EXTRACTING SEGMENTS

Jyotsna Doonga

Date: 05-10-2022

Abstract

Market segmentation is a marketing term that refers to aggregating prospective buyers into groups or segments with common needs and who respond similarly to a marketing action. Market segmentation enables companies to target different categories of consumers who perceive the full value of certain products and services differently from one another.

Purpose

The purpose of market segmentation is to match the genuine needs and desires of consumers with the offers of suppliers particularly suited to satisfy those needs and desires, benefits consumers and suppliers, and drives an organisation's marketing planning process. Marketing planning is series of activities leading to the setting of marketing objectives and the formulation of plans to achieving them. Two components are a strategic and a tactical marketing plan. Market segmentation lies at the heart of successful marketing (McDonald 2010), tools such as segmentation have the largest impact on marketing decisions (Roberts et al. 2014).

Benefits of Market Segmentation

Effective in sales management because it allows direct sales efforts to be targeted at groups of consumers rather than each consumer individually. Better understanding of differences between consumers, which improves the match of organisational strengths and consumer needs. Micro marketing or hyper-segmentation, finer segmentation where each consumer represents their own market segment.

Analysis of Market Segmentation

A. Layering, process of grouping consumers into naturally existing or artificially created segments of consumers who share similar product preferences or characteristics.

B. Various approaches such as based on Organisational Constraints, Choice of Variables etc

C. Data Structure and Data–Driven Market Segmentation Approaches

Steps of Market Segmentation Analysis

Step 1: Deciding (not) to segment, adv. and disadvantage.
Step 2: specify characteristics of their ideal market segment.
Step 3: data collection.
Step 4: Exploring data.
Step 5: Extracting segments
Step 6: Profiling segments.
Step 7: Describing segments.
Step 8: Selecting (the) target segment(s).
Step 9: Customising the marketing mix.
Step 10: Evaluation and monitoring.

Extracting Segments Brief

Many segmentation methods can be used to extract market segments using clustering algorithms, where cluster signifies the segments. As mentioned by Henning and Liao (2013) selecting a suitable clustering method is required to match the requirement of a user group like researcher. Hence it is required to 4 extract the segments from different clustering algorithms like Hierarchical methods, K-means method, Hybrid approach (the centroids of K-means cluster is user in Hierarchical Method to form dendrogram to decide number of segments) etc can be used and then to decide the analytical-result of which algorithm suits as per requirement of the end user as there is no one single best algorithm for all data set. While deciding the optimum algorithm for extracting the segments we need to consider Data Set characteristics like Nos of consumers, Nos of segmentation variables, scale of segmentation variable etc and Segmentation characteristics like similarities in the segment and differences between the segments, Number and size of segments etc. A helpful solution to decide the best clustering method is to extract the groups using all suitable clustering methods and then perform global stability and segment level stability analysis to decide the best clustering method and segments and possibly pass the segments created by the best clustering algorithm to the next step for profiling segments.

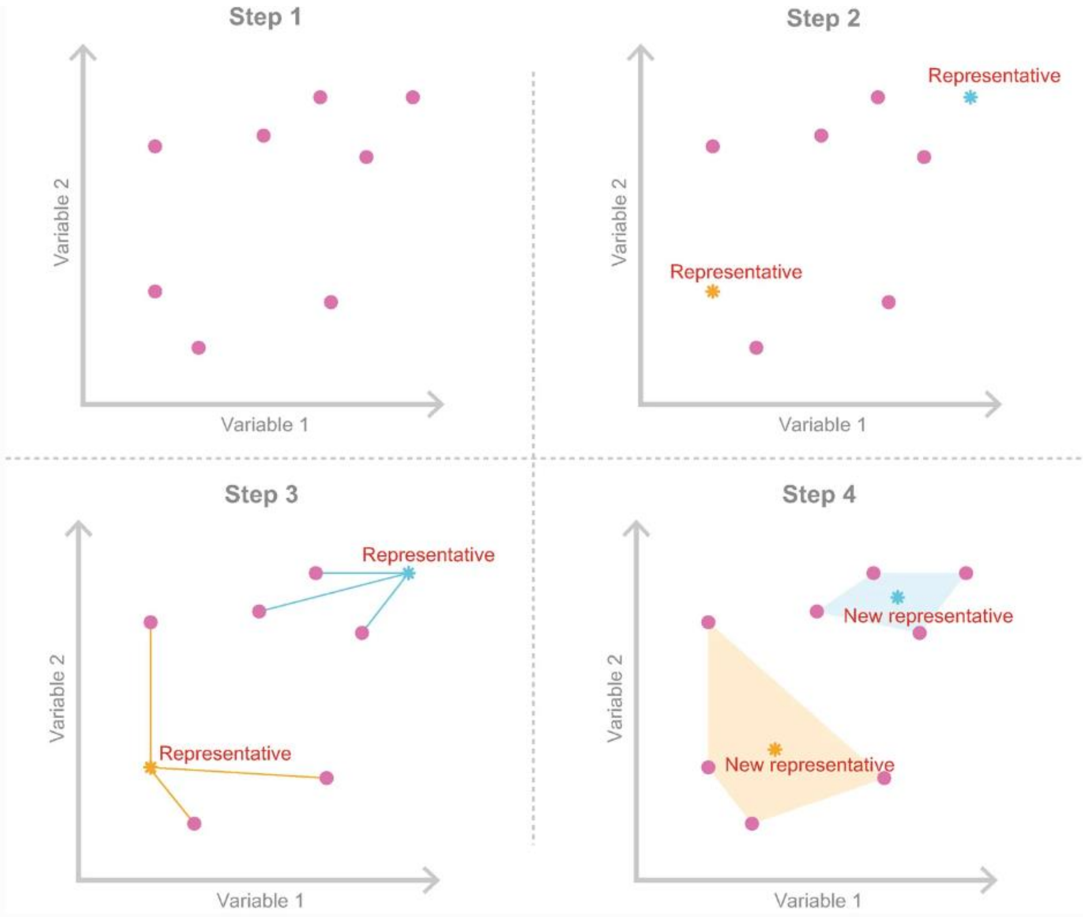
K-Means and K-Centroid Clustering

The most popular partitioning method is *k*-means clustering. Within this method, a number of algorithms are available. R function kmeans() implements the algorithms by Forgy (1965), Hartigan and Wong (1979), Lloyd (1982) and MacQueen (1967). These algorithms use the squared Euclidean distance . A generalisation to other distance measures, also referred to as *k*-centroid clustering, is provided in R package flexclust.

Let X={x1,…,xn}, X={x1,…,xn} be a set of observations (consumers) in a data set. Partitioning clustering methods divide these consumers into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible. The representative of a market segment is referred to in many partitioning clustering algorithms as the centroid . For the *k*-means algorithm based on the squared Euclidean distance , the centroid consists of the column-wise mean values across all members of the market segment. The data set contains observations (consumers) in rows, and variables (behavioural information or answers to survey questions) in columns. The column-wise mean, therefore, is the average response pattern across all segmentation variables for all members of the segment.

The following generic algorithm represents a heuristic for solving the optimisation problem of dividing consumers into a given number of segments such that consumers are similar to their fellow segment members, but dissimilar to members of other segments. This algorithm is iterative; it improves the partition in each step, and is bound to converge, but not necessarily to the global optimum. It involves 5 steps, four are visualized as follows –

1. Specify the desired number of segments *k*.
2. Randomly select *k* observations (consumers) from data set XX (see Step 2 in Fig) and use them as initial set of cluster centroids C={c1,…,ck}C={c1,…,ck}. If five market segments are being extracted, then five consumers are randomly drawn from the data set, and declared the representatives of the five market segments. Of course, these randomly chosen consumers will – at this early stage of the process – not be representing the optimal segmentation solution. They are needed to get the step wise (iterative) partitioning algorithm started.



Simplified visualisation of the *k*-means clustering algorithm

3. Assign each observation xi to the closest cluster centroid (segment representative, see Step 3 in Fig) to form a partition of the data, that is, k market segments S1,…,Sk where

$$Sj=\{x\in X|d(x,cj)\leq d(x,ch),1\leq h\leq k\}$$

This means that each consumer in the data set is assigned to one of the initial segment representatives. This is achieved by calculating the distance between each consumer and each segment representative, and then assigning the consumer to the market segment with the most similar representative. If two segment representatives are equally close, one needs to be randomly selected. The result of this step is an initial – suboptimal – segmentation solution. All consumers in the data set are assigned to a segment. But the segments do not yet comply with the criterion that members of the same segment are as similar as possible, and members of different segments are as dissimilar as possible.

4. Recompute the cluster centroids (segment representatives) by holding cluster membership fixed, and minimising the distance from each consumer to the corresponding cluster centroid (representative see Step 4 in Fig):

$$\mathbf{c}_j = \arg\min_{\mathbf{c}} \sum_{\mathbf{x}\in\mathcal{S}_j} d(\mathbf{x},\mathbf{c})$$

For squared Euclidean distance , the optimal centroids are the cluster-wise means, for Manhattan distance cluster-wise medians, resulting in the so-called *k*-means and *k*-medians procedures, respectively. In less mathematical terms: what happens here is that – acknowledging that the initial segmentation solution is not optimal – better segment representatives need to be identified. This is exactly what is achieved in this step: using the initial segmentation solution, one new representative is “elected” for each of the market segments. When squared Euclidean distance is used, this is done by calculating the average across all segment members, effectively finding the most typical, hypothetical segment members and declaring them to be the new representatives.

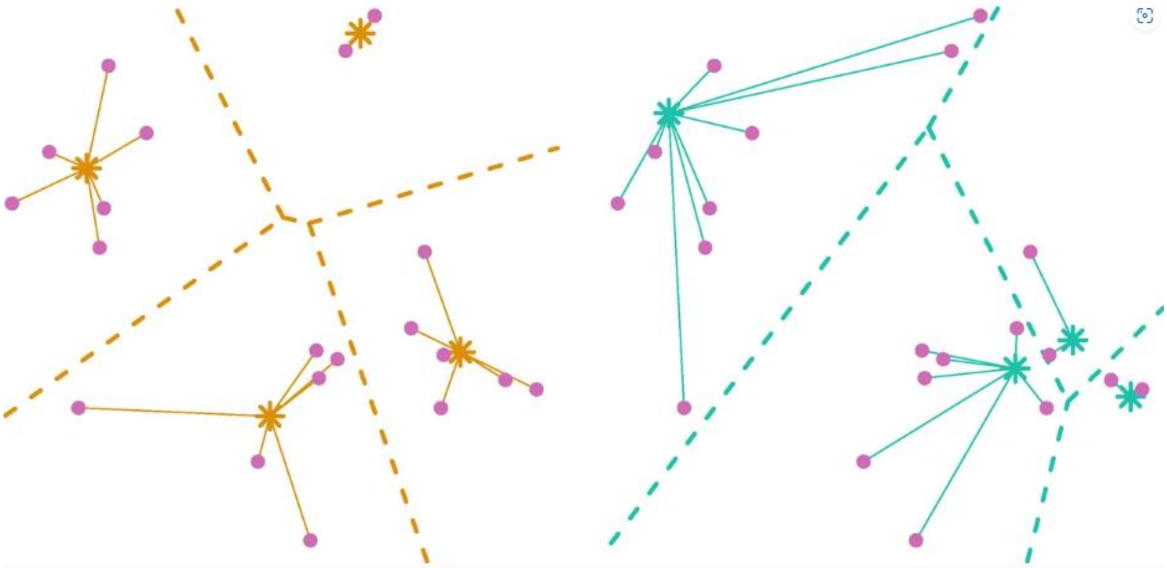
5. Repeat from step 3 until convergence or a pre-specified maximum number of iterations is reached. This means that the steps of assigning consumers to their closest representative, and electing new representatives is repeated until the point is reached where the segment representatives stay the same. This is when the stepwise process of the partitioning algorithm stops and the segmentation solution is declared to be the final one.

The algorithm will always converge: the stepwise process used in a partitioning clustering algorithm will always lead to a solution.

Improved k-Means

Many attempts have been made to refine and improve the *k*-means clustering algorithm. The simplest improvement is to initialise *k*-means using "smart" starting values, rather than randomly drawing *k* consumers from the data set and using them as starting points. Using randomly drawn consumers is suboptimal because it may result in some of those randomly drawn consumers being located very close to one another, and thus not being representative of the data space. Using starting points that are not representative of the data space increases the likelihood of the *k*-means algorithm getting stuck in what is referred to as a *local optimum*. A local optimum is a good solution, but not the best possible solution. One way of avoiding the problem of the algorithm getting stuck in a local optimum is to initialise it using starting points evenly spread across the entire data space. Such starting points better represent the entire data set.

Steinley and Brusco (2007) compare 12 different strategies proposed to initialise the *k*-means algorithm. Based on an extensive simulation study using artificial data sets of known structure, Steinley and Brusco conclude that the best approach is to randomly draw many starting points, and select the best set. The best starting points are those that best represent the data. Good representatives are close to their segment members; the total distance of all segment members to their representatives is small (as illustrated on the left side of Fig below). Bad representatives are far away from their segment members; the total distance of all segment members to their representatives is high (as illustrated on the right side of Fig below).



Examples of good (*left*) and bad (*right*) starting points for *k*-means clustering

Regression Analysis

Regression analysis is the basis of prediction models. Regression analysis assumes that a dependent variable y can be predicted using independent variables or regressors x1,..., xp:

$$y \approx f\,(x1,...,xp).$$

Regression models differ with respect to the function f (·), the distribution assumed for y, and the deviations between y and f (x1,...,xp). The basic regression model is the linear regression model. The linear regression model assumes that function f (·) is linear, and that y follows a normal distribution with mean f (x1,...,xp) and variance $\sigma 2$. The relationship between the dependent variable y and the independent variables x1,...,xp is given by:

$$y = \beta 0 + \beta 1x1 + ... + \beta pxp + ,$$

Binary Logistic Regression.

Binary logistic regression is used to perform a logistic regression on a binary response value. binary variable has only 2 possible values. Binary logistic regression models the relationship between a set of independent variables and a binary dependent variable. It’s useful when the dependent variable is dichotomous in nature, like death or survival, absence or presence, pass or fail and so on. Independent variables can be categorical or continuous, for example, gender, age, income or geographical region. Binary logistic regression models a dependent variable as a logit of p, where p is the probability that the dependent variables take a value of 1. It is used to identify potential buyers of a product. These objectives are based on information such as age, gender, occupation, premium amount, purchase frequency, etc and in all these objectives, the dependent variable is binary, whereas independent variables are categorical or continuous.

We can formulate a regression model for binary data using generalized linear models by assuming that f $(y|\mu)$ is the Bernoulli distribution with success probability μ, and by choosing the logit link that maps the success probability

$$\mu \in (0, 1) \text{ onto } (-\infty,\infty) \text{ by } g(\mu) = \eta = \log\,[\mu/(1-\mu)].$$

Tree Based Methods

Classification and regression trees are an alternative modelling approach for predicting a binary or categorical dependent variable given a set of independent variables. Here a stepwise procedure is used to fit the model and in each step the consumers are split into independent groups. Tree-based methods can be used for both regression and classification. The idea behind these algorithms is to divide the predictor space into multiple non-overlapping regions and to assign new observations to their respective region. These methods are also known as decision tree methods.

After splitting the predictor space into n regions, we calculate a statistic (for instance, the mean) for the response variable in that region. New predictions will be placed into their respective regions and the region's statistic will be used as the prediction.

Decision Trees for classification behaves similarly to regression. Assume that Jim decides to use the same predictors (on a different dataset) to predict whether someone will make an offer on a house. The blue observations corresponds to those that made offers while red corresponds to no offer. The algorithm divides the predictor space into nn regions and each region is mapped to a class. Random Forest ensembles decision trees using bagging. In addition to randomly sampling the training data for each tree, Random Forests also randomly sample the features used for each tree.