Monitoring Data Cleaning of Urban Tunnels by Fusing PCA and CLARA Algorithms

1, 2 Kun Hao Tang, 1 Luo Zhong, 1 Lin Li, 1 Guang Yang
1 Department of Computer Science and Technology, Wuhan University of Technology, 4300702
2 Department of Computer and Information Science, Hunan Institute of Technology School, 421002
Tel.: 13975466676
E-mail: 13587263@qq.com

Received: 13 September 2013   /Accepted: 25 October 2013   /Published: 30 November 2013

Abstract: With the development of urban construction, urban traffic is becoming more and more complex. And the status of tunnel is more important than before. As time goes on, urban road tunnel monitoring system will accumulate a large amount of data. To mining the knowledge from those data for urban management, the processing of data is necessary. This paper put forwards PCCA algorithm, which combines PCA and CLARA algorithms for a set of real-time environmental data collected by an intelligent tunnel monitoring system. The proposed algorithm is able to reduce data dimensions while analyzing data effectively. In our experimental results, we show that our proposed algorithm can improve the accuracy of traditional data cleaning algorithms and reduce their running times. In addition, the abnormal data is listed by their reasons.

Keywords: PCA, CLARA dimension reduction, Processing of data.

1. Introduction

Benefit from the ministry of public security policy, such as ‘road traffic safety ‘twelfth five-year’ plan” and “road traffic technology development ‘twelfth five-year’ plan”, 182 billion will be invested in intelligent transportation [1]. So far there are 19 cities announce their intelligent transportation investment projects, involving 7.805 billion. Thus the urban tunnel becomes more and more important [2]. In view of the characteristics such as the narrow space, the insufficient light, ventilation, the bad air quality and so on, the tunnel monitoring system is particularly important in the process of tunnel management.

Tunnel environment data not only reflect the indexes, including visibility, wild speed, temperature, and humidity of tunnel, but also can be used to analyze the abnormal situation of tunnel, including traffic accidents and equipment accidents. And the tunnel environment data is big data with high dimension, how to process data repetitiously and analyze thoroughly from the data ocean for enabling us to gain knowledge effectively and dig out the key information needed to successfully, becomes a research focus in data mining at the current time. In the whole process of data mining, data preprocessing time accounted for about Sixty percent, and the final data mining work only takes up about ten percent of all time. The related data is so big that it must be achieved through the program and algorithm.

The purpose of data preprocessing is using relevant technology to extract data from various data sources so that find out the dirty data or clean up it. Common technical means include Binning smooth, regression and clustering.
Literature [3] proposes CLARA algorithm, a clustering algorithm for dealing with large data, based on dividing. This paper proposes solve the problem of large data by sampling. And it tells that the theory of CLARA is using samples to instead of the overall data set and getting best clustering result by random sampling many times for the purpose of deal with data by clustering.

Literature [4] proposes to reduce dimensions with PCA so that cut down the amount of information for the purpose of analyze data conveniently. It shows that the PCA algorithm can perform a covariance matrix, correlation matrix or any weighted covariance matrix. And it illustrates how to eliminate the intensity dependence and standardized dimensionless variables by principal component analysis so that we can preprocess data by using it.

Literature [5] analyzes the method of detecting isolated point in the instance. In addition, it introduces the related concept, the method of classification and the way of evaluation in detail. At the same time the system designs two algorithms. The one is based on pruning strategy while the other one supports vector machine and improves it. Finally the algorithm turns out to be suitable for dealing with the low dimensional data for its accuracy and high efficiency. The idea of this literature can be used to pretreat the tunnel data.


In the view of analyzing the data of one tunnel in Wuhan, which have complex formats, inconsistent norms and unified data parameters. Besides, the internal environment is terrible, the equipments are easy to make mistakes and damage so that abnormal data come into being. Literature [3] deals with the large data and literature [4] analysis the method to reduce the dimensions of the data having many dimensions. Furthermore, literature [5] solve the dirty data by analyze the isolated point. In a word, literature can solve a problem effectively while it can’t deal with the complex situation including many different problems. In order to solve the problem of this situation, the new algorithm PCCA is put forward.

2. Urban Tunnel Monitoring Data Pretreatment and Solution

The urban tunnel monitoring data can obtained a variety of variables by PLC devices [10], in general, it can be divided into vehicle inspection data and environmental data as the following Table 1.

<table>
<thead>
<tr>
<th>Data sources</th>
<th>Data type</th>
<th>Pretreatment method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tunnel car inspection data</td>
<td>Time</td>
<td>Clustering and SOM</td>
</tr>
<tr>
<td>Wild speed</td>
<td>Traffic</td>
<td></td>
</tr>
<tr>
<td>Driving speed</td>
<td>Driveway occupancy rate</td>
<td></td>
</tr>
<tr>
<td>Tunnel car environmental data</td>
<td>Time</td>
<td>PCA or improved algorithm</td>
</tr>
<tr>
<td>Wild speed</td>
<td>Wild speed</td>
<td></td>
</tr>
<tr>
<td>Temperate</td>
<td>Humidity</td>
<td></td>
</tr>
<tr>
<td>Humidity</td>
<td>Light intensity</td>
<td></td>
</tr>
<tr>
<td>CO concentration</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Due to the high demand for real-time, the collected data storage density is large. And the environmental data is so large that data mining must be after the data cleaning. Besides, the dimension has different reference value and has diverse effect in mining data. So, we should pretreat data in advance. This paper collect and analysis data such as CO, temperature, wind speed, light intensity and so on. In order to analysis data later, data are divided into six modules as the following Fig. 1, Fig. 2.

**Fig. 1.** Ideal state figure of the divided of tunnel environment variables.

**Fig. 2.** Actual state figure of the tunnel environment variables.
Fig. 1, called as ideal state figure, means dividing the variables into six basic variables which divide collected information equally. And the six variables have the same effect. On the other hand, Fig. 2, called as actual figure, means the data change according to the number of equipments. The CO concentration and temperature changes remarkably so that other type data change along with them. Finally, the Fig. 2 is proved to be more in line with the simulation results.

3. Algorithm Description

In recent years, the status of data mining is improved. Moreover, data is the basis of data mining. In order to input data as the type we needed, we should preprocess the data. And there are many approaches to preprocess the data. For the purpose of reducing the dimension, we should preprocess with the PCA algorithm, then we will cluster the result and find out the abnormal data in order to clean up the data.

3.1. Principal Component Analysis Description

Principal component analysis is also known as principal component analysis. It is a multivariate statistical method that reduces many variables to a few principal components by reducing dimension. The principal component reflects most information of original variables. They usually expressed as a linear combination of the original variables. In order to make the information contained in these principal component non-overlapping, it demand that the variables are unrelated. In general, when the condition involves a lot of variables and relationship between variables are obvious, the PCA is suited. It is easy to grasp the main aspects of things so as to simplify the problems. The steps are as follows:

Step 1: Draw random samples from overall data, name the observation value matrix as $X$. The each row of $X$ corresponds to a sample of record and each column corresponds to a variable.

Step 2: Calculate the sample covariance matrix $S$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

Step 3: Solve main component according to the sample covariance matrix $S$. The eigenvalues are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, and the eigenvectors are $t_1, t_2, \ldots, t_i$. The principal components are $y_i = t_i^T$. 

Step 4: Calculate the contribution of each principal component $a_i$

$$a_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Step 5: Calculate the scores of principal component $F_i$

$$F_i = a_1x_1 + a_2x_2 + \ldots + a_px_p$$

3.2. CLARA Clustering Algorithm Description

In the clustering algorithms base on partition, k-mean algorithm is widely used. It is an algorithm used to cluster according to the distances between the point in the data set and the clustering centers. However it can't suit for the large data situation. In order to deal with large data, Kaufman and Rousseeuw put forward the CLARA algorithm. CLARA did not find out representative objects from the overall data, but extract a part of samples and choose the representative point with the method of replacement. If the samples are selected in a random manner, the sample can represent the whole data set well. In order to get the better result, CLARA algorithm dram samples several times and chooses the best result as the cluster. Specific steps are as follows.

Step 1: Make sure the cluster number $k$ according to the concentration distribution of data. Draw random samples from whole data set and get the sample of $40 + 2k$ objects.

Step 2: Find out the sample center by using the method of PAM and calculate average dissimilarity degree as the threshold.

Step 3: Find out the nearest center for each object in the overall data set. Then classify the object by the center and write down the cluster number.

Step 4: Calculate the average dissimilarity degree of each cluster. If one is less than the minimum, the minimum should be replaced by it. And take the corresponding center as the best cluster center.

Step 5: Repeat step 1 to 4 several times so that ensure that there the result less than the threshold. And choose the experimental result which has the least dissimilarity degree, according to four decimal places.

3.3. Principal Clustering Component Analysis (PCCA) Algorithm Description

Because of the situations that there are high dimension and the data have obvious difference in the different dimension, PCA is needed at first for reducing the dimensions. So it can carry the point of getting reasonable data. Then we can perform the CLARA cluster to find out the dirty data. Moreover, we can find the reason according to the distance’s judgment. The specific steps are as follow:

Step 1: Pretreat the whole data set and remove the defect data.
Step 2: Use the principal component analysis in the whole data set and calculate the contribution in order to reduce the dimension.

Step 3: Cluster the data by the CLARA algorithm after reducing the dimension.

Step 4: Set the maximum and minimum values for all the dimension of record that is, set the limit.

Step 5: Find out the maximum value of the distance between the various data in each dimension for each cluster center.

\[ d_{i,j} = \max \{|a_{i,j} - m_j|, |a_{i,j} - n_j|\} \]  

Step 6: According to the most value of the distance between each dimension data and edge, and the weight of each dimension to estimate reasonable range of the dissimilarity between the normal point and the cluster center, called distance (i):

\[ \text{distance}(i) = \sqrt{\sum_{j=1}^{m} (d_{i,j}^2 * Q_j)} \]  

Step 7: According to the reasonable range distance (i), find out the dirty data (the corresponding dissimilarity is beyond the range).

Step 8: Test the dirty data and judge the fault type according to the number of dimension which is beyond the limit (if the number is one, it means equipment damage, or it stands for accident).

4. Experimental Result

4.1. The Collection of Experimental Data

This study uses MATLAB as the development tool, performing on 65535 data record with the input data provided in the form of EXCEL file.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Data type</th>
<th>Length</th>
<th>Null or not</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Int</td>
<td>4</td>
<td>Not</td>
</tr>
<tr>
<td>Time</td>
<td>Char</td>
<td>20</td>
<td>Not</td>
</tr>
<tr>
<td>CO</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>V1</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>FS</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>GQ</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>SD</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>WD</td>
<td>Varchar</td>
<td>50</td>
<td>Yes</td>
</tr>
</tbody>
</table>

4.2. The Experimental Evaluation

4.2.1. The Average Dissimilarity

This paper measure the average dissimilarity between the various data basing on Euclidean distance measurement. The meaning of average dissimilarity is the average value of the Euclidean distance between the data in the cluster and the corresponding cluster center. The smaller the average dissimilarity is, the more ideal the experimental result will be.

The calculation formula of Euclidean distance is as follows:

\[ d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \]  

the \( m \) is dimension of data

4.2.2. Silhouette

Silhouette is a kind of used for evaluation of clustering validity evaluation index. The higher silhouette index shows the better clustering results and it also can reflect the effectiveness of the clustering results in the form of contour diagram clearly.

For an arbitrary division \( X = X_1 X_2 \ldots X_k \), global Silhouette value is defined as:

\[ GSu = \frac{1}{k} \sum_{j=1}^{k} S_j, \]  

Where \( k \) is the number of category. And \( S_j \) is the Silhouette value of cluster \( j \), \( S_j \) is defined as:

\[ S_j = \frac{1}{m} \sum_{i=1}^{m} S(i), \]  

\( m \) is the sample size of cluster \( j \). And \( S(i) \) is a quality measure in the sample \( i \) of the cluster \( j \). \( S(i) \) is defined as:

\[ \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}, \]  

\( a(i) \) is the average distance between the sample \( i \) and other sample in the same cluster. And \( b(i) \) is the minimum value of the distance between sample \( i \) and all the samples included in the Cluster \( X_t (t = 1, 2, \ldots, k; t \neq j) \).

It is proved that \( s(i) [-1,1] \). When the value of \( s(i) \) is close to 1, the sample is clustered well. If the value of \( s(i) \) is close to 0, it indicates that the sample can be divided into the nearest the class.

If the value of \( s (i) \) is close to -1, the samples were wrong divided.

4.2.3. Contribution Rate

The proportion of \( y_i \), which is the \( i \) of main component, in the total variance is called as the contribution rate \( a_i \).
\[ a_i = \lambda_i \sum_{j=1}^{p} \lambda_j, \text{ the } i = 1, 2, ..., p. \]

The contribution rates of the principal component reflect the primitive variable information comprehensive ability of it. And the ability can also be understood as the power to explain primitive variable. We call the sum of the contribution rate from the component ahead as the cumulative contribution rate. It stands for the primitive variable formation comprehensive ability of them. According to the contribution rates, the Principal Component Analysis algorithm can be used to reduce the dimensionality.

4.3. The Experimental Results and Analysis

4.3.1. The Outline Figure

Each point represents a record, plotted by Silhouette on the horizontal axis and cluster on the vertical. The silhouette index of the algorithm PCCA is better than CLARA a lot and it already very close to 1. Besides, it means that the clustering result is very good. Moreover, the negative situation turns down a lot. It represents the probability of misclassification become lower (Figs. 3-5).

4.3.2. Contribution Rate

Contribution rate reflect the original variable information comprehensive ability. According to the Table 3, the accumulating contribution rate from the first five comes to 91%. It shows that the first five can reflect the original information well. So that we can reduce the dimensions in accordance with contribution rate, meanwhile, reduce the amount of information. Furthermore, it can clean up date effectively.

4.3.3. Average Dissimilarity Histogram

The horizontal axis of histogram (Figs. 6-8) shows the clustering number while the vertical axis shows the average dissimilarity. Comparing the algorithms, it turns out that the PCCA algorithm has the best cluster result. The result, average dissimilarity of each cluster is no more than 5, tells its quantity of no obvious abnormal point and the distances between point and its corresponding cluster center. And we can also find that all the average dissimilarity is small.

4.3.4. Dirty Data Cleaning

For the 65535 records from the urban tunnel monitoring environment data, three reasonable clusters can be got by using algorithm PCCA united PCA and CLARA. According to the different distances between point in the cluster and the cluster center, we can deal with the clustering result, find out the dirty data and analysis the reason of it.

The experimental result is in the Table 4.

5. Conclusions

The experimental results show that use the PCCA algorithm to deal with the data in different dimensions for the environmental data which has large amount of data and multifarious data in different dimensions. Moreover, it can make a difference in clustering and cleaning up dirty data. Besides, we can mine data and get the potential relationship of data while finding the abnormal data efficiently, so that we can analysis the accident in tunnel conveniently.
Table 3. Contribution rate.

<table>
<thead>
<tr>
<th>Eigen values</th>
<th>Difference</th>
<th>Contribution rate</th>
<th>Accumulating contribution rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.0326</td>
<td>7.4229</td>
<td>44.2136</td>
<td>44.2136</td>
</tr>
<tr>
<td>7.6097</td>
<td>3.084</td>
<td>22.3815</td>
<td>66.5951</td>
</tr>
<tr>
<td>4.3013</td>
<td>1.9238</td>
<td>12.6508</td>
<td>79.2460</td>
</tr>
<tr>
<td>2.3775</td>
<td>0.9414</td>
<td>6.9926</td>
<td>86.2386</td>
</tr>
<tr>
<td>1.8861</td>
<td>0.9081</td>
<td>5.5474</td>
<td>91.7859</td>
</tr>
<tr>
<td>0.9781</td>
<td>0.4365</td>
<td>2.8767</td>
<td>94.6626</td>
</tr>
<tr>
<td>0.5416</td>
<td>0.2096</td>
<td>1.5929</td>
<td>96.2555</td>
</tr>
<tr>
<td>0.3320</td>
<td>0.0549</td>
<td>0.9764</td>
<td>97.2319</td>
</tr>
</tbody>
</table>

Fig. 6. K-mean histogram.
Fig. 7. CLARA histogram.
Fig. 8. PCCA histogram.

Table 4. Deal with dirty data.

<table>
<thead>
<tr>
<th>Number</th>
<th>Cluster number</th>
<th>Dissimilarity</th>
<th>Reasonable range</th>
<th>Wrong dimension</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>19.7188</td>
<td>4.30</td>
<td>1 to 3</td>
<td>Equipment damage</td>
</tr>
<tr>
<td>2411</td>
<td>2</td>
<td>4.4156</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>2412</td>
<td>2</td>
<td>4.4296</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>2413</td>
<td>2</td>
<td>6.9926</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>8112</td>
<td>3</td>
<td>5.5474</td>
<td>4.30</td>
<td>1 to 3</td>
<td>Equipment damage</td>
</tr>
<tr>
<td>42445</td>
<td>1</td>
<td>2.8767</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>42446</td>
<td>1</td>
<td>1.5929</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>42447</td>
<td>1</td>
<td>0.9764</td>
<td>4.30</td>
<td>More than 3</td>
<td>Accident</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

Acknowledgements

Wuhan technology department, innovation project (2013070204005); Hengyang technology department (2012KG76); Hunan education department (11C0366); Hunan technology department (2013GK3037).

References

[5]. Run Hua Miao, Data preprocessing method based on clustering and outlier detection, Beijing Traffic University, 2012.

2013 Copyright ©, International Frequency Sensor Association (IFSA). All rights reserved.
(http://www.sensorsportal.com)