IOSR Journal of Mobile Computing & Application (IOSR-JMCA)
e- ISSN: 2394-0050, P-ISSN: 2394-0042.Volume 9, Issue 6 (Nov. - Dec. 2022), PP 51-53
www.iosrjournals.org

Research on customer clustering of new energy automobile industry based on cluster analysis and data mining technology

Chu Fang
College of Economics and Management, Zhaoqing University, Zhaoqing City, Guangdong, China

Abstract: *In recent years, with the improvement of people's living standards, cars have gradually entered people's daily life, thus driving the rapid development of the automotive industry. At the same time, there are many kinds of cars in the market, and there are many performance indicators to measure the quality of cars. Therefore, this study will explore the internal relationship between the types of cars and their performance indicators through multivariate statistical methods, and cluster all kinds of cars according to these indicators. For example, they are divided into three grades of good quality, medium quality and good quality, which can provide guidance for the majority of car buyers and sellers.*

Key Word*: Cluster analysis; new energy automobile industry ; data mining technology*

I. Introduction

This data mainly comes from 20 groups of data and their descriptions, which records the data of several vehicle performance indicators, including 22 samples. Each sample is mainly described by 8 variable indicators, namely economy, service, value, price, design, sport, safety and easyiness.

II. Research model and hypothesis

Before data analysis, first, judge the general level of vehicle performance indicators in the market by describing statistical analysis methods, mainly starting from the calculation and description of basic statistics (such as mean, variance, standard deviation, maximum / minimum value, skewness, kurtosis, etc.) and assisted by the graphic function provided by SPSS, so as to grasp the basic characteristics of the data and the overall distribution characteristics.

In this case, by comparing the mean value and maximum / minimum value of performance indicators of different vehicle models (such as A100, bmw3, ciax, etc.), we can judge which vehicle models have better performance and which have worse performance in general.

III. Empirical Research

	CARMARK	ECONOMY	SERVICE	VALUE	PRICE	DESIGN	SPORT	SAFETY	EASYINESS
1	A100	3.90	2.80	2.20	4.20	3.00	3.10	2.40	2.80
2	BMW3	4.80	1.60	1.90	5.00	2.00	2.50	1.60	2.80
3	CiAX	3.00	3.80	3.80	2.70	4.00	4.40	4.00	2.60
4	Ferr	5.30	2.90	2.20	5.90	1.70	1.10	3.30	4.30
5	FiUn	2.10	3.90	4.00	2.60	4.50	4.40	4.40	2.20
6	FoFi	2.30	3.10	3.40	2.60	3.20	3.30	3.60	2.80
7	Hyun	2.50	3.40	3.20	2.20	3.30	3.30	3.30	2.40
8	Jagu	4.60	2.40	1.60	5.50	1.30	1.60	2.80	3.60
9	Lada	3.20	3.90	4.30	2.00	4.30	4.50	4.70	2.90
10	Mazd	2.60	3.30	3.70	2.80	3.70	3.00	3.70	3.10
11	M200	4.10	1.70	1.80	4.60	2.40	3.20	1.40	2.40
12	Mits	3.20	2.90	3.20	3.50	3.10	3.10	2.90	2.60
13	NiSu	2.60	3.30	3.90	2.10	3.50	3.90	3.80	2.40
14	OpCo	2.20	2.40	3.00	2.60	3.20	4.00	2.90	2.40
15	OpVe	3.10	2.60	2.30	3.60	2.80	2.90	2.40	2.40
16	P306	2.90	3.50	3.60	2.80	3.20	3.80	3.20	2.60
17	Re19	2.70	3.30	3.40	3.00	3.10	3.40	3.00	2.70
18	Rove	3.90	2.80	2.60	4.00	2.60	3.00	3.20	3.00
19	ToCo	2.50	2.90	3.40	3.00	3.20	3.10	3.20	2.80
20	Trab	3.60	4.70	5.50	1.50	4.10	5.80	5.90	3.10
21	VWGo	3.80	2.30	1.90	4.20	3.10	3.60	1.60	2.40
22	VWPa	3.10	2.20	2.10	3.20	3.50	3.50	2.80	1.80

Cluster analysis is a method to classify research objects according to their individual characteristics. Classification is widely used in the fields of economy, management, sociology, medicine and so on. Cluster

analysis can automatically classify a batch of sample (or variable) data according to its many characteristics and the degree of affinity in nature without prior knowledge, and produce multiple classification results. There are similarities among individual characteristics within classes, and there are great differences among individual characteristics among different classes.

After the basic description and statistics of the data, we need to carry out cluster analysis on 23 vehicle types, which can be roughly divided into 3-5 categories, with performance ranging from poor to excellent. In this case, we will use two methods for clustering: one is the system clustering method, and the other is the k-means method (fast clustering method).

The basic principle of the system clustering method: first, a certain number of samples or indicators are regarded as one category, then the two categories with the highest degree of familiarity are merged according to the degree of familiarity of the samples (or indicators), and then the degree of familiarity between the merged category and other categories is considered before merging. Repeat this process until all samples (or indicators) are merged into one category.

Systematic clustering is divided into Q-type clustering and R-type clustering: Q-type clustering is to cluster samples, which makes samples with similar characteristics gather together and separate samples with large differences; R-type clustering is to cluster variables. It separates variables with great differences and gathers similar variables together. In this way, a few representative variables can be selected from similar variables to participate in other analysis, so as to reduce the number of variables and reduce the dimension of variables.

In this case, Q-type clustering is performed.

There are mainly the following methods for calculating the distance between classes:

(1) The nearest neighbor method refers to the minimum value of each individual distance between two classes;

(2) Farthest neighbor refers to the maximum value of each individual distance between two classes;

(3) Between groups linkage refers to the average value of the distance between individuals of two classes;

(4) Within groups linkage refers to taking into account the distance between all individuals of two categories;

(5) Centroid clustering refers to the distance between two class center points;

(6) For the ward method, the sum of the squares of the deviations of similar samples should be small, and the sum of the squares of the deviations between classes should be large.

K-means method (also known as fast clustering method) was proposed by Macqueen in 1967. It regards data as points in k-dimensional space, takes distance as an indicator to measure the "affinity" of individuals, and gains high execution efficiency at the expense of multiple solutions. However, the k-means method can only produce clustering results with a specified number of classes, and the determination of the number of classes is inseparable from the accumulation of practical experience.

The basic idea of fast cluster analysis is: first, select a batch of condensation points (focus) according to a certain method, then let the samples condense to the nearest condensation point to form an initial classification, and then modify the unreasonable classification according to the principle of the nearest distance until it is reasonable. Therefore, in fast clustering, the user should first be asked to give the number of classes to be clustered, and finally only the unique solution about it can be output. Fast clustering is an iterative classification process. In the process of clustering, the class of the sample will be adjusted until it finally reaches stability.

IV. Result

K-means clustering

Select "analysis (a)" → "classification (f)" → "K-means clustering (k)" in the main menu of the data editing window, and the "K-means clustering analysis" dialog box pops up. Select the "region" variable into the "case marking basis (b)" and other variables into the "variable box (V)", as shown in the following figure. Select "iteration and classification" in the "method" radio box, and fill "3" in the "cluster number", indicating that the clustering results will be divided into 3 categories.

Click iterations (I) to open the K-means cluster analysis: iterations dialog box. Fill in 10 (the default value) in the maximum iterations (m), as shown in the following figure, indicating that the set maximum iterations is 10.

（1） Descriptive statistical analysis

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ECONOMY	23	2.10	5.30	3.2913	.18013	.86388	.746	.700	.481	-.132	.935
SERVICE	23	1.60	4.70	3.0609	.16878	.80945	.655	.293	.481	.035	.935
VALUE	23	1.60	5.50	3.1522	.22702	1.08873	1.185	.585	.481	-.004	.935
PRICE	23	1.50	5.90	3.2739	.24777	1.18828	1.412	.678	.481	-.203	.935
DESIGN	23	1.30	4.80	3.2000	.17715	.84960	.722	-.302	.481	.298	.935
SPORT	23	1.10	5.80	3.4652	.21400	1.02629	1.053	-.010	.481	1.179	.935
SAFETY	23	1.40	5.90	3.2870	.23400	1.12222	1.259	.542	.481	.565	.935
EASYINESS	23	1.80	4.30	2.7870	.11767	.56432	.318	1.199	.481	1.886	.935
Valid N (listwise)	23										

Analysis of output results: it can be seen from the analysis of the output results of descriptive statistics that the higher average performance indicators are dynamic (3.4652), economic (3.293) and safety (3.2870); The lowest average performance index was comfort (2.7870). From the perspective of maximum value and minimum value, among all models, the maximum value is price (5.90) and safety (5.90), and the minimum value is power (1.10). The index performance fluctuates greatly in price (1.412) and safety (1.259), and the stability of comfort (0.318) is the best.

The three indicators with high average performance level are power (3.4652), economy (3.293) and safety (3.2870); The lowest average performance level is comfort (2.7870). The index performance fluctuates greatly in price (1.412) and safety (1.259), and the stability of comfort (0.318) is the best.

Comparing the output results of system clustering method and K-means clustering method, it can be seen that the clustering results are roughly the same. A more reasonable clustering method is to divide all samples into three categories. The first category includes 14 vehicle types, such as A100, fofi, Hyun, Mazd, MITs, nisu, OpCo, opve, p306, re19, Rove, toco, vwgo and vwpa. The second category includes 4 vehicle types, such as bmw3, Ferrer, Jagu and m200. The rest of the vehicle types belong to category 3. Obviously, the performance level of Category 3 is the best, category 1 is in the middle, and category 2 is the second.

References

[1]. Zheng, J.; Sun, S. An Improved Algorithm and Theoretical Proof of Hopfield Network for TSP.
[2]. Gong, S.; Zhang, Y.; Wu, H.; Wei, G. A method based on particle swarm optimization and Hopfield network to solve TSP problem.
[3]. Wang, X.; Song, W. Application of Neural Network in Dynamic Production Scheduling of Flexible Production Line.
[4]. Li, W. A Garbage Collection Scheduling Method Based on Neural Network.
[5]. Li, S. A Real-Time Grid Scheduling Algorithm Based on Neural Network.
[6]. Wang, W.; Wu, Q.; Xu, X. Job Shop Scheduling Method Based on Hopfield Neural Network.
[7]. Park, J.; Ha, S. Co-Creation of Service Recovery: Utilitarian and Hedonic Value and Post-Recovery Responses. *J. Retail. Consum. Serv.* **2016**, *28*, 310–316.
[8]. Anderson, K.C.; Knight, D.K.; Pookulangara, S.; Josiam, B. Influence of Hedonic and Utilitarian Motivations on Retailer Loyalty and Purchase Intention: A Facebook Perspective. *J. Retail. Consum. Serv.* **2014**, *21*, 773–779.
[9]. Chiu, C.M.; Wang, E.T.; Fang, Y.H.; Huang, H.Y. Understanding Customers' Repeat Purchase Intentions in B2C E-Commerce: the Roles of Utilitarian Value, Hedonic Value and Perceived Risk. *Inf. Syst. J.* **2014**, *24*, 85–114.
[10]. Qiao, Y.; Zhang, J. TSP Solution Based on an Improved Genetic Simulated Annealing Algorithm.
[11]. Zhou, J.; Deng, Y.; Huang, Y. A Simulated Annealing Algorithm with Memory for TSP Problem.
[12]. Xiang, H.; Zhang, J. An Improved Quantum Genetic Simulated Annealing Algorithm and Its Application in Neural Network Intelligent Fault Diagnosis.
[13]. Yubo, D.; Weiwei, Q.; Qun, Z.; et al. Short Term Load Forecasting Using Recurrent Artificial Neural Network. *J. Jiamusi Univ.* (NATURAL SCIENCE EDITION **2020**.
[14]. Shaofei, J.; Chunli, Z.; Shantong. Z. Discussion on the Improvement Method of BP Network Model. *J. Harbin Jianzhu Univ.* **2020**.
[15]. Fan, C.; Chen, W.; Xi, Y. A Path Planning Method Based on Hopfield Neural Network in Dynamic Unknown Environment.

Acknowledgements

The research result of this subject is "the 2022 subject of Guangzhou philosophy and Social Sciences Planning" and the subject No.: 2022GZGJ88.