Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (8): 1251-1255, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.1251.1255

Combination Clustering Analysis Method and its Application

Yang Liu, Qin-Liang Li, Li-Yuan Dong and Bang-Chun Wen
Department of Mechanical Engineering and Automation, Northeastern University,
Shenyang, Liaoning, 110004, China

Abstract: The traditional clustering analysis method can not automatically determine the optimal clustering number. In this study, we provided a new clustering analysis method which is combination clustering analysis method to solve this problem. Through analyzed 25 kinds of automobile data samples by combination clustering analysis method, the correctness of the analysis result was verified. It showed that combination clustering analysis method could objectively determine the number of clustering first, and then the members of each clustering could be got. It can be found that the result was identical with the objective reality. Through the comparison of each clustering interior standard deviation and overall standard deviation, it also proved the feasibility of this new clustering method.

Key words: Two step clustering analysis, hierarchical clustering analysis, combination clustering analysis, standard deviation

INTRODUCTION

Clustering analysis is the multivariate statistical analysis method by which the research object are classified according to their characteristics. There are many methods based on different clustering algorithms, such as hierarchical clustering method, density-based method, grid-based method, isolated point analysis method, two-step clustering analysis method and so on (Vijaya *et al*., 2006; Pawlak, 1998). These clustering methods all have their own advantages and disadvantages. The hierarchical clustering method can divide the data set by several clustering levels, but the optimal clustering number can not be automatically determined. The two-step clustering method can get judgment information of different clustering numbers and other descriptive statistics. But it is difficult to get each number of every clustering set by two-step clustering method easily.

In this study, Combination Clustering Analysis method (CCA) was proposed. So, we can evaluate the quality of mechanical product by CCA method. The specific process is as follows: First, the optimal clustering number (n) can be determined by two-step clustering method. And then each member of n clustering set can be got by hierarchical clustering method, in which decomposition process of every layer can be visually showed by tree diagram and composition diagram. At last, the final result would be checked, through the comparison of each clustering interior standard deviation and overall standard deviation. The CCA method will give full play to the advantages of hierarchical clustering method and two-step clustering method.

MATERIALS AND METHODS

Sample selection: In this study, we took several Chinese automobiles as example for clustering analysis by CCA method, according to the information from the website of relevant automobile manufacturers. Through investigation and analysis, seven key indicators were selected as the basis for clustering analysis. They were prices (ten thousand yuan), maximum speed (km h^{-1}), fuel consumption (L 100 km^{-1}), wheelbase (mm), total mass (kg), displacement (mL) and the maximum power (kw). The clustering analysis would be done by the combination method (CCA) of two-step clustering and hierarchical clustering for the data of 25 kinds of automobile on three levels. The sample data are shown in Table 1.

Combination clustering analysis method: CCA method combines the advantages of hierarchical clustering method and two-step clustering method. The specific steps are as follows:

- The best clustering number can automatically be determined according to the Schwarz Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) in two-step clustering method. Specific formula is as follows (Hu *et al*., 2009):

Corresponding Author: Yang Liu, Wenhua Road No. 3, No. 11 Lane, Heping District, Northeastern University, Shenyang, Liaoning, Province, 110004, China

Table 1: Price, top speed, fuel consumption, wheelbase, total quality, displacement and maximum power data of vehicle samples

Serial No.	Price	Top speed	Fuel consumption	Wheelbase	Total quality	Displacement	Maximum power
1	4.58	160	4.8	2340	880	1206	0.0105
2	3.69	151	4.2	2340	870	1000	0.0083
3	3.99	158	5.8	2345	870	999	0.0091
4	4.98	156	4.6	2390	1040	1210	0.0102
5	13.28	185	6.0	2578	1353	1595	0.0123
6	14.88	203	8.38	2578	1405	1390	0.0192
7	10.48	195	7.2	2640	1321	1798	0.0152
8	12.78	180	8.1	2640	1330	1798	0.0152
9	14.28	200	7.5	2640	1345	1899	0.0173
10	10.79	180	8.0	2600	1250	1598	0.013
11	10.49	172	8.3	2600	1260	1598	0.013
12	12.88	200	6.2	2610	1275	1490	0.0182
13	11.88	180	6.8	2610	1265	1598	0.0132
14	11.99	180	6.55	2685	1360	1598	0.0143
15	11.68	180	8.2	2600	1175	1598	0.0133
16	10.88	195	8.5	2600	1145	1598	0.0147
17	10.88	173	8.6	2610	1250	1600	0.0137
18	12.68	182	8.5	2610	1300	1700	0.016
19	18.98	180	8.5	2660	1495	1998	0.0183
20	21.78	175	9.1	2660	1540	1998	0.0183
21	20.48	177	9.8	2630	1486	1997	0.0196
22	19.78	160	9.4	2650	1675	2388	0.0196
23	20.48	155	9.9	2650	1815	2388	0.0196
24	19.58	165	10.5	2630	1625	1975	0.0173
25	21.78	160	8.9	2740	1795	1998	0.026
$\bar{x}$	13.198	176.08	7.6932	2585.4	1325	1680.6	0.0154
s_j	5.5119	15.1683	1.694	0109.167	0254.443	0360.2962	0.004

$\bar{x}$: average, s_j: Standard deviation

$$AIC = -2\ln\left[L\left(D\mid\hat{\theta}\right)\right] + 2m \quad (1)$$

$$BIC = -2\ln\left[L\left(D\mid\hat{\theta}\right)\right] + \ln(n)m \quad (2)$$

In which, $L(D|\hat{\theta})$ is the maximum likelihood value of the observed object D (xi, I = 1,..., n); $\hat{\theta}$ is the maximum likelihood estimator of the clustering number, m is the clustering number, n is the number of observed objects

- The n samples are considered as n clusters. And then two most nearest clusters are merged into a cluster according to "the nearest principle". So, n-1 clusters are got. This process will not finish until there is only one cluster. At the same time, the clustering pedigree chart (tree) will be drawn (Kim *et al*., 2004; Berman *et al*., 2007; Liu and Li, 2006)
- The final clustering number and clustering result are determined according to AIC or BIC and the clustering pedigree chart respectively (Niu *et al*., 2007; Hoai An *et al*., 2007)
- The final result would be checked, through the comparison of each clustering interior standard deviation and overall standard deviation

RESULT ANALYSIS

Before clustering analysis for the data of 25 kinds of automobile, we needed to carry out dimensionless processing for the raw data. So, the standardization data calculated by Eq. 3 were shown in Table 2:

$$b_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \ (i = 1,2,\cdots,n; j = 1,2,\cdots,p) \quad (3)$$

In which, x_{ij} (i = 1,2,....,n; j = 1,2,..., p) is the value of the j-th observation indicator of the i-th sample. $\bar{x}$ and s_j are, respectively the mean and the standard deviation of each observation indicator.

First, BIC in two-step clustering method was applied to the clustering analysis for the sample data. And then the abnormal samples would be excluded. Here we used the log-likelihood distance measure as the distance measure between both samples and setting confidence for 95%. Specific calculation process can be done with software SPSS. So the best clustering number will be got by BIC criterion. The automatic clustering process was done with software SPSS was shown in Table 3.

There are 5 columns in the automatic clustering table. The first column is the clustering number. CCA method learns from the cohesion hierarchical clustering method. So, the default maximum clustering number is 15 and then two clusters with the minimum distance will not be merged until only one cluster forms ultimately. The final clustering result has a relationship with the clustering number, the type and number of the sample, the number of members in each cluster, variance estimates of members in each cluster and so on. The second column is Schwartz Bayesian Criterion (BIC) which was the complexity of clustering result. BIC usually decreases firstly and then increases as the clustering number increases. When BIC is the least, the clustering result is generally considered to

Table 2: Price, top speed, fuel consumption, wheelbase, total quality, displacement and maximum power data after standardization

Serial No.	Price	Top speed	Fuel consumption	Wheelbase	Total quality	Displacement	Maximum power
1	-1.5635	-1.0601	-1.7079	-2.2479	-1.7489	-1.3172	-1.225
2	-1.725	-1.6534	-2.0621	-2.2479	-1.7882	-1.889	-1.775
3	-1.6706	-1.192	-1.1176	-2.2021	-1.7882	-1.8918	-1.575
4	-1.491	-1.3238	-1.826	-1.7899	-1.1201	-1.3061	-1.3
5	0.0149	0.5881	-0.9995	-0.0678	0.11	-0.2376	-0.775
6	0.3052	1.7748	0.4054	-0.0678	0.3144	-0.8066	0.95
7	-0.4931	1.2473	-0.2911	0.5002	-0.0157	0.3258	-0.05
8	-0.0758	0.2584	0.2401	0.5002	0.0197	0.3258	-0.05
9	0.1963	1.577	-0.114	0.5002	0.0786	0.6062	0.475
10	-0.4369	0.2584	0.1811	0.1337	-0.2948	-0.2293	-0.6
11	-0.4913	-0.269	0.3582	0.1337	-0.2555	-0.2293	-0.6
12	-0.0577	1.577	-0.8815	0.2253	-0.1965	-0.529	0.7
13	-0.2391	0.2584	-0.5273	0.2253	-0.2358	-0.2293	-0.55
14	-0.2192	0.2584	-0.6749	0.9124	0.1376	-0.2293	-0.275
15	-0.2754	0.2584	0.2992	0.1337	-0.5895	-0.2293	-0.525
16	-0.4205	1.2473	0.4763	0.1337	-0.7074	-0.2293	-0.175
17	-0.4205	-0.2031	0.5353	0.2253	-0.2948	-0.2237	-0.425
18	-0.094	0.3903	0.4763	0.2253	-0.0983	0.0538	0.15
19	1.049	0.2584	0.4763	0.6834	0.6681	0.8809	0.725
20	1.557	-0.0712	0.8305	0.6834	0.845	0.8809	0.725
21	1.3211	0.0607	1.2437	0.4085	0.6328	0.8782	1.05
22	1.1941	-1.0601	1.0076	0.5918	1.3756	1.9634	1.05
23	1.3211	-1.3897	1.3027	0.5918	1.9258	1.9634	1.05
24	1.1579	-0.7305	1.6569	0.4085	1.179	0.8171	0.475
25	1.557	-1.0601	0.7124	1.4162	1.8472	0.8809	2.65

Table 3: Process of automatic clustering of sample data

Clustering No.	BIC	BIC change value	BIC change rate	Minimum distance change rate
1	163.100			
2	152.171	-10.929	1.000	2.647
3	162.515	10.344	-0.947	1.959
4	203.791	41.276	-3.777	1.237
5	245.595	41.803	-3.825	1.306
6	286.959	41.365	-3.785	1.399
7	327.535	40.576	-3.713	1.239
8	371.459	43.924	-4.019	1.064
9	415.086	43.626	-3.992	1.013
10	458.960	43.874	-4.015	1.227
11	501.528	42.568	-3.895	1.565
12	545.950	44.422	-4.065	0.864
13	590.349	44.400	-4.063	1.060
14	634.945	44.595	-4.081	1.049
15	679.429	44.484	-4.070	1.079

be the best. But the following condition may happen sometimes, namely BIC will continuously decrease as the clustering number increases, so there is no minimum. The third column is BIC change value, which is the BIC change value of two clustering results before and after merger (Srinivasan *et al*., 1994). Generally, it is believed that the larger the absolute value of BIC change value is, the more ideal clustering result is. The fourth column is BIC change rate. It is the measure of BIC change speed. The value range of BIC change rate is in [-1, 1]. It plays an important role in determining the clustering number. The fifth column is the minimum distance change rate. It is generally believed that the larger its value is, the more ideal the clustering result is. And the minimum distance change rate can be deduced by BIC change rate.

After the above four indicators got, the optimal clustering number can be determined by two steps: The first step is called "rough estimate". First, when the clustering number is 2 in the first column of the Table 3, if the BIC change value is larger than 0, the optimal clustering number is 1. Otherwise, the minimum clustering number (assumed to be J) is the optimal clustering number, when the value of BIC change rate is less than 0.04. If all the values of BIC change rate are larger than 0.04, the maximum clustering number in the first column of the Table 3 is the optimal clustering number.

The second step is called "accurate estimate". Searching from the clustering number 2-J-1 in the first column of the Table 3, the two largest value of the minimum distance change rate in the fifth column of the Table 3 can be got. If the ratio of the maximum and the second largest value is larger than 1.15, the clustering number corresponding to the larger minimum distance change rate is the optimal clustering number. Otherwise,

Table 4: Standard deviation of three kind of sample data

SD	Price	Top speed	Fuel consumption	Wheelbase	Total quality	Displacement	Maximum power
I	0.580	3.862	0.681	24.281	83.467	120.389	0.001
II	1.398	10.396	0.916	28.430	71.754	129.815	0.002
III	1.072	9.813	0.678	37.417	135.707	192.822	0.003
Global	5.512	15.168	1.694	109.167	254.444	360.296	0.004

SD: Standard deviation

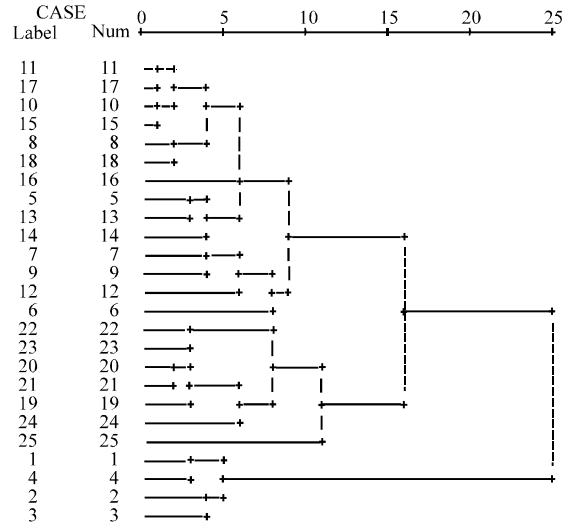

Fig. 1: Arborescence of clustering process

the larger clustering number between them is the optimal clustering number. It should be noted that the two thresholds 0.04 and 1.15 are experience values by a large number of simulation computation of scholars (Srinivasan *et al*., 1999; Liu *et al*., 2012; Liu and Zhu, 2011).

In this study, BIC change value (-10.929) in the third column of the Table 3 was less than 0 when the clustering number was 2. So, it was necessary to look over BIC change rate. The minimum clustering number was 3, while the BIC change rate was less than 0.04. So, the minimum distance change rate should be examined from its corresponding clustering number 2-J-1. Now J was 3, so there was only one value of minimum distance change rate while the clustering number was 2. It could not further carry on "accurate estimate". So, it can be seen that the optimal clustering number was 3 in this case. That is to say, all the 25 samples would eventually be divided into three clusters.

The 25 kinds of car samples would be classified specifically by the hierarchical clustering method after the optimal clustering number got. The sample distance was calculated by the Euclidean distance. And the distance between clusters was calculated by the group connection method. The tree diagram of clustering process was shown in Fig. 1.

Through analyzing the tree diagram, it could be found that the serial number 1-4, 5-18, 19-25 individually formed a cluster when the step length was between 11 and 15. There were a total of three clusters. Through comparing the price, three kinds of clusters has the following characteristics: The average price of cluster I was only 43,100 yuan, less than the price of cluster II and III. This was mainly because there was a strong correlation between the price indicators. The average price of cluster II was 121,300 yuan. And its price indicator was larger than cluster I but less than cluster III. The average price of cluster III was 204,100 yuan. Its price indicator was larger than cluster I and II.

Although, some indicators of samples in one cluster still had some differences, the overall standard deviation of every indicator of all samples was larger than the standard deviation in one cluster (Table 4). So, it could be seen that the clustering result was scientific and rational by CCA method. The clustering result was also consistent with the actual situation. In these car samples, the serial numbers 1-4 were compact cars, 5-18 were economy cars, and 19-25 were off-road vehicles. So, we believed that clustering analysis could be done by CCA was feasible.

CONCLUSION

In this study, we proposed CCA method and illustrated the correctness of the method. The advantage of two-step clustering method is to determine the clustering number. And hierarchical clustering method has visual clustering process. Through the combination of the advantages of two above methods, first the clustering number is scientifically and objectively determined by CCA method. And then clustering result will be got. At last, the rationality of final result will be checked. CCA method is able to complete the task of clustering analysis. It offers a simple solution for clustering analysis.

ACKNOWLEDGMENTS

This study was financially supported by the National Natural Science Foundation of China for Young Scientists (Grant No. 51105065), Exploration-oriented Key Scientific and Technological Innovation Project from Ministry of Education of China (Grant No. N110203001).

REFERENCES

Berman, P., B. DasGupta, M.Y. Kao and J. Wang, 2007. On constructing an optimal consensus clustering from multiple clusterings. Inform. Process. Lett., 104: 137-145.

Hoai An, L.T., L.H. Minh and P.D. Tao, 2007. Optimization based DC programming and DCA for hierarchical clustering. Eur. J. Oper. Res., 183: 1067-1085.

Hu, J.J., J.H. Zhou, J.P. Zhang and H.Q. Yang, 2009. Application of two-step cluster analysis for appearance quality evaluation of flue-cured tobacco. Trans. Chin. Soc. Agric. Mach., 40: 143-146.

Kim, Y.I., D.W. Kim, K.H. Lee and D. Lee, 2004. A cluster validation index for GK cluster analysis based on relative degree of sharing. Inform. Sci., 168: 225-242.

Liu, A.S. and Q. Zhu, 2011. Automatic modulation classification based on the combination of clustering and neural network. J. China Univ. Posts Telecommun., 18: 13-38.

Liu, D.W. and Q. Li, 2006. Application of systematic cluster analysis in the comprehensive appraisal of agricultural production efficiency. Agric. Technol., 26: 36-38.

Liu, G., Y. Li, X. Nie and H. Zheng, 2012. A novel clustering-based differential evolution with 2 multi-parent crossovers for global optimization. Applied Soft Comput., 12: 663-681.

Niu, Z.Y., D.H. Ji and C.L. Tan, 2007. Using cluster validation criterion to identify optimal feature subset and cluster number for document clustering. Inform. Process. Manage., 43: 730-739.

Pawlak, Z., 1998. Rough set theory and its applications to data analysis. Cybern. Sys: Int. J., 29: 661-688.

Srinivasan, D., A.C. Liew and C.S. Chang, 1994. Forecasting daily load curves using a hybrid fuzzy-neural approach. IEE Proc. Gener., Transm. Distrib., 141: 561-567.

Srinivasan, D., S.S. Tan, C.S. Cheng and E.K. Chan, 1999. Parallel neural network-fuzzy expert system strategy for short-term load forecasting: System implementation and performance evaluation. IEEE Trans. Power Syst., 14: 1100-1106.

Vijaya, P.A., M. Narasimha Murty and D.K. Subramanian, 2006. Efficient bottom-up hybrid hierarchical clustering techniques for protein sequence classification. Pattern Recogn., 39: 2344-2355.