Application of clustering algorithms and classification methods for identifying stench data of automobiles

Chinuk Lee[1], Byeongmin Mun[1], Junseop Lee[1], Hyoseok Kim[1], Hansin Lee[1], Sukjoo Bae[1]

[1]*Department of Industrial Engineering, Hanyang University, Seoul, Republic of Korea*

Culee@psm.hanyang.ac.kr
Stat@ psm.hanyang.ac.kr
Hustledo3@psm.hanyang.ac.kr
Irron2004@psm.hanyang.ac.kr
Hanshine8888@psm.hanyang.ac.kr
sjbae@psm.hanyang.ac.kr

ABSTRACT

There has been steady increase in consumer's complaint about the affective quality for automobiles such as the stench inside an automobile. In order to improve the affective quality of smell for consumer, it is crucial to cluster or classify the types of smell at first. We apply several clustering algorithms, such as hierarchical algorithm and K-means algorithm, and classification models, such as decision tree and artificial neural network, and support vector machine, to 26 indices provided from 'A' automobile company. After determining optimal number of clusters, we apply three classification methods and compare their performance in terms of computational time and accuracy.

1. BACKGROUND

In the past, consumers highly care about product quality. Automobile is one of product which becomes necessity for life due to its usefulness. There has been some development and progress of automobile which switch its role of necessity into more luxurious role. As consumer's interest of automobile is increasing, its quality is increasing as well. Mechanical quality of automobile is improving and studies such as initial quality study and vehicle dependability study improve the affective quality as well. However, stench from air conditioner of automobile has been constant problem for corporation and cause increase of complaints of consumer. There are many reasons which cause stench and algorithm to distinguish and classify stench from one another.

2. DATA EXPLANATION

Data is measured by 16 sensors and its size of data is 681. Below box plot shows boxplot of observed data and next plot shows the plot of each data.

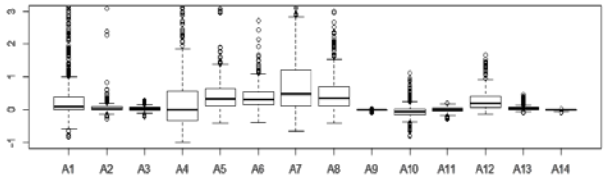

Figure 1. Box plot of observed data set

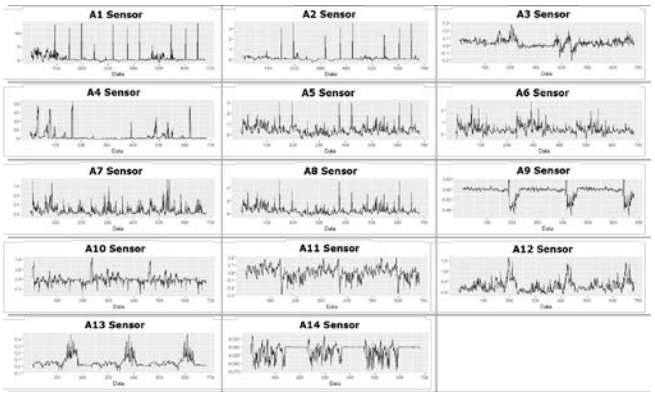

Figure 2. Plot observed data set

3. METHOD

Due to its different dispersion of each data of sensors, we apply the normalization to equalize the dispersion. Below table shows variances of each data set. By applying the normalization, mean of each data set is equalized as 0 and variance as 1.

Table 1. Variance of observed data set

Variable	A1	A2	A3	A4	A5
Variance	3.064	0.121	.003	131.0	0.195
Variable	A6	A7	A8	A9	A10
Variance	0.139	1.691 4	0.404	0.000 3	0.049
Variable	A11	A12	A13	A14	-
Variance	0.006	0.083	0.006	0.000 5	-

3.1. Clustering analysis

Two clustering analyses have been applied while there are not exact standard to distinguish stench data for each other. Clustering analysis measure features of each variables and classify the data by clusters with high similarities. K-means algorithm is use of centroid point and nearby points as clusters and hierarchical cluster categorize by close similarities.

3.2. Deciding the number of clusters

By application of two clustering algorithm, we can set the sufficient number of clusters. By application of multiple indices, we can choose sufficient number of clusters. By Milligan and Copper (1985), quality of index is defends on the features of data and no index are shows preferable performance for every data set. So we applied 26 indices method to decide the optimized number of clusters. Below tables shows the analysis result which include optimized number of clusters and computation time and figure 3 shows 5 clusters which is optimized number of K-mean algorithms decided by application of 26 indices.

Table 2. Result of cluster analysis

K-mean algorithm		Hierarchical cluster analysis	
Computation time	Number of Cluster	Computation time	Number of Cluster
74.32 sec	5	74.30 sec	6

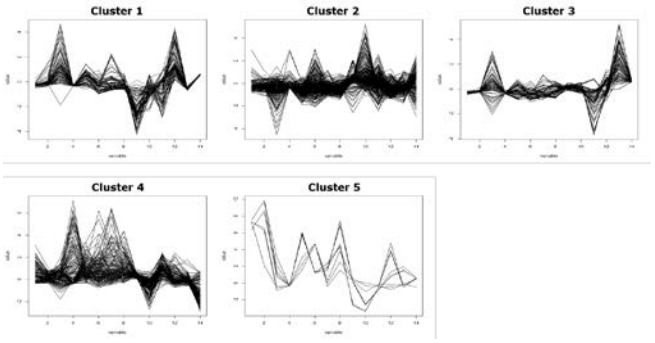

Figure 3. Shows result of clustering analysis by K-means algorithms.

3.3. Classification method

After deciding optimized number of clusters, we applied classification method to categorize the data into each clusters. After optimization of clusters, we divided data by training data set and test data set to evaluate the classification method. 70% of data is set as training data and 30% of data set as test data. Artificial neural network, Decision tree, Support vector machine has been applied to classify data into each clusters by K means algorithm and hierarchical cluster analysis. Table 3 shows accuracy of classification method to evaluate how much test data set has been correctly classified and its computation time for each algorithm.

Table 3. Result of classification method

	K-mean algorithm		Hierarchical cluster analysis	
	Time	Accuracy	Time	Accuracy
Decision tree	0.33	0.8235	0.52	0.7010
Artificial Neural network	5.31	0.9067	12.32	0.8088
SVM	0.19	0.9608	0.91	0.9167

4. CONCLUSION

Analysis of data shows that optimized number of clustering for 5 clusters and 6 clusters for K-mean algorithm and Hierarchical clustering analysis. Support vector machine was preferable classification method comparing to Decision tree and artificial neural network. For further research, other clustering method will be applied and signal processing methods such as Fourier transform and wavelet transform will be used.

5. ACKNOWLEDGEMENT

This work was supported by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20154030200900)

Reference

Glenn W. Milligan & Martha C. Cooper. (1985), An examination of procedures for determining the number of clusters in data set. *Psychometrika*, 50(2), pp159-179, 10.1007/BF02294245

Malay K. Pakhira, Sanghamitra Bandyopadhyay & ujjwal Maulik, (2004), Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(2), pp 487-501, 10.1016/j.patcog.2003.06.005

M. Halkidi, M. Vazirgiannis, & Y. Batistakis, (2002), Quality scheme assessment in the clustering process, *Principles of Data Mining and Knowledge Discovery*, pp 265-276, 10.1007/3-540-45372-5_26

Pang Ning Tang, Michael Stenbach, Vipin Kumar. (2006)., *Introduction To Data Mining*, Addison-Wesley Longman Publishing Co., Inc