Proceedings of the 5th WSEAS Int. Conf. on Instrumentation, Measurement, Circuits and Systems, Hangzhou, China, April 16-18, 2006 (pp29-33)

Systematic analysis of OCS testing data

CHEN TANGLONG XIAO JIAN
School of Electrical Engineering
Southwest Jiaotong University
Chengdu 610031
CHINA

Abstract: - The testing parameters of OCS (Overhead Contact System) are keen to evaluate the safety of electrified railway transportation and the quality of current collection between pantograph and catenary. Since the test of parameters depends on the velocity of train and the OCS conditions, carrying on systematic analysis for testing data is of great significance. In this paper, the testing data are preprocessed through centering, de-dimension and standardization first; then the classification of the testing data has been made by clustering algorithm in terms of spatial location. Thus, we could estimate the parameters appropriately and dealing with these parameters respectively according to their characteristics. The feasibility of the approach proposed is verified by the simulation results.

Key-Words: - Overhead catenary system; Overhead catenary system detection; Catenary testing car; Railway electrification; Clustering; Linear regression

1 Introduction

The OCS is a large investment and high quality required system. It is essential for carrying out OCS on-line testing to guarantee the safe running and the quality of current collection.

The testing of OCS is to obtain the geometric and kinetic parameters of the contact line such as hard spot, pull-off value, contact pressure and velocity of train etc., through the sensors fixed up under the pantograph when train is running. The hard spot is an important parameter used to reflect the dynamic impact of pantograph and catenary. The bigger the value of hard spot is, the less smooth of the contact between the pantograph and catenary is, and the bigger offline of the dynamic contact between the catenary and pantograph is[3]. This would leads to the unstableness of the current collection of the locomotive [2]. Thus, it will affect the traction speed of the train, and increase the abrasion of pantogragh's slippery strip and contact line, which, in consequence, will reduce their service life.

The testing value sampled by vehicular testing devices differs when train run at different speed. The question of how to determine the threshold value of each parameter at an given speed is still open. All over the world, there is a lack of a technical criterion for the power supply and maintenance departments in railway companies [1].

In this paper, the OCS data, which were sampled from the section from Beijng-Guangzhou Railway Line in Dec.2004, are used to carry out analytical research by data-mining technique. Our goal is to determine the implied relationship among these parameters, so as to provide an approach to set up the mathematical model for the data processing. By this way, it could normalize the parameter values sampled at different speed to the corresponding values at specified speed of electrified section so one can evaluate the physical characteristics of the hard spot comprehensively, and is helpful for power supply and maintenance departments to adopt the most effective maintenance plan.

Furthermore, through the mathematical model established in this paper, the estimated value of these parameters when train running at high speed can be estimated from the parameters measured by OCS test equipment, when the train running at low speed. This can supply an important theoretical basis for the reconstruction of low speed railway to meet the needs of high speed train, which is a heavy investment in China now. Besides, it can be used to evaluate the OCS condition so as to determine the highest speed allowable in a specified section of railway; under which, the quality of current collection can be guaranteed.

2 Data preparation

There are hundreds of properties associated with mass measured data, most of them, however, are redundant, since a large part of them have nothing to do with the mining mission [5]. If analyzing these complex data directly, it will not only take a lot of time but also affect the accuracy of data-mining.

Therefore, it's necessary to eliminate noise, flaw and duplicate record of original data and to carry out de-dimension, centering and standardization processing in the stage of data preparation, under the premise of keeping integrity of the original data.

The centering processing of data is to carry out phase transition transform, defined as:

$$x_{ij}^{*} = x_{ij} - \overline{x}_j \tag{1}$$

Where $i = 1,2,...,n$, $j = 1,2,...,p$, x_{ij}^{*} is the transformed coordinate and $\overline{x}_j$ is the center-point (mean) of the column vector. This transform could make the origin of new coordinate system coincide with the center of gravity of the sample data set. Neither the mutual position of every sample points nor the correlation of each variables will be changed by the above transform.

After centering, the variance of the data can be written as :

$$Var(x_j) = \tfrac{1}{n}\left\|x_j\right\|^2 = \tfrac{1}{n}x_j' x_j = \tfrac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \tag{2}$$

In this paper, Euclidean distance is used to measure the distances of the center-points in the sample data space, we have:

$$d^2(e_i - e_k) = \left\|e_i - e_k\right\|^2 = \sum_{j=1}^{p}(x_{ik} - x_{jk})^2 \tag{3}$$

In practical problems, different units are generally adopted for different variables, the numeric value of testing data differ with each other extremely. Simply using the Euclidean distance is inappropriate [5]. For instance, the altitude-difference of contact line is generally ten millimeters more or less, while the voltage-difference is about ten kilovolts. Since the variation of the numerical value of altitude-difference is comparatively large, while the variation of the pressure-difference is insignificant, adopting common distance operator $d^2(i, j)$ will exaggerate the effect of the pressure variable and could not reflect the change of data truthfully. In order to eliminate the adverse effect of pseudo-variation, the technique of de-dimension is usually adopted to carry out compression processing for various of variables, viz. to make every variables has same variance 1, namely:

$$x_{ij}^{*} = x_{ij} / s_j \tag{4}$$

where. $s_j = Var(x_j)$.

The standardization processing of data is to carry on centering and compression processing simultaneously, namely:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \tag{5}$$

Where $i = 1,2,...,n$, $j = 1,2,...,p$; take the new data list as $X^{*} = (x_{ij}^{*})_{n\times p} = (x_1^{*}, x_2^{*}, ..., x_p^{*})$. It could be proved that the property exists in variable space E, all of the data have same variance 1, namely: $Var(x_j^{*}) = \tfrac{1}{n}\left\|x_j^{*}\right\|^2 = \tfrac{1}{n}(x_j^{*})' x_j^{*} = 1$.

Therefore, the points of variable set distribute on a super-sphere whose diameter is $\sqrt{n}$, viz. $\left\|x_j^{*}\right\|^2 = n$.

3 The realization of Algorithm

The OCS testing data are un-classification record, clustering analysis is to classify the record appropriately by some criterion and determine each record belongs to witch class..

k-means algorithm and Hierarchical clustering algorithm are 2 kinds of common approaches to deal with clustering process. k-means algorithm try to search k classification which would make squares sum of error be small as far as possible [6]. When these clusters obtained are close and defer with each other clearly indicates the result is good. But, k-means need to confirm the number of classes preliminary, so it's necessary to determine the number of classes using Hierarchical clustering algorithm before carrying on k-means clustering process.

Suppose the observing value of p variable which belong to the i_{th} sample be $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$

$$d_{ij} = (\sum_{k=1}^{p}(x_{ik} - x_{jk})^2)^{1/2} \tag{6}$$

The similarity coefficient of the two samples can be obtained by:

$$a_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2}} \tag{7}$$

3.1 Hierarchical clustering algorithm

When including a subsection you must use, for its heading, small letters, 12pt, left justified, bold, Times New Roman as here.

There are 648 groups of OCS testing data. Merge every two groups into one if they are close enough, and use their center of gravity to represent the new group. Go on merging until the distances of data are greater than the specified threshold value *T*, and complete the classification of all testing data.

It's can be seen that the selection of threshold T plays a important part in the Hierarchical clustering algorithm. If determining an appropriate threshold, then the algorithm could be success. The selection of threshold depends on the practical condition and need to adjust continuously: If the number of classes is comparatively bigger, we need to reduce the threshold. Otherwise, we need to increase the threshold.

The hierarchical clustering algorithm is the most basic and practical algorithm and be defined strictly. If a merging step has been done, it could not be cancelled. This strict define is useful in saving the computation time and cost for it need not be considered about the influence of selection with different merging number.

Through experiments several times, the optimal clustering number is obtained to be 5, and the single points number is 6. The distance of these single points are far from other points and far from each other too.

3.2 k-means clustering algorithm

The k-means algorithm classifies n object into k clusters, so as to make there is a high degree of similarity between data within intra-class, and low similarity degree between inter-classes [4]. Take the clustering number k＝5 as example to explain the procedure of k-means algorithm as follows:

Step 1: The processed data can be considered as a sample space X, and 5 points are randomly selected to be the prototypes: $\{w_1, w_2, w_3, w_4, w_5\}$. The selection principle is let the similarity degree of each point be small as far as possible.

Step 2: Suppose the number of clusters is 5, denoted as $\{C_1, C_2, C_3, C_4, C_5\}$, and are corresponding to 5 prototypes respectively, viz. the center-point of the first class C_1 is w_1, and w_2 for the second class C_2, etc. By removing 5 prototypes from original sample space X, forms new sample space X_1.

Step 3: Read the first group of data $x_1^{'} = (1.8132, 1.3307, 0.1847, -0.1912)^T$ from the sample space X_1, figure out the distance $d_1' = 3.7156$, $d_2' = 3.3211$, $d_3' = 2.6495$, $d_4' = 3.5033$, $d_5' = 1.1797$ from x_1' to each prototypes $\{w_1, w_2, w_3, w_4, w_5\}$ respectively, then determine the minimum distance $d'_{\min} = d_5' = 1.1797$. Thus, consider that x_1' belongs to the 5th class. Regard the gravity center w_5' of w_5 and x_1' as the new prototype $\{w_1, w_2, w_3, w_4, w_5'\}$.

Step 4: Repeat the step 3, until the sample space is empty. Terminate the circulation. Then obtain the final prototypes $\{w_1^*, w_2^*, w_3^*, w_4^*, w_5^*\}$, which are the center of gravity of these classes.

Step 5: Read a group of data x_i from the original sample space X, figure out the distances d_1, d_2, d_3, d_4, d_5 from x_i to each prototypes $\{w_1^*, w_2^*, w_3^*, w_4^*, w_5^*\}$ respectively, and determine the minimum distance $d_{\min}$. For example, suppose $x_i = (0.0022, 1.3972, 0.1235, 3.9485)^T$, and the distances to each prototypes are:

$d_1 = 5.6672$, $d_2 = 1.5747$, $d_3 = 4.3201$, $d_4 = 5.6354$, $d_5 = 4.5462$. Then the minimum distance $d_{\min} = d_2 = 1.5747$, so x_i belongs to the second class. Continue process in this way, and complete the classification.

3.3 The analysis of clustering results

This computation for appraise the clustering results can be done by:

$$S_i = \frac{\min(b(i,:) - a(i))}{\max(a(i), \min(b(i,:)))} \tag{8}$$

where a_i represents the average distance from the i_{th} point to another point of the same class; $b(i,k)$ represents the average distance from the i_{th} point of a certain class to another class (the k_{th} class). Let x-axis represent the appraisal range, whose interval is [-1,+1], where ＋1 denotes that the distance from considered point of a specified class to the adjacent classes is very far, 0 represents that it is not clear that the considered point belongs to which class, -1 represents the result of clustering may be wrong. Y-axis represents the number of classes and the number of data within a class.

As can be seen from the figures bellow, one straight line represents a cluster is just composed by few points, the clustering result is bad, so the simulated effect of clustering with number 5 is the best result.

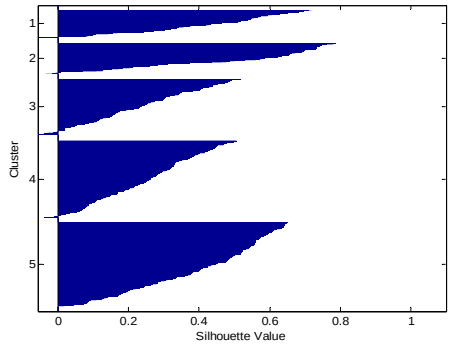

Fig.1 The simulated plot of clustering with 5 clusters

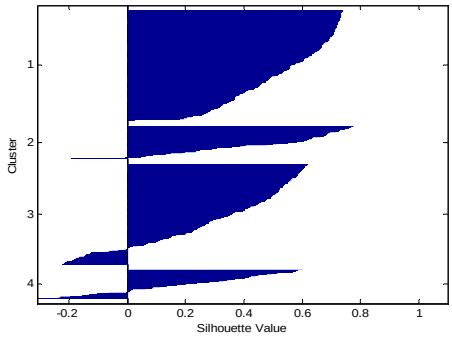

Fig.2 The simulated plot of clustering with 6 clusters

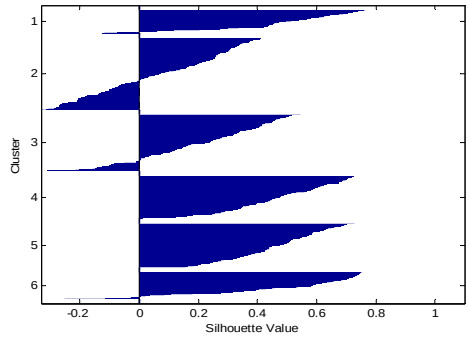

Fig.3 The simulated plot of clustering with 7 clusters

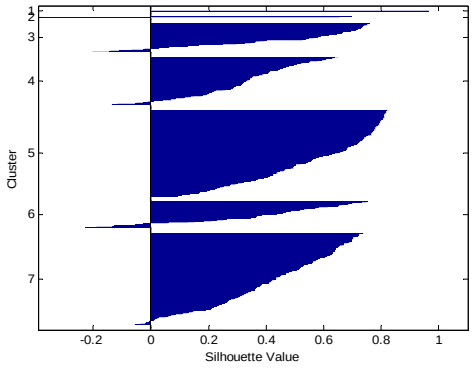

Fig.4 The simulated plot of clustering with 8 clusters

4 Regression Analysis

Through the analysis above, the OCS testing data are clustering to 5 classes in terms of spatial location in order to determine the relationship of hard spot, altitude-difference, velocity and voltage-difference etc. from the OCS testing parameters. Experiments indicate that the testing data exhibit linear distribution in the space. Therefore, it is reasonable to establish the mathematical model by linear regression method [4].

In the process of regression analysis, dependent argument y is assigned to denote the hard spot, while other parameters are expressed with arguments $x_1, x_2, \cdots, x_p$. The general form of linear model is shown as below:

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N_n(0, \sigma^2 I_n) \end{cases} \tag{9}$$

Where y is a $n \times 1$ dependent variable vector; X is a $n \times p$ argument regression matrix; β is the regression coefficient of a $p \times 1$ parameter vector; ε is a $n \times 1$ random disturbance vector. β and ε are independent, and submit Gaussian distribution $N(0, \sigma^2)$ and σ^2 is unknown. $N_n(\mu, \Sigma)$ represents a n-dimensional Gaussian distribution with the mathematical expect value μ and covariance matrix Σ. Observing y for n times independently, viz. to obtain the observed value $y_1, \cdots, y_n$ under the condition of $x_{i1}, \cdots, x_{ip}$ and $i = 1, \cdots, n$. According to Eq.(11), it could be written as:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n \end{cases} \tag{10}$$

The least-squares estimation of parameters could be obtained by the equation below:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{11}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12}$$

Where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$ is the estimated value of y_i, $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY = (y_1, \cdots, y_n)'$ is the estimated value of Y. X' is the transposed matrix of X, $H = X(X'X)^{-1}X'$ is the projection matrix, $e = Y - \hat{Y}$ is called residual error, Q is the sum of residual error with degree of freedom (n-p-1).

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

U is called regression squares sum, whose degree of freedom is p.

$$U = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad (14)$$

It's can be proved that $\hat{\beta}$ is β's optimum linear unbiased estimation and

$$\hat{\beta} \sim N_{p+1}(\beta, \sigma^2 (X^{'} X)^{-1}).$$

The equation below is called general variation squares sum, $S = Q + U$, whose degree of freedom is (n-1).

$$S = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad (15)$$

Obviously, we have $\frac{S}{\sigma^2} \sim \chi^2(n-1)$.

To test whether there exists linear relationship between dependent variable y and argument $x_1, \cdots x_p$ as shown in Eq.(10), we need to examine the assumption as follows:

$$H_0: \beta_1 = \cdots = \beta_p = 0 \quad (16)$$

The statistic used here is:

$$F = \frac{U/P}{Q/(n-p-1)} \quad (17)$$

If H_0 exists, then $\frac{Q}{\sigma^2} \sim \chi^2(n-p-1)$, $\frac{U}{\sigma^2} \sim \chi^2(p)$, and are independent with each other. So, the rejection field with confidence level α is

$$P\{F > F_{1-\alpha}(p, n-p-1)\} = \alpha \quad (18)$$

This means that if observed value F is greater than $F_{\alpha}(p, n-p-1)$, we reject H_0 under confidence level $1-\partial$, and believe that the linear relationship is significant.

5 Conclusion

This paper presents an OCS testing data analysis method based on data mining technique. Mass data are collected by OCS testing equipment. The original OCS testing data are then separated into 2 parts, one for analysis, and the other part for the verification. Through the preprocessing and clustering, we could determine the gravity center of each class respectively. We have determined the classes for each group of data by the combination of hierarchical clustering and k-means clustering algorithm. These steps optimize the results of clustering that can summarize certain number of different physical characteristics of electrified railway section effectively, which contains a lot of information such as curve section, straight-line section, newly built electrified railway section etc.. Then regression analysis is used to search corresponding regression equations that represent the mathematical model of these parameters. Statistics analysis and Simulation results indicate that preprocessing of the original data is necessary to ensure the accuracy of the clustering and regression analysis.

Since the OCS testing data considered here obtains relatively less quantity of the data tested at some fixed points of railway section and observed at different time by the OCS testing car, the mathematical model established here needs to go on further improvement and enhancement. We believe, however, that comprehensive clustering analysis is an effective method to dealing with OCS testing data.

References:

[1] X2. Wan-Ju Yu, *Catenary systems of high-speed electrified railway*, Southwest Jiaotong University Publishing House, 2003.

[2] X1. Gukow, Kiessling, Puschmann, Schmieder, Schmidt, Fahrleitungen elektrischer Bahnen, *B. G. Teubner Stuttgart*, 1997.

[3] X1. Ikeda M, Precise contact force measuring method for current collection system, *Railway Technology Avalanche of RTRI*, Vol.1, No.1, 2003.

[4] X2. Agrawal. R, Automatic Subspace Clustering of High Dimensional Data for Data Mining Application, *Proc. ACM SIGMOD'98 Int. Conf. On Management of Data*, 1998.

[5] X1.Richard J.Roiger, Michael W.Geatz, *Data Mining: A Tutorial-bases Prime*, 2003.

[6] X1.Margaret H. Dunham, *Data Mining Introductory and Advanced Topics*, Tsinghua University Publishing House, 2005.