Comparison of K-means, Normal Mixtures and Probabilistic-D Clustering for B2B Segmentation

Satish Garla, Goutam Chakraborty, Oklahoma State University, Stillwater, OK, US
Gary Gaeth, University of Iowa, Iowa City, Iowa, US

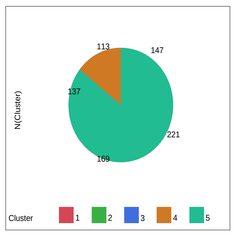
Introduction

Cluster Analysis is a popular technique used by organizations for market segmentation. Clustering splits customers in a market into groups such that the customers within a group are similar and customers between the groups are dissimilar. Several clustering algorithms were suggested in the literature based on a variety of similarity measures. This poster describes a comparative study of three clustering methods (K-means, Normal Mixtures and Probabilistic-D) for segment profiling of customers in a business-to-business (B2B) market. Data collected from a survey conducted by a supplier of hydraulic and pneumatic products was used in this study. Ten variables that measure customers' perceptions of important attributes in selecting a supplier were used for clustering.

The results from each method are evaluated based on cluster purity and cluster profiles. SAS® Enterprise Miner is used for probabilistic-D clustering and for profiling clusters while JMP® Pro 9 is used for K-Means and Normal Mixtures.

K-Means

This is an iterative method where the number of clusters are specified *a priori*. It is a hard clustering in the sense that each observation is assigned to only one cluster. Five clusters as identified from Ward method was used as input to K-means.



Segment	Most Important Predictive Variables
1	warranty policy, return policy, technical documentation
2	electronic payment option, reliability, delivery timeliness
3	credit policy, return policy, price
4	credit policy, price, technical documentation
5	breadth of products, technical documentation

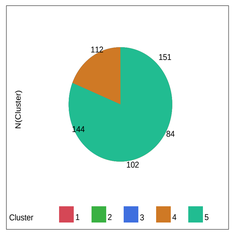
Figure 1. Sizes and profiles of the five clusters from K-means

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	8.65	8.66	6.33	6.76	7.57	4.90	2.35	4.82	5.91	8.22
2	8.08	8.07	7.07	7.51	7.58	6.52	5.71	7.19	7.70	7.97
3	8.54	8.59	7.82	8.21	8.53	7.80	2.37	7.98	8.51	8.35
4	8.77	8.63	7.52	8.37	6.99	4.00	2.29	6.58	8.19	8.58
5	8.73	8.65	4.64	7.43	7.64	5.73	2.31	7.21	8.30	8.51

Figure 2. Mean ratings for all variables from k-means (Red : Highest, Blue: Lowest)

Probabilistic - D

Probabilistic-D clustering is an iterative soft clustering technique in which the cluster memberships of a data point are based on the distances (Euclidean) from the cluster centers. The probability of cluster membership at any point is assumed to be inversely proportional to the distance from the center of the cluster. We used a SAS macro which uses the distances to calculate cluster membership probabilities. In order to compare Probabilistic-D clustering results with other cluster techniques each observation needs to be assigned to only one cluster. After trial-and-errors, we used a probability cut-off of 0.28 to classify 593 observations in five clusters.



Segment	Most Important Predictive Variables
1	electronic payment option, delivery timeliness, reliability
2	breadth of products, return policy, warranty policy
3	warranty policy, return policy, technical documentation
4	price, return plocy, credit policy
5	price, credit policy, return policy

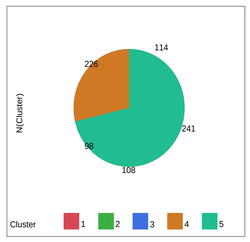
Figure 3. Sizes and profiles of the five clusters from probabilistic-D

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	7.98	7.97	7.52	7.79	7.75	7.01	6.18	7.44	7.86	8.10
2	8.61	8.56	4.71	7.20	7.75	6.07	2.82	7.76	8.48	8.50
3	8.73	8.75	6.12	6.46	7.80	5.10	2.66	5.10	5.73	8.00
4	8.74	8.79	8.10	8.51	8.73	7.67	2.22	8.13	8.67	8.67
5	8.80	8.75	6.96	8.25	6.27	4.10	2.11	5.53	7.72	8.65

Figure 4. Mean ratings for all variables from Probabilistic-D (Red : Highest, Blue: Lowest)

Normal Mixtures

This method also starts with a predefined value for the number of clusters. We used five clusters as identified by Ward's method. JMP uses the EM algorithm, an iterative optimization method that estimates probabilities for each observation to belong to each cluster by assuming that the joint probability distribution of the clustering variables can be approximated by a mixture of multivariate Normal distributions, which represent different clusters.



Segment	Most Important Predictive Variables
1	credit policy, delivery timeliness, technical documentation
2	direct access to salesreps, delivery timeliness, reliability
3	reliability, credit policy, delivery timeliness
4	return policy, credit policy, technical documentation
5	direct access to salesreps, delivery timeliness, warranty policy

Figure 5. Sizes and profiles of the five clusters from Normal Mixtures

Segment	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
1	8.19	7.88	5.90	6.80	7.08	3.89	2.08	5.98	6.96	8.04
2	9.00	9.00	7.05	7.97	7.71	5.90	3.21	6.74	8.00	9.00
3	7.87	8.05	6.78	7.65	7.73	6.74	3.35	6.83	7.92	8.15
4	8.97	8.97	8.34	8.97	8.97	8.00	4.48	8.97	8.97	8.97
5	8.20	8.20	6.41	7.20	7.41	5.74	3.42	6.37	7.14	7.42

Figure 6. Mean ratings for all variables from Normal Mixtures (Red : Highest, Blue: Lowest)

Comparison

Comparison of cluster means from all the three techniques (Figures 2, 4 and 6) we can observe that probabilistic-D and Normal Mixtures tend to separate the means across clusters better than the k-means. Better separation makes profile of segments easier to understand and easier to act upon for developing tailored marketing communications. In addition to straightforward mean comparisons, difference in the results from the clustering methods can also be identified by looking at the range of means (maximum mean rating – minimum mean rating) for each attribute across the clusters. Figure 7 shows the range of the attribute means reported in Figures 2, 4 and 6.

Method	reliab	time	av_br	av_spec	price	credit	av_pay	return	warranty	talk_dir
K-means	0.69	0.59	3.19	1.61	1.54	3.80	3.42	3.15	2.60	0.61
Probabilistic-D	0.82	0.82	3.39	2.05	2.46	3.58	4.07	3.03	2.94	0.67
Normal Mixtures	1.13	1.12	2.43	2.17	1.89	4.11	2.40	2.99	2.01	1.58

Figure 4. Range of means for each variable from all the three techniques (Blue: Highest)

Overall satisfaction rating with the supplier, measured on a 11 point scale, was also used to compare the purity of the clusters. We measured the percentage of customers in each cluster who are highly satisfied with the current supplier. Normal Mixtures seems to separate the clusters better than k-means or probabilistic-D methods.

Cluster	K-means	Probabilistic-D	Normal Mixtures
1	38%	40%	41%
2	39%	44%	43%
3	40%	40%	28%
4	43%	46%	51%
5	38%	46%	37%

Figure 5. Percent classification of customers who rated high on satisfaction

Conclusion

Our results show a very wide difference in the profiles of clusters generated from each method. In most practical applications, the shapes of clusters, the distributions of clustering variables, number of clusters, etc. are unknown. Therefore, it is not possible to theoretically justify one clustering method over another because of the assumptions of each of these methods. Therefore, at the end of the day, the value of each clustering method has to be evaluated by domain experts to judge the usefulness of each solution. Using a descriptor variable, such as the overall satisfaction that was not used in deriving the clusters, can also help to a certain extent in validating the cluster results. Given this criterion, our analysis shows that Normal Mixtures is performing slightly better when compared to other methods. This suggests that analysts may gain valuable insights by routinely including the Normal Mixtures along with other cluster techniques.