K-means Clustering Relational Analysis of The Relations Among Vehicle Models, Sizes and Regions

Jian CHEN, Wenrong JIANG
School of Economic and Management Shanghai Second Polytechnic University, Shanghai, China
{chenjian & wrjiang}@sspu.edu.cn

Abstract - With the enhancement of living standards, Car has become the most out of essential transport tools for us all. But in big cities, one to the rush, the road will be crowded; you can see all kinds of different brands, models, colors and sizes of vehicles. This is not just the status quo in China; most of the countries will have this happen. Then, in different regions, the model of the vehicle, the size will vary it? In this regard, we use K-means clustering algorithm to analysis relations among attributes of vehicles.

Index Terms - K-means, Relational analysis, Vehicle Models.

1. Introduction

1.1 The relationship between vehicle models, sizes and regions

According to the downloaded data from the data hall, analyze the relationship between vehicle models, sizes and regions. Models can be divided into three countries, four countries and five countries, mainly from the tire size and the number of point of view, there are 4, 6, 8, 10, 12, 13, 16 several. China's major provinces in the region are mainly divided into a total of 25 provinces, in a total of 1111 data based in Figure 1.

产品名称	产品商标	批次	更改	发动机	生产企业	排量	功率	前轮距	后轮距
自卸汽车底盘	福建牌	250	0	YN38CRD1YC D4F2L-115	昆明云内动力股份有限公司/广西玉柴动力机械有限	37603922	9585	1660,1750	1650,1650
载货汽车底盘	福田牌	250	1	F3.8s3122	北京福田康明斯发动机有限公司	3760	90	1530,1730	1485,1590
越野载货汽车底盘	北奔牌	257	0	WP10.336E40	潍柴动力股份有限公司	9726	247	1990	1800/1800
载货汽车底盘	解放牌	257	0	CA4DD1-13E4R	中国第一汽车集团公司/中国第一汽车集团公司	2999/2999	99/84	1580	1525
自卸汽车底盘	福田牌	244	1	CA6DF3-18E3	一汽解放汽车有限公司无锡柴油机厂	6740/6000	143/156	1916/1916	1851
自卸汽车底盘	欧曼牌	258	1	WP10.336E40	潍柴动力股份有限公司/潍柴动力股份有限公司	9726/11596	247/276	2005	1865/1865,1880/1880
载货汽车底盘	长安牌	258	1	BH4650/P-AAP10-06	山西成功淮海发动机有限公司	997/997	45/36 39/31	1360	1300

Fig.1 Source data format.

Analyzed by region, Shandong vehicles, followed by Beijing and Anhui, Jiangxi minimal. By models to analyze, to the country three or four main countries, only a small number of vehicles are five of the country. By size analysis, six vehicles are the most common and the most, only a very few of the 13 or 16.

2.K-Means Clustering Analysis

2.1General Introduce for K-Means algorithm

Assuming K-means clustering problem is to have a set of N data set X = {x1, x2, x3, ..., xn} be clustering. K-means clustering problem is to find the value of a division of Pk X = {C1, C2, C3, ..., Ck}, so that the objective function

$$f\ (P_k) = \sum_{i=1}^{k} \sum_{x_i \in C_i} d(x_i, m_i) \quad (1)$$

is minimum . Where, $m^i = 1 / n^i$, Indicates the i-th cluster center position, i = 1, ..., k; ni is the number of data items in the cluster C^i; d (xi, mi) is the distance of x^i to m^i. Typically spatial clustering algorithm is built on the basis of various distances, such as the Euclidean distance, Manhattan distance and clear Cowes distance. Among them, the most commonly used is the Euclidean distance.

The basic idea of K-means algorithm is: Given an n data object database contains, and the need to generate the number of clusters k, k randomly selected objects as the k initial cluster centers; and then the remainder of each sample to calculate each distance from the cluster center, to return to the sample from the class of its nearest cluster center that is located on the way to a new class using the average of the adjusted calculate the new cluster centers; twice if the adjacent cluster centers no change, indicating that the adjustment is completed and the cluster sample average error criterion function has converged. In each iteration of the algorithm must examine each sample classification is correct, if correct, will have to adjust. Modify the cluster centers in all the samples after adjustment is completed, one iteration to the next. If the first iteration, all samples are correctly classified, there will be no adjustment, there will be no change in the cluster center. In the algorithm iteration values decreasing, eventually converge to a fixed value. The guidelines also a measure of the algorithm is based on the correct one.

K-means algorithm process can be described as:

Algorithm: divide and calculate the average cluster object-based.

Input: the number of clusters K and database containing n objects.

Output: K clusters under the square sum of the minimum error conditions.

Methods:

1) K arbitrarily selected objects as the initial cluster centers;

2) division of all objects corresponding to the cluster;

3) calculate the average value of each cluster of objects, all objects will reassign similar clusters;

4) Repeat;

5) until no more changes.

Defective K-means clustering method is that it generates rigid division, that is, each data point is uniquely assigned to one and only one cluster. Thanks in advance not know the actual situation of clustering, so this could be a serious limitation. Meanwhile, K-means algorithm is very easy to fall into local minima and thus can not obtain the global optimal

solution, in a large space vector space search performance. Meanwhile, K-means method for the isolation and abnormal data is sensitive and non-spherical cluster may fail.

2.2 Clustering Results

The initial value of k as 17, which is this 1111 data, according to the models and sizes are divided into 17 different clusters, as the following form.

TABLE I. Initial Center of Clustering

Initial Center of Clustering1~8								
	clusters							
	1	2	3	4	5	6	7	8
x	4	4	4	3	4	4	3	5
d	6	10	4	6	8	12	12	6

Initial Center of Clustering9~15							
	cluster						
	9	10	11	12	13	14	15
x	3	3	3	5	4	5	5
d	8	10	4	12	13	10	8

Initial Center of Clustering16~17		
	cluster	
	16	17
x	4	5
d	16	4

B. After an iterative process, since no changes in the cluster center or minor changes to achieve convergence. Absolute maximum coordinate any changes to the center of .000. Current iteration is 1. The initial minimum distance between the centers is 1.000. Therefore, this process does not bring any change to the results. So this is not a map. And because the data is too large, so the cluster members are not setting them up.

C. After the end of the iterative process, the formation of the distance between the final cluster centers, and final cluster centers. Seen in the following figure.

TABLE 2. Distances among cluster centers

Distances among cluster centers 1~8								
cluster	1	2	3	4	5	6	7	8
1		4.000	2.000	1.000	2.000	6.000	6.083	1.000
2	4.000		6.000	4.123	2.000	2.000	2.236	4.123
3	2.000	6.000		2.236	4.000	8.000	8.062	2.236
4	1.000	4.123	2.236		2.236	6.083	6.000	2.000
5	2.000	2.000	4.000	2.236		4.000	4.123	2.236
6	6.000	2.000	8.000	6.083	4.000		1.000	6.083
7	6.083	2.236	8.062	6.000	4.123	1.000		6.325
8	1.000	4.123	2.236	2.000	2.236	6.083	6.325	
9	2.236	2.236	4.123	2.000	1.000	4.123	4.000	2.828
10	4.123	1.000	6.083	4.000	2.236	2.236	2.000	4.472
11	2.236	6.083	1.000	2.000	4.123	8.062	8.000	2.828
12	6.083	2.236	8.062	6.325	4.123	1.000	2.000	6.000
13	7.000	3.000	9.000	7.071	5.000	1.000	1.414	7.071
14	4.123	1.000	6.083	4.472	2.236	2.236	2.828	4.000
15	2.236	2.236	4.123	2.828	1.000	4.123	4.472	2.000
16	10.000	6.000	12.000	10.050	8.000	4.000	4.123	10.050
17	2.236	6.083	1.000	2.828	4.123	8.062	8.246	2.000

Distances among cluster centers 9~15							
cluster	9	10	11	12	13	14	15
1	2.236	4.123	2.236	6.083	7.000	4.123	2.236
2	2.236	1.000	6.083	2.236	3.000	1.000	2.236
3	4.123	6.083	1.000	8.062	9.000	6.083	4.123
4	2.000	4.000	2.000	6.325	7.071	4.472	2.828
5	1.000	2.236	4.123	4.123	5.000	2.236	1.000
6	4.123	2.236	8.062	1.000	1.000	2.236	4.123
7	4.000	2.000	8.000	2.000	1.414	2.828	4.472
8	2.828	4.472	2.828	6.000	7.071	4.000	2.000
9		2.000	4.000	4.472	5.099	2.828	2.000
10	2.000		6.000	2.828	3.162	2.000	2.828
11	4.000	6.000		8.246	9.055	6.325	4.472
12	4.472	2.828	8.246		1.414	2.000	4.000
13	5.099	3.162	9.055	1.414		3.162	5.099
14	2.828	2.000	6.325	2.000	3.162		2.000
15	2.000	2.828	4.472	4.000	5.099	2.000	
16	8.062	6.083	12.042	4.123	3.000	6.083	8.062
17	4.472	6.325	2.000	8.000	9.055	6.000	4.000

Distances among cluster centers 16~17		
cluster	16	17
1	10.000	2.236
2	6.000	6.083
3	12.000	1.000
4	10.050	
5	8.000	4.123
6	4.000	8.062
7	4.123	8.246
8	10.050	2.000
9	8.062	4.472
10	6.083	6.325
11	12.042	2.000
12	4.123	8.000
13	3.000	9.055
14	6.083	6.000
15	8.062	4.000
16		12.042
17	12.042	

D. When the end of all the algorithms, you can see the number of cases in each cluster.

TABLE 3. Number of Data-sets

Number of data-sets		
cluster	1	308.000
	2	51.000
	3	46.000
	4	365.000
	5	19.000
	6	53.000
	7	79.000
	8	39.000
	9	34.000
	10	79.000
	11	18.000
	12	10.000
	13	1.000
	14	6.000
	15	1.000
	16	1.000
	17	1.000
	effect	1111.000
	missing	.000

3. CONCLUSION

According spss software, existing data, k-means analysis, produced a total of 17 clusters. The first category proportional to the fourth category accounted for the largest shows in most areas of the country, three and four countries most vehicles, and especially the six-wheeled vehicle occupies the top spot, in fact, in the hearts of everyone, the general model, the size of it enough, no how much the car really good, as long as it can play a role in the daily life is enough of. In contrast, these types of vehicles like 13,15,16,17 least 1,000 pieces of data, and only one-thousandth, indicating the country in fact five cars are not common, especially in the four and eight, for the country, four of the car, it is 13 and 16 cars are rare, for the general public is concerned, this car is not the first choice of their hearts. Overall, data analysis or data mining is just one-sided, it does not cover all situations. Only from the data, the vehicle model, the size of the region is the presence of a certain relationship, as in Shandong, Anhui, Beijing and other places, the vehicle is the largest, but also includes almost all models, and in Jiangxi, Shanxi, Xinjiang, Hunan , the vehicle is minimal, only three or four countries national small car. Although these data is not enough to prove anything, but relatively speaking, large population, rapid development of the region, vehicle models, size is also relatively high. Regardless of the results of k-means clustering algorithm, regardless of their own area is not included on all models, as long as they buy their favorite car on it, the other is not important.

References

[1] Bae, E, Bailey, J, Dong, G (2010) A clustering comparison measure using density profiles and its application to the discovery of alternate clusterings. Data Min. Knowl. Discov. 21: pp. 427-477

[2] Birtolo, C., Diessa, V.,etc(2013) Customer churn detection system: Identifying customers who wish to leave a merchant. Applied Artificial Intelligence pp. 411-420

[3] Datta, S, Giannella,etc(2009) Approximate distributed K-means clustering over a peer-to-peer network. IEEE Trans. Knowl. Data Eng. 21: pp. 1372-1388

[4] Karaboga, D., Basturk, B. (2008) On the performance of Artificial Bee Colony (ABC) algorithm. Applied Soft Computing 8: pp. 687-697

[5] Grabmeier, J, Rudolph, A (2002) Techniques of cluster algorithms in data mining. Data Min. Knowl. Discov. 6: pp. 303-360

[6] Linoff, GS, Berry, MJ (2011) Data Mining Techniques. Wiley Publishing Inc, Indianapolis, IN

[7] Strehl, A., Ghosh, J (2002) Relationship-based clustering and visualization for high-dimensional data mining. INFORMS J. Comput. 15, 1–23

[8] Trappey, CV, Trappey, etc (2010) Clustering analysis prioritization of automobile logistics services. Ind. Manag. Data Syst. 110: pp. 731-743

[9] Wu, H, Wang (2013) Div-clustering: exploring active users for social collaborative recommendation. J. Netw. Comput. Appl. 36: pp. 1642-1650