Research on Technical State Evaluation of Vehicle Equipment based on BIC Cluster Analysis

Xiang-bo Zhang
Department of Technological Support Engineering,
Academy of Armored Force Engineering,
Beijing 100072, P.R.China
e-mail: zxb2002226@sohu.com

Run-hua Qian
Department of Machine Engineering,
Academy of Armored Force Engineering,
Beijing 100072, P.R.China
e-mail: qianrh@163.com

***Abstract*—The technical state evaluation of Vehicle Equipment is a necessary step to operate and support. Considered conditions such as technical characters, operate environments and support elements, this paperresearches its technical state cluster, which is based on BIC(Schwarz's Bayesian Criterion). The conclusion reveals that BIC is accurate and concise to cluster the technical state of vehicle equipment.**

Keywords-Vehicle Equipment; Technical State Evaluation; Cluster Analysis; BIC

I. INTRODUCTION

The technological condition of vehicle equipment is the characters of equipment quality, and reveals its operation ability. For the comprehensive architecture, operation environment, numbers of indexes, the technical state evaluation of vehicle equipment is various and not easy[1-3]. How to assessment it by the flexible and effective methods is important and necessary.

II. CLUSTER

A. Two-step Cluster

Cluster Analysis is the basic method to research the clusters of something[4-7]. It distinguishes anything by quantitative characters, so it is combined with number analysis and multi-variables statistical technique. Although it is rough and not mature method, it can be used easily and the result can be accepted. This method often is applied as exploring tool[8-15].

In the process of cluster analysis, the relation of each sample is based on its quantity, and the relation is measured by its distance. Once the distance is defined, the samples are classed with the same cluster whose distance is near. In tradition, the variables are defined as number. Supposed x_{ik} is k index of i sample, and there are p variables, so the distance D_{ij} between x_i and x_j is defined as

$$D_{ij}(q)=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{q}\right)^{1/q} \quad (1)$$

Equation (1) is called with the Minkowshi distance, where $q>0$.

When $q=1$,

$$D_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \quad (2)$$

Equation (2) is called with the absolute distance or Manhattan distance,

When $q=1$,

$$D_{ij}(2)=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{2}\right)^{1/2} \quad (3)$$

Equation (3) is called with the Euclidean distance.

When $q=\infty$,

$$D_{ij}(\infty)=\max_{1\le k\le p}\left|x_{ik}-x_{jk}\right| \quad (4)$$

Equation (4) is called with the Chebychev distance.

On the other, the distance between variables can also be defined. There are two definitions. One is cosine angle, the other is correlation.

For variable x_i and x_j, cosine angle C_{ij} is defined as

$$C_{ij}=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[\left(\sum_{k=1}^{n}x_{ki}^{2}\right)\left(\sum_{k=1}^{n}x_{kj}^{2}\right)\right]^{1/2}} \quad (5)$$

If $C_{ij}=1$, it shows that x_i and x_j are absolute like.

If $C_{ij}\to 1$, it means that x_i and x_j are closely like.

If $C_{ij}=0$, it means that x_i and x_j are absolute unlike.

If $C_{ij}\to 0$, it means that x_i and x_j are unlike in some degree.

978-1-5090-3619-6/17/$31.00 ©2017 IEEE

Variable x_i and variable x_j, correlation r_{ij} is defined as

$$C_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}-\overline{x_i}\right)\left(x_{kj}-\overline{x_j}\right)}{\left[\left(\sum_{k=1}^{n}\left(x_{ki}-\overline{x_i}\right)^2\right)\left(\sum_{k=1}^{n}\left(x_{kj}-\overline{x_j}\right)^2\right)\right]^{1/2}} \quad (6)$$

C_{ij} or r_{ij} is called with appropriate coefficient. The distance D_{ij} between variables is defined as

$$D_{ij}=\sqrt{1-C_{ij}^2} \quad (7)$$

Or

$$D_{ij}=\sqrt{1-r_{ij}^2} \quad (8)$$

Cluster analysis can solve samples cluster and variables cluster. Samples cluster is called with Q-cluster, and variables cluster is R-cluster. We can select Hierarchical Cluster or K-Means Cluster which is based on the samples quantity. The latter method is a fast cluster. When the samples quantity is large, which exists number variables and cluster variables, two-step cluster can also be applied.

Two-step cluster is a new hierarchical algorithm, so it is usually is applied in mode class, which is the cross field between data mining and multi-variables statistics. The algorithm can be applied in various scale variables.

In cluster analysis, the cluster number is vital. Today, there are two criterions, AIC(Akaike's information criterion) and BIC(Schwarz's Bayesian Criterion) [8-9]. In this paper, BIC is applied widely and maturely.

B. AIC

AIC is firstly proposed by H. Akaike[16-18]. The principle of parsimony is its advantage[19-22]. It can be defined as

$$AIC = -2\ln(\textit{the likelihoodfunction}) + 2(\text{thenumbero freeparameters}) \quad (9)$$

For sample data, $\{X_i \mid i=1,2,\cdots,K\}$, the cluster number is $L_m(m=1,2,\cdots,N)$, and the central position is $\{C_m \mid m=1,2,\cdots,N\}$. The deviation is as $\{D_m \mid m=1,2,\cdots,N\}$, and the number of samples is $\{Q(m) \mid m=1,2,\cdots,N\}$. The distribution density in one cluster is as

$$f(D_m)=\frac{\frac{Q(m)}{K}}{\frac{d_{\max}-d_{\min}}{N}}=\frac{N}{K}\frac{Q(m)}{d_{\max}-d_{\min}} \quad (10)$$

In which,

$$d_{\max}=\max\{D_m \mid m=1,2,\cdots,N\};$$

$$d_{\min}=\min\{D_m \mid m=1,2,\cdots,N\}$$

So can be obtained as

$$l(D \mid C_1,C_2,\cdots,C_N)=\ln L(D_m \mid C_1,C_2,\cdots,C_N) = -N\ln\frac{K}{N}-\sum_{m=1}^{N}\ln\frac{d_{d\max}-d_{\min}}{Q(m)} \quad (11)$$

For AIC,

$$AIC=-2\left(-N\ln\frac{K}{N}-\sum_{m=1}^{N}\ln\frac{d_{\max}-d_{\min}}{Q(m)}\right)+2N = 2\sum_{m=1}^{N}\ln\frac{d_{\max}-d_{\min}}{Q(m)}+2N\left(1+\ln\frac{K}{N}\right) \quad (12)$$

In above $K>N$.

The Equation (11) is AIC in the cluster process. The second item is an increasing function of N, and the error in cluster of the first item is sharply decreased with increasing N, so the first item is down. It shows that the error is smaller, when the clusters are more up. On the other hand, the parameters to estimate become to increase, so the second item is up. Otherwise, the second is down. So the number of clusters should consider the model applicability and compact ability, the goal is to make the minimum AIC.

In a word, when AIC is small, the cluster number is well accepted.

C. BIC

BIC is Schwarz's Bayesian Information Criterion, which is come from the Bayesian theory[23-27]. This criterion is used for model selection. The thought is to balance between the model complexity and data sets describing.

Supposed that $x_1,x_2,\cdots,x_n$, which is independent and same distribution. The behind validated probability is $f(\bullet \mid \theta)$, in which f is a function with k parameters. $N(k)$ is defined as

$$N(k):\{f(\bullet \mid \theta) \mid \theta=(y_1,y_2,\cdots,y_k),\theta\in\Theta_k\} \quad (13)$$

In which, Θ_k is the model space. It supposed that the free parameters number in the model is k. The problem is to select f, which can be best describeded in $N(k)$. For $f\in N(k)$, BIC is defined as:

$$B(f)=\ln L_{\hat{\theta}_k}(X)-k\ln n \quad (14)$$

In which, $\ln L_{\hat{\theta}_k}(X)$ is the maximum likelihood degree of $f(\bullet \mid \hat{\theta})$ in sample data X. $k\ln n$ is the punishment item.