Customer Segmentation through K-Means and Hierarchical Clustering Techniques

Anika Singh[1] Prasanna Balaji Venkataraman[2]
[1,2]Department of Marketing Management
[1,2]Great Lakes Institute of Management India

Abstract— Clustering techniques are used to group data/observations in a few segments so that data within any segment are similar while data across segments are not similar. Defining what similar and different observations mean is an important part of cluster analysis which often requires a lot of contextual knowledge and creativity beyond what statistical tools can provide. Based on how we define similarities and differences between data observations like customers or assets, which can also be defined mathematically using distance metrics, one can find different segmentation solutions.

Key words: Clustering Techniques, Customer Segmentation through K-Means and Hierarchical Clustering Techniques

I. Introduction

Clustering techniques are used to identify segments in the data. For customer data, even though we may have data from millions of customers, these customers may only belong to a few segments: customers are similar within each segment but different across segments. We may often want to analyze each segment separately, as they may behave differently for example different market segments may have different product preferences and behavioural patterns.

Key component of clustering and segmentation is definition of distance metrics (between observations), which need to be defined creatively based on contextual knowledge and not only using 'black box' mathematical equations and techniques. It is often necessary to split the data into segments and perform subsequent analysis within each segment in order to develop (potentially more refined) segment-specific insights.

The purpose of this research is to know the types of people who are, or who could be, visiting their respective mall. Segmenting the market will help us designing and positioning the shopping mall services better in order to attract a few profitable market segments, or to differentiate the services (like invitations to events, discounts, etc.) across market segments. The goal of this research paper is to have clear view of clustering and segmentation.

II. Objectives

- Identification of potential clients through Data Analytics techniques: K-Means and Hierarchical techniques
- Comparison between K-Means and Hierarchical cluster outputs
- Visualisation of segmented clusters through above mentioned techniques

III. Data Collection

Market research survey was run with few potential customers. In this paper a small survey was run, where each person answered six attitudinal questions and a question regarding how often they visit the mall, all on a scale 1-7, as well as one question regarding their household income. Following are the questions asked:

Name	Description	Scale
V1	Shopping is fun	1-7
V2	Shopping is bad for your budget	1-7
V3	I combine shopping with eating out	1-7
V4	I try to get the best buys while shopping	1-7
V5	I don't care about shopping	1-7
V6	You can save lot of money by comparing prices	1-7
Income	The household income of the respondent	Dollars
Mall.Visits	How often they visit the mall	1-7

Fig. 1: Questions Asked

ID	V1	V2	V3	V4	V5	V6	Income	Mall.Visits
1	6	4	7	3	2	3	60000	3
2	2	3	1	4	5	4	30000	1
3	7	2	6	4	1	3	70000	3
4	4	6	4	5	3	6	30000	7
5	1	3	2	2	6	4	60000	1
6	6	4	6	3	3	4	50000	2
7	5	3	6	3	3	4	65000	3
8	7	3	7	4	1	4	55000	4
9	2	4	3	3	6	3	70000	0
10	3	5	3	6	4	6	25000	6

Fig. 2: Response Received

Hundred people responded to above questions. Following are the responses for first 10 people:

IV. Statistical Analysis

Now we are to find how this data of 40 people can be segmented, or are there actually any segments in data

V. Methodology

There is not one single way to for clustering and segmentation. We follow the following process:

- Confirm data in metric
- Decide whether to scale or standardize the data
- Decide which variables to use for clustering
- Define similarity or dissimilarity measures between observations
- Visualize Individual Attributes and Pair-wise Distances between the Observations
- Select the clustering method to use and decide how many clusters to have
- Profile and interpret the clusters
- Assess the robustness of our clusters

A. Confirm Data is Metric

Mostly clustering methods require data to be metric. This means that data should have actual numeric meaning, i.e., 1 should be less than 2, which should be less than 3 etc. Main reason for this is that researcher needs to define distances between observations. The main reason for this is that one needs to define distances between observations and often ("black box" mathematical) distances (e.g. the "Euclidean distance") are defined only with metric data. Descriptive statistics of our data is given below:

All rights reserved by www.ijsrd.com

	min	25 percent	median	mean	75 percent	max	std
ID	1	25.75	50.5	50.5	75.25	100	11.69
V1	1	2.00	4.0	3.85	5.25	7	1.87
V2	2	3.00	4.0	4.10	5.00	7	1.39
V3	1	2.00	4.0	3.95	6.00	7	1.99
V4	2	3.00	4.0	4.10	5.25	7	1.50
V5	1	2.00	3.5	3.45	4.25	7	1.74
V6	2	3.00	4.0	4.35	5.25	7	1.48
Income	25000	30000.00	42500.0	46000.00	60000.00	80000	17216.57
Mall.Visits	0	2.00	3.0	3.25	4.25	7	1.94

Fig. 3: Descriptive Statistics

B. Scaling/Standardizing

Six of the survey questions are on similar scale (1-7), but income variable is different. Having variables at different scale in data creates problems as most of the results may be driven by a few large values. To avoid such issues, data is standardized by making raw attributes have mean 0 and standard deviation 1, (e.g. scaledIncome = (Income-mean (Income)) / sd(Income)), or scaling them between 0 and 1 (e.g. scaledIncome = (Income-min(Income)) / (max(Income)-min(Income))).

Here is for example the R code for the first approach, if we want to standardize all attributes:

```
ProjectData_scaled = apply(ProjectData, 2, function(r) {
    if (sd(r) != 0)
        res = (r - mean(r))/sd(r) else res = 0 * r
    res
})
```

Fig. 4: Sample R Code

	min	25 percent	median	mean	75 percent	max	std
ID	-1.67	-0.83	0.00	0	0.83	1.67	1
V1	-1.52	-0.99	0.08	0	0.75	1.68	1
V2	-1.51	-0.79	-0.07	0	0.65	2.08	1
V3	-1.49	-0.98	0.03	0	1.03	1.54	1
V4	-1.40	-0.73	-0.07	0	0.77	1.93	1
V5	-1.41	-0.83	0.03	0	0.46	2.04	1
V6	-1.59	-0.91	-0.24	0	0.61	1.79	1
Income	-1.22	-0.93	-0.20	0	0.81	1.97	1
Mall.Visits	-1.67	-0.64	-0.13	0	0.51	1.93	1

Fig. 5: Standardized Data

All variables have mean 0 and standard deviation 1.

Data will not be standardized when we want analytics findings to be driven mainly by a few attributes that take large values or when different scales of attributes imply something about those attributes.

C. Variables Selection for Segmentation

Choosing variables which gives us a good sense of what variables may distinguish people or products or assets or regions is critical.

We use only a few of the data attributes for segmentation (the segmentation attributes) and use some of the remaining ones (the profiling attributes) only to profile the clusters. In this paper, we use the 6 attitudinal questions for segmentation, and the remaining 2 (Income and Mall. Visits) for profiling later.

D. Define Similarity Measure

Clustering is done to group observations based on similarity. It is therefore important that we understand what makes two observations (e.g. customers, products, companies, assets, investments, etc.) similar. Most statistical methods for clustering and segmentation use common mathematical measures of distance. Typical measures are, for example, the Euclidean distance or the Manhattan distance. The Euclidean distance between two observations (customers in our paper) is the square root of the average of the square difference between the attributes of the two observations (customers).

For example, the distance of the first customer in our data from customers 2-10 (summarized above), using their responses to the 6 attitudinal questions is:

1	2	3	4	5	6	7	8	9	10
0									
8	0								
3	8	0							
6	6	7	0						
8	3	9	7	0					
2	7	3	4	7	0				
2	6	3	5	6	1	0			
2	9	2	6	9	3	3	0		
7	3	8	6	2	6	5	8	0	
7	4	7	2	6	6	6	7	5	0

Fig. 6: Response O Attitudinal Questions

If we use the Manhattan distance metric, these distances change as follows:

1	2	3	4	5	6	7	8	9	10
0									
16	0								
6	16	0							
13	13	15	0						
17	5	19	16	0					
3	13	7	10	14	0				
5	11	7	10	12	2	0			
5	15	3	14	18	6	6	0		
12	6	16	13	5	11	11	17	0	
16	10	18	5	13	13	13	17	10	0

Fig. 7: Using Manhattan Metrics

This step may require multiple iterations.

E. Visualization of Pair Wise Distance

Using various distance metrics, distance between two pairs to be visualized now. If there are indeed multiple segments in data, some of these plots should show "mountains and valleys", with the mountains being potential segments. For example, in our case we can see the histogram of, the first 2 variables:

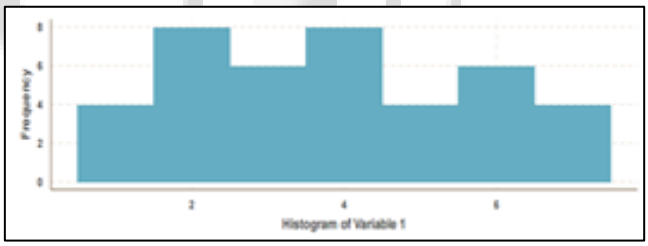

Fig. 8: Histogram of First Variable

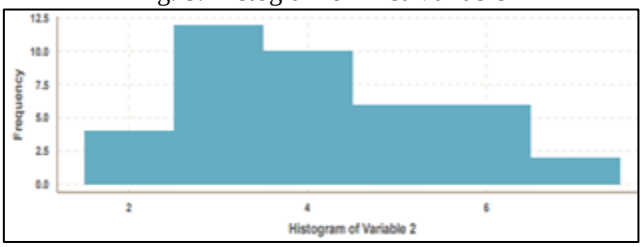

Fig. 9: Histogram of Second Variable

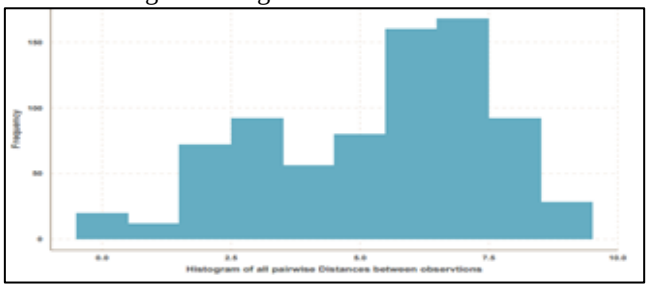

Fig. 10: Histogram of All Pair Wise Distances for the Euclidean Distance

All rights reserved by *www.ijsrd.com*

F. Number of Segments and Method

Method chosen should be statistically robust, interpretable, and actionable. In this paper we will use the K means Clustering Method, and the Hierarchical Clustering Method. If we do not consider methodology, the only difference to highlight is that K-means requires the user to define how many segments to create, while Hierarchical Clustering does not Various Types of clustering are explained below:.

1) Hierarchical Clustering

We will first use hierarchical Clustering method, as we do not know how many segments there are in the data. It will help us visualise how the data may be clustering together. Plot called the Dendrogram will be generated. Hclust command was used in R to create it (using Euclidean distance and the ward.D hierarchical clustering option as well as help(hclust) in R).

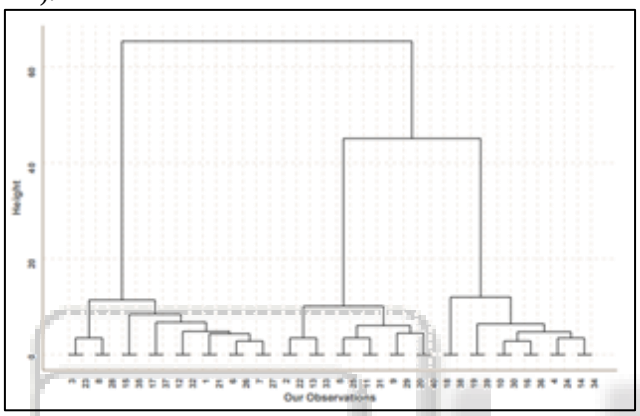

Fig. 11: Observed Data 1

We can draw as many clusters are we want here we choose to make 3 clusters around the branches of the Dendrogram.

In Dendrogram, observations from data are grouped together, starting from pairs of individual observations that are the closest to each other, and merging smaller groups into larger ones, depending on which are the groups closest to each other. Eventually all the data are merged into single segment. The heights of branches of the tree indicate how much difference is there in the clusters merged at that level of the tree. Longer lines indicate that clusters below are very different. The heights of the tree branches increase as we traverse the tree from the end leaves to the tree root- this method merges data points/groups from the closest ones to the furthest ones.

Dendrograms are very helpful visualization tool for segmentation, even if the number of observations is very large. The tree typically grows logarithmically with the number of data. However, these can be misleading. Once two data points are merged into the same segment they remain in the same segment throughout the tree. This tendency of rigidity of the Hierarchical Clustering method may lead to segmentations which are suboptimal in numerous ways. Though, the dendrograms are useful in practice to help us get some understanding of the data, including the potential number of segments we would have in the data. Moreover, there are various ways to construct the dendrograms, not only depending on the distance metric we defined in the earlier steps above, but also depending on how the data are aggregated into clusters (help command in R- help (hclust), for example, which provides the following options for the way the tree is constructed: ward, single, complete, average, mcquitty, median or centroid). We can also plot distances travelled before we need to merge any of the lower and smaller in size clusters into larger ones - the heights of the tree branches that link the clusters as we traverse the tree from its leaves to its root. If we have n observations, this plot has n-1 numbers.

As a thumb rule, we can select the number of clusters as the "elbow" of this plot. Elbow is the place in the tree where, if we traverse the tree from the leaves to its root, we need to make the 'longest jump' before we merge further the segments at that tree level. Obviously actual number of segments can be very different from what this rule of thumb may indicate, in practice we usually explore different numbers of segments, possibly starting with what a hierarchical clustering Dendrogram may indicate, and then we select the final segmentation solution using both statistical and qualitative criteria, as discussed below.

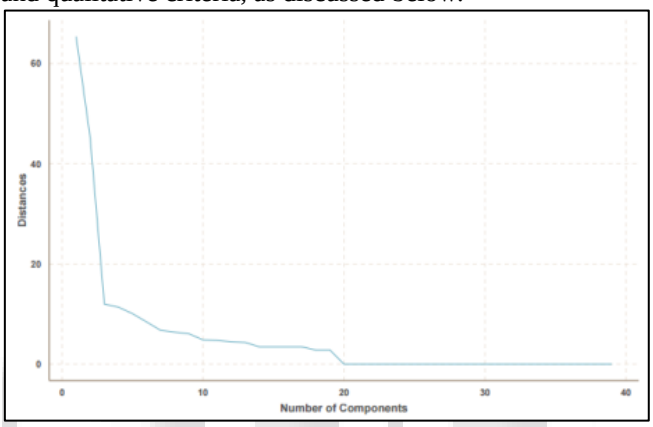

Fig. 12: The Components

Different numbers of segments should be explored, that is multiple iterations, and the final choice should be made based on both statistical and qualitative criteria. For this paper we consider the 3-segments solution found by the Hierarchical Clustering method (using the Euclidean distance and the hclust option ward.D).

We can also see which segment each observation (respondent) belongs to for the first 10 people:

Observation Number	Cluster_Membership
1	1
2	2
3	1
4	3
5	2
6	1
7	1
8	1
9	2
10	3

Fig. 13: Observed Data 2

2) K-Means Clustering

K-means can also be performed using numerous distance metrics, same like hierarchical clustering. There are variations of K-means (like Hartigan-Wong, Lloyd, or MacQueen, for knowing more we can see help(k-means) in R). K-means does not necessarily lead to the same solution every time you run it. Here are the clusters our observations belong to when we select 3 clusters and the Lloyd k-means method, for the first 10 people (cluster IDs may differ from those from hierarchical clustering):

All rights reserved by *www.ijsrd.com*

Observation Number	Cluster_Membership
1	1
2	2
3	1
4	3
5	2
6	1
7	1
8	1
9	2
10	3

Fig. 14: Observed Data 3

The observations need not be in the same clusters as different methods are used, neither do the segment profiles. Though, a characteristic of statistically robust segmentation is that observations will be grouped in similar segments independent of the approach used. Also, the profiles of the segments should not differ much when different approaches are used.

Now, profile segments need to be calculated. Before that we should know that most of the observations should belong in the same clusters independent of how the clusters are found. The segments found should be relatively robust to changes in the clustering methodology and data subsets used. Additionally, the profiles of the clusters found using different approaches should be as consistent across different approaches as possible.

Large changes may indicate that our segmentation is not valid. Judging the quality of segmentation depends on both robustness of the statistical characteristics of the segments (like changes from different methods and data used) as well as a matter of many qualitative criteria: action ability, stability over time, interpretability, etc.

G. Profile and Interpretation

Once number of clusters is decided, we get a better understanding of who the customers in those clusters are and interpret the segments. Researcher can now see how the summary statistics (e.g. averages, standard deviations, etc.) of profiling attributes differ across the segments.

In this paper we use the 3 segments found using hclust (similar profiling can be done with the results of other segmentation methods), differences in responses of survey can be seen across segments.

The average values of the data for the total population as well as within each customer segment are:

	Population	Segment 1	Segment 2	Segment 3
V1	3.85	5.75	1.67	3.50
V2	4.10	3.62	3.00	5.83
V3	3.95	6.00	1.83	3.33
V4	4.10	3.12	3.50	6.00
V5	3.45	1.88	5.50	3.50
V6	4.35	3.88	3.33	6.00
Income	46000.00	60000.00	42500.00	30833.33
Mall.Visits	3.25	3.25	1.00	5.50

Fig. 15: Average Values

We use snake plots to visualize segments for each cluster. For example, we plot the means of the profiling variables for each of the clusters to better visualize differences between segments.

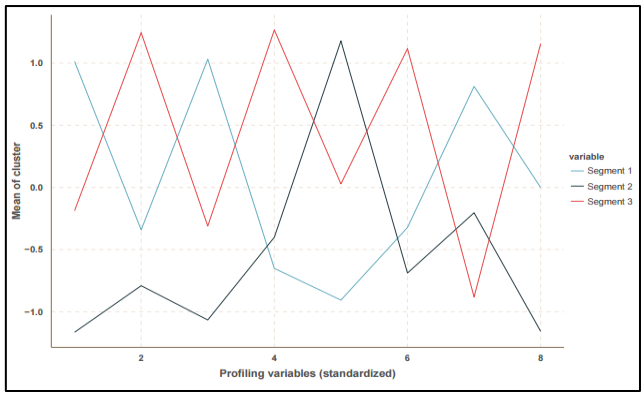

Fig. 16: The Plot of Standardized Profiling Variables

This plot helps us seeing differences between the segments, if the segments differ in terms of their average household income and in terms of their visit to mall, are there some other additional attributes to be noticed about these parameters. We can also compare the averages of the profiling variables of each segment relative to the average of the variables across the whole population.

This helps us better understand whether there are indeed clusters in data (for example if all segments are much like the overall population, there may be no segments). Like, we can measure the ratios of the average for each cluster to the average of the population (e.g. avg(cluster) / avg (population)) and explore a metrics as the following one:

	Segment 1	Segment 2	Segment 3
V1	1.49	0.43	0.91
V2	0.88	0.73	1.42
V3	1.52	0.46	0.84
V4	0.76	0.85	1.46
V5	0.54	1.59	1.01
V6	0.89	0.77	1.38
Income	1.30	0.92	0.67
Mall.Visits	1.00	0.31	1.69

Fig. 17: Average of Each Cluster

The further a ratio is from 1, more important that attribute is for a segment relative to the total population. Both the snake plot as well as this metrics of relative values of the profiling attributes for each cluster help us visualize our segments and interpret them.

H. Robustness

Segmentation is an iterative process with many variations of data, methods, number of clusters, and profiles generated until a satisfying solution is reached. It can be done using variations of the original segmentation attributes, using different subsets of the original data, using different numbers of clusters, using different distance metrics, using different segmentation methods.

We can't explore all variations here, but a few techniques were performed to test the statistical robustness and stability of interpretation are:

- How much overlap is there between the clusters found using different approaches? Specifically, for what percentage of our observations the clusters they belong to are the same across different clustering solutions?
- How similar are the profiles of the segments found? Specifically, how similar are the averages of the profiling attributes of the clusters found using different approaches? As we can have the cluster memberships of our observations for all clustering methods, we can

All rights reserved by *www.ijsrd.com*

measure both the total percentage of observations that remain in the same cluster, as well as this percentage for each cluster separately.

VI. CONCLUSION

Use of different number of clusters leads to different percentages of overlap (can be tried using 2 clusters) - robustness of solution may also indicate how many clusters there are in data. However, in general there is no 'correct percentage of overlap', as this depends on how difficult clustering is: the robustness of solution is often "relative to other solutions". Additionally, sound segmentation requires eventually robustness of our decisions across many good clustering approaches used. Only after a number of such robustness checks, profiling and interpretations, researcher can end with final segmentation. We may need to return to our original raw data at any point and select new raw attributes as well as new clusters before reaching to our final segmentation as data analytics is an iterative process.

ACKNOWLEDGEMENT

We are highly indebted to our parents and all our faculty members at Great Lakes Institute of Management.

REFERENCES

[1] MacKay and David, "An Example Inference Task: Clustering," Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.
[2] I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.
[3] S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.
[4] H. Zha, C. Ding, M. Gu, X. He, and H. D. Simon, "Spectral Relaxation for K-means Clustering," Neural Information Processing Systems, Vancouver, Canada, vol.14, pp. 1057-1064, 2001.
[5] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in Proc.10th ACM Symposium on Computational Geometry, 1994, pp. 332-339.
[6] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hard Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.
[7] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The Planar K-Means Problem is NP-Hard," LNCS, Springer, vol. 5431, pp. 274-285, 2009.
[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.
[9] C. Elkan, "Using the triangle inequality to accelerate K-means," in Proc. the 12th International Conference on Machine Learning (ICML), 2003.

All rights reserved by *www.ijsrd.com*