow Forecasting Method Based on Non-parameter Regression Spline Fitting[J]. Computer Engineering and Applications, 2006(26):218-220.

[8] WANG Xiao-yuan; LIU Hai-hong. Short-time Traffic Flow Forecasting Based on Projection Pursuit Auto Regression[J].Systems Engineering, 2006,24(3):20-24.

[9] Jiang Guiyan.Technologies and applications of the identification of road traffic conditions [M].Beijing: China Communications Press,2004,103-113.

[10] JIANG Gui-yan; GANG Long-hui; ZHANG Xiao-dong; WANG Jiang-feng. Malfunction identifying and modifying of dynamic traffic data[J]. Journal of Traffic and Transportation Engineering, 2004,4(1)121-125.

[11] CHEN De-wang;. ZHENG Chang-qing, ZHANG Chang-biao . An Algorithm for Judging Abnormal Data of Expressway Traffic Flow and Its Validation[J]. China Safety Science Journal (CSSJ), 2006,16(7):122-127.

[12] GENG Yan-bin, YU Lei ,ZHAO Hui . ITS Data Quality Control Techniques and Applications[J]. China Safety Science Journal,2005,15 (1):82-87.

[13] JIANG Rui, WANG Jun. Method to Verify and Repair Road Traffic Flow Data[J]. Computer and Communications, 2006,24(6):65-67.

[14] QIN Ling, GUO Yan-mei, WU Peng, Cross-section of data traffic detection and pre-test the key technology research[J]. Journal of Highway and Transportation Research and Development, 2006,11,39-41.

[15] GUO Zhi-mao; ZHOU Ao-ying. Research on Data Quality and Data Cleaning: a Survey[J]. Journal of Software, 2002,13(11):2076-2082.

[16] CHEN Wei; WANG Hao; ZHU Wen-ming. Cleaning Method of Incorrectness Data Based on Outlier Detection[J]. Application Research of Computers, 2005(11):71-73.

[17] JIANG Lingmin. Clustering Algorithm to Check Outlier Based on Similar Coefficient Sum[J]. Computer Engineering, 2003,29(11):183-185.

[18] ZHANG Feng-rong,WANG Li-li,Quality management and control[M].Beijing: China Machine Press,2006.

[19] ZHANG Gong-xu. Management and control the quality of the election map[M].Beijing: Posts, 1983.