Cluster Analysis of Container Station Based on Container Application Data

Shilin Li[1], Lingxi Zhu[1], Jun Liu[1], Qingying Lai[1] and Xu Wang[2]

[1] School of Beijing Jiaotong University, Beijing 100044, China

[2] China Railway Container Transportation, Beijing 100055, China

Keywords: Railway Transportation, Data Mining, Analysis of Stations, Clustering Algorithm

Abstract. With the increase of the number of railway container manufacturers, the average maintenance cost of each factory's container is uneven. In order to evaluate the quality of the container, it is necessary to do cluster analysis of railway container stations, thus classify the container by the different stations the containers go through. This paper is based on the data of container application for nearly 10 years and constructs the index system of station cluster analysis. Through the cargo ticket and maintenance information of the container, the index data of the station is obtained. Then compares the different results of each clustering algorithm and selects the best clustering algorithm. Finally analyze the data characteristics and the reason of grouping results.

1. Introduction

In recent years, the number of container manufacturer and inventory is increasing, due to the different manufacturer production equipment, production technology and personnel quality, these factors make the qualities of the container are different and also affect the use efficiency of container. Therefore, it is an urgent problem to propose a method to evaluate the container manufacturers.

However, the station, goods category, weight and loading and unloading times of the container are different, so it is necessary to conduct cluster analysis and evaluation on containers with different application conditions. Therefore, it is necessary to cluster the station first to get the containers of different kinds of stations. Therefore, the cluster analysis of the station is the first and necessary work. According to the container property parameter, freight invoice, maintenance information, this paper firstly transforms the form of the tables. Then uses software to get the station such as the Tableau of clustering analysis index and chooses a variety of clustering method to cluster. According to the result of different clustering method, then compares the results of each algorithm to choose the best one. Finally, this paper analyzes the data characters of typical group and proves that the result is reliable.

2. The Process of Station Cluster Analysis

2.1 Station Cluster Analysis Architecture

At present, when the research literature classifies the station, the index used is generally the freight volume, the loading frequency, the economic indicators of the location, etc. But these indicators reflect the total number of goods and do not distinguish between categories of goods. When doing cluster analysis about container station, we need to consider the different impact of the different goods to the container, because under the same operation conditions, different category collision can cause varying degrees of damage to the containers. Therefore, the weight and frequency of loading and unloading should be related to the category of goods, i.e. the weight and number of different goods.

According to the current shortage of clustering analysis, the selection method of clustering indexes adopted in this paper is as follows:

1) For loading and unloading operation, the damage to the container is different due to different categories, so the classification should be considered. According to the way which used to classify the kinds of goods in railway, we take the former two commodity code, from 01 ~ 25 and 99

corresponding goods category, followed by coal, oil, coke, other kinds. For each goods, select the weight of each item and the number of containers to be loaded and unloaded.

2) In the case of loading and unloading operation, due to the different operation mode and mechanical equipment, the selection index should be considered separately.

From what has been discussed above, selected indicators are as follows: loading weight of each goods, loading container number, unloading weight of each goods, unloading container number.

The indicator architecture of the station is shown in figure 1.

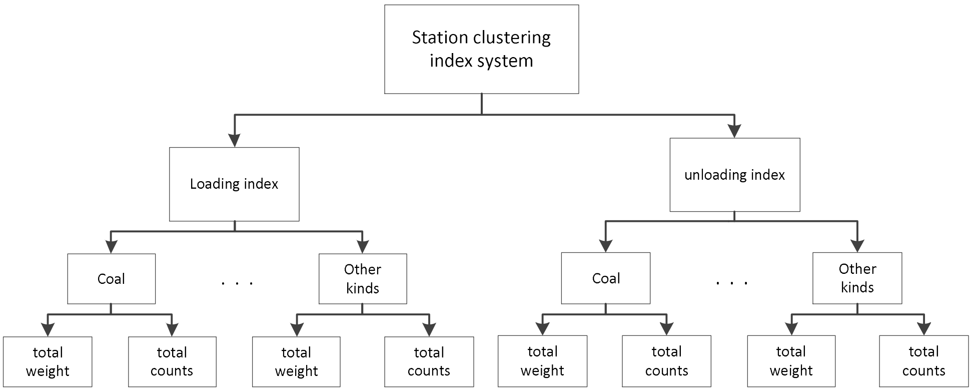

Fig. 1 the structure of station cluster analysis

2.2 Data Processing

The data for the station cluster analysis are from the container equipment management information system and the bill management and fund clearing system of China Railway Container Company. The name, time, and size of the data table are shown in table 1.

Table 1 the size and amount of data source

Table name	Table name in database	Time range		The number of data
Assets Ledger	TY_TLJZXZCTZB	1993.01.01	2017.12.31	503271
Main table of cargo ticket	JXFSHP	2007.01.01	2017.11.21	28424240
Sub-table of container	JXFSXH	2007.01.01	2017.11.21	45743734
sub-table of goods	JXFSPM	2007.01.01	2017.11.21	27914885
Maintenance table	TY_SB_JZXXL_ZB	2007.01.02	2017.12.03	1045268
Assets Ledger	TY_SB_JZXXL_XMXB	2007.01.02	2017.12.03	11853938

We can count the weight and number of every kind of goods according to each stations to get the data. The data for clustering is shown in table 2.

Table 2 Station clustering index data

station	loading weight of coal	loading number of coal	…	loading weight of other kinds	loading weight of other kinds	…	unloading weight of other kinds	unloading weight of other kinds
Fengtai	20000	1	…	40000	2	…	0	0
Wanzhuang	0	0	…	0	0	…	0	0
North of Langfang	0	0	…	0	0	…	0	0
…	…	…	…	…	…	…	…	…
Nancang	432000	16	…	16711000	619	…	945000	35

The aggregation coefficient of each group is obtained by using hierarchical clustering algorithm. The calculation process is as follows:

(1) Calculate the cosine coefficient between sample $X_i(x_1,x_2,...,x_n)$ and $Y_i(y_1,y_2,...,y_n)$. The formula is as follows:

$$COSINE(X,Y)=\frac{\sum_{i=1}^{k}(x_i y_i)^2}{\sqrt{(\sum_{i=1}^{k}x_i^2)(\sum_{i=1}^{k}y_i^2)}} \tag{1}$$

(2) Select the two samples with the smallest cosine coefficient and cluster them into one class, called group 1.

(3) Choose another sample $Z_i(z_1,z_2,...,z_n)$ and calculate the cosine coefficient between this sample and group 1, choose the smallest cosine coefficient and combine them into one group.

(4) Cluster in turn and get the cosine coefficient with the number of clusters.

Use SPSS to get the condensed cluster table according to the station index data, and use EXCEL to get the relationship between number of clusters and condensation coefficient. The clustering gravel diagram is shown in figure 2.

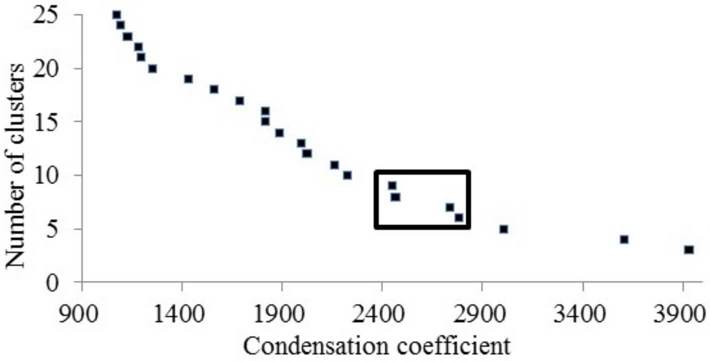

Fig. 2 the rubble map of station cluster analysis

In the figure, the number of groups in the black box is 6~10 groups, indicating the inflection point of the gravel graph. The optimal cluster number is within this range, and silhouette coefficient is used to test.

The silhouette coefficient combines the cohesion and separation degree of clustering which is used to evaluate the effect of clustering. The value of silhouette coefficient is between -1 and 1. When the value is bigger, the effect of clustering is better. The calculation method is as follows.

(1) Assemble all the stations into six groups.

(2) Select sample points in one set of stations $X_i(x_1,x_2,...,x_n)$, Calculation the Euclidean distance between point X_i and all other elements in the same area, write for $a(X_i)$, called the cohesion of this class. The formula of Euclidean distance between sample X_i and sample X_j is as follows:

$$d_{ij}=\sqrt{\sum_{k=1}^{106}(x_{ik}-x_{jk})^2} \tag{2}$$

(3) Select another class b except from the class where sample X_i in, calculates the average Euclidean distance between X_i and the entire samples in class b, iterating through all the other groups, and find the nearest average distance, write for $b(X_i)$. This class is called the neighbor class of sample X_i, $b(X_i)$ is used to quantify the degree of separation between classes.

(4) For sample X_i, the silhouette coefficient $s(i)=(b(i)\text{-}a(i))/max\{a(i),b(i)\}$

(5) Calculate the average silhouette coefficient of all stations, that is The overall contour coefficient under the current category, the coefficient measures the closeness of the data clustering.

(6) Use clustering algorithm to divide all stations into seven groups, eight groups, nine groups, ten groups, Redo the steps (2) ~ (5), then compare silhouette coefficient $s(i)$, the highest value represents the best group number.

Use R program to get the silhouette coefficient is shown in figure 3.

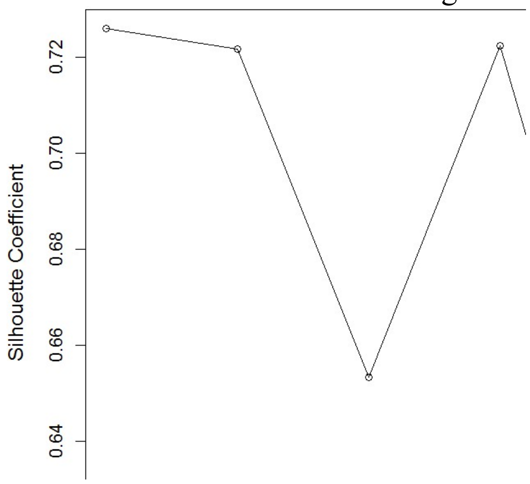

Fig. 3 the silhouette coefficient of station cluster analysis

The group has the highest contour coefficient at 6, so it is better to divide into 6 groups.

3. The Result of Station Cluster Analysis

We use the more commonly used clustering algorithms, the clustering analysis is carried out by the more commonly used clustering algorithms, and the results are shown in Table 3.

Statistical analysis was performed on the number of objects of different clustering algorithms, as shown in table 3.

Table 3 the results of different algorithms

Cluster	K-Means	EM	Ward	TwoSteps
1	1174	1061	1068	964
2	5	15	119	72
3	2	26	6	46
4	16	3	5	71
5	3	83	2	32
6	1	13	1	16

The results of each algorithm are analyzed, and it can be found that TwoSteps has a more balanced grouping quantity of each class and the result is ideal. For index data value, the distribution of goods is extremely uneven. The algorithm, such as k-means, EM and Ward, when dealing with sparse matrices, tends to cluster the outlier points into one class, resulting in a serious deviation. The TwoSteps will first select the regions with relatively concentrated points, and automatically treat the outliers, so the distribution of each group is relatively uniform.

The results of the first and sixth group of TwoSteps are shown in figure 4 and figure 5.

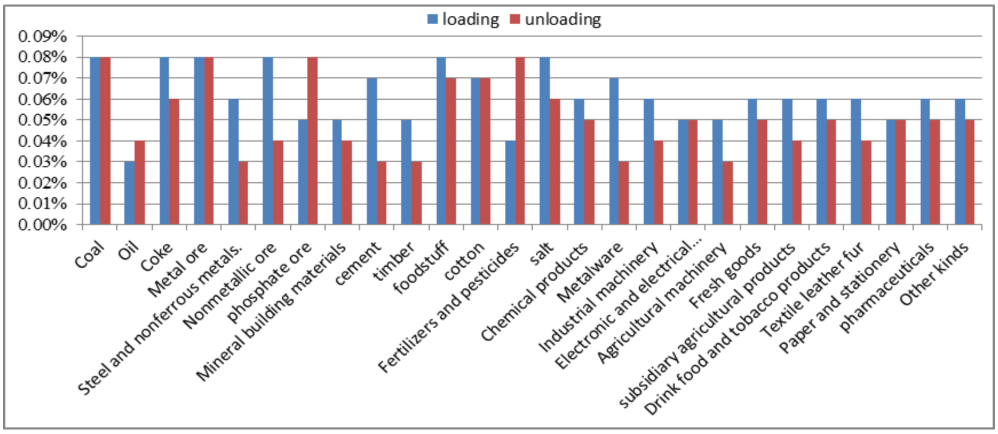

Fig. 4 the average workloads of the 1th group

According to figure 4, we can see that although the first station are numerous, but the workloads are extremely limited, no more than 0.09%, even if there's a certain station of assignment one kind of goods is bigger, but the other kinds of category of this station is very small. Thus, this type represents the station which has the smallest one. In the actual use of container, this kind of station accounted for 96.67%.

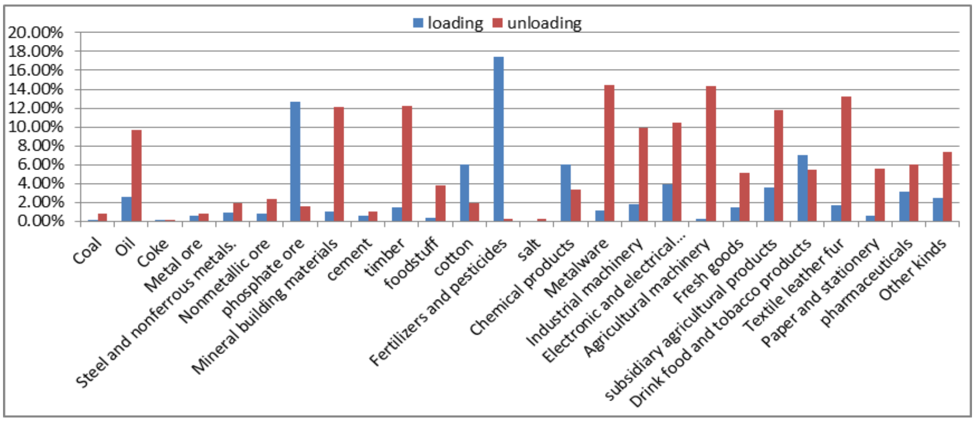

Fig. 5 the average workloads of the 6th group

According to the figure 5, you can see that although only 16 sixth classes at the station, but the workload are very huge. There are nine categories of goods are more than 10.00%, the loading of fertilizers and pesticides production accounted for the largest, more than 16.00%. This group represents the largest station and has the most important position in the national station.

Above all, it can be seen that although the number of each category is very unbalanced, the analysis of actual data shows that the classification is reliable. The reason for such a result is that the distribution of container transport in China is extremely unbalanced at present, and a large number of container operations are concentrated in a few stations.

References

[1]. Tangrong Jian, Mi Lin. The cause of damage of the roof of the container and the countermeasures. Container Transport. Vol. 27 (2016) No. 2, p. 22-25.

[2]. Xinghua Li, Yuntao Ma. Oracle development of classics examples(the second edition). Beijing: Tsinghua University Press, 2009, p. 448-476

[3]. Xue Wei. Statistical analysis and the application of SPSS (the forth edition). Beijing: China Renmin University Press, 2014, p. 128-132.

[4]. Wu Qiang. Railway Container Transportation (the first edition). Beijing: China Railway Press, 2011, p. 248-254