Application of Cluster Analysis For Data Driven Market Segmentation

AKOMOLAFE .A. A And ADEBOLA F.B

Department of Statistics, Federal University of Technology, Akure, Ondo state Nigeria
[akomolafe01@yahoo.com][femi_adebola@yahoo.com]

Abstract

Despite the wide variety of techniques available for grouping individuals intomarket segments on the basis of multivariate survey information, clusteringremains the most popular and most widely applied method.Clustering is a popular and widely used method for identifying or constructing data based market segments. Over decades of applying cluster analysis procedures for the purpose of searching for homogenous subgroups among consumers, questionable standards of using the techniques have emerged one of such is the black-box approach ignoring crucial parameters of the algorithm applied or the lack of harmonization of methodology chosen and data conditions. This research work is all out to capture: which standard of application of cluster analysis have emerged in the academic marketing literature, compare their standards of applying the methodological knowledge about clustering procedures and delineate sudden changes in clustering habits. These goals are achieved by systematically reviewing some data-driven segmentation studies that apply cluster analysis for partitioning purposes.

Keywords: Clustering, Black-Box Approach, Data-driven market segmentation and Homogeneous subgroup.

Introduction

The analysis of the consumer behavior is not only a key factor for the success of companies, but also gives a good insight in the way the society in which the consumers live, its orientation and its values. As Solomon et al. mention without understanding the way the consumer feel and think, it is almost impossible for a company to offer the products he wants and in the way he wants. This fact is amplified in the recent years by the fact that the bought products don't represent anymore just some things which cover their needs, but they also describe the way the consumer is. Cluster analysis is a term that refers to a large number of techniques for grouping respondents based on similarity or dissimilarity between each other. Each technique is different; has specific properties, which typically (this is assuming that the data does not contain strong cluster structure) lead to different segmentation solutions. . As Aldenderfer and Blashfield (1984, p.16) say: "Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing." It is therefore very important to carefully select the algorithm that is to be imposed on the data. For instance, hierarchical procedures might not be feasible when large data sets are used due to the high number of distance computations needed in every single step of merging respondents. Market segmentation is a central strategic issue in marketing, these issues depend on the quality of the market segments constructed or identified. Market segmentation is a marketing strategy which involves dividing a broad target market into subsets of consumers, businesses, or countries that have, or are perceived to have, common needs, interests, and priorities, and then designing and implementing strategies to target them. Market segmentation strategies are generally used to identify and further define the target customers, and provide supporting data for marketing plan elements such as positioning to achieve certain marketing plan objectives. Businesses may develop product differentiation strategies, or an undifferentiated approach, involving specific products or product lines depending on the specific demand and attributes of the target segment. This research focuses on clustering techniques exclusively which was first applied by Myers and Tauber (1977) and have since developed to become the major tool for segmentation purposes according to Wedel and Kamakura (1998). Knowing the behavior and the complex motives which drive the consumer to buy products helps not only the producers, but also the retailers to develop their strategies (Dabija, 2011).

The aim of this research is to showcase the common practice for the purpose of market segmentation. The assumption underlying this investigation is that cluster analysis is typically used in a non-explorative manner (a black-box manner with lack of match using data conditions). Organizing data into clusters such that there is high intra-cluster similarity, low inter-cluster similarity or, finding natural groupings among objects. Area of application includes: Sociology, Biology, Psychology, Economics, and Engineering and medical or paramedical sciences.

Source Of Data

The data used is purely a secondary data culled from already prepared document of standard sales outlet in south western part of Nigeria. All the data were analyzed with respect to predefined criteria mirroring the issues known to be most crucial.

Analysis And Results

Among the segmentation studies investigated, the smallest sample size detected contains only 10 elements, the biggest one 20,000 (Table 1). Half of the studies (123, on some of the research work reported on more than one solution) with samples including fewer than 300 objects, data sets smaller than 100 were used by 52 studies (22%). The media sample size amounts to 293.

Table 1. Sample size statistics.

Mean	698
Median	293
Standard deviation	1697
Minimum	10
Maximum	20,000

Table 2. Statistics on the number of variable.

Mean	17
Median	15
Standard deviation	11.48
Minimum	10
Maximum	66

The range varies between 10 and 66 (Table 2). Nearly the thirds of the studies use less than 20 variables as segmentation base. About one fifth uses one to five variables; another fifth bases the segmentation solution on 11 to 15 variables. The number of sample size and the number of variable used is expected to be correlated, as large number of variables (high data dimensionality) requires large data set.

Surprisingly, both Pearson's and spearman's correlation coefficients render insignificant results leading to the inference that even very small sample sizes are used for clustering in very high dimensional attribute space. Due to lack of rules, the only recommendation that can be given concerning size of the ad variables is to critically question if the dimensionality is not too high for the number of cases to be grouped.

Clustering Algorithm

Clustering algorithm can be used in identifying the cancerous data set. Initially we take known samples of cancerous and non-cancerous data set. Label both the samples data set. We then randomly mix both samples and apply different clustering algorithms into the mixed samples data set (this is known as learning phase of clustering algorithm) and accordingly check the result for how many data set we are getting the correct results (since this is known samples we already know the results beforehand) and hence we can calculate the percentage of correct results obtained. Now, for some arbitrary sample data set if we apply the same algorithm we can expect the result to be the same percentage correct as we got during the learning phase of the particular algorithm. On this basis we can search for the best suitable clustering algorithm for our data samples.

It has been found through experiment that cancerous data set gives best results with unsupervised nonlinear clustering algorithms and hence we can conclude the nonlinear nature of the cancerous data set.Varieties of clustering algorithms exist, some of them have restrictions in terms of the maximum number of cases in the data in other to keep calculative feasible e.g, hierarchical approaches (Aldenderfer and Blashfield, 1984), and others are known to structures. More clustering techniques are been developed permanently (e.g. neural networks suggested by Kohonen (1997), Martinetz and Schulten (1994).

Table 3. Percentage explore based on demographic variables.

Variable	Percentage
Needs values	42
Brand loyalty, using of media	20
Age, sex	13
Ordinal manner	14
Dichotomous data	9
Others	2

Table 4. Frequency table of linkage methods (agglomerative hierarchical clustering).

Parameter	Frequency	Percentage
Single linkage	5	6.0
Complete linkage	8	9.6
Average linkage	6	7.2
Nearest centroid sorting	5	6.0
Ward	47	56.6
Not stated	8	9.6
Multiple	4	4.8

methods are nearly balanced (46 to 44%).

In most of the research carried out using hierarchical studies uses ward's method as shown in Table 4. The other techniques like complete linkage dusting, style linkage, clustering, average linkage clustering and nearest centroid sorting do not enjoy this extent of popularity. Among the partitioning algorithms, k-means wins in terms of popularity (76%) (Table 5) sporadically, other types are applied.

Surprisingly, no interaction between data characteristic and algorithm chosen is detected. Although hierarchical methods are limited in data size due to destines computation between all pains of subjects at each step, ANOVA indicates that both sample size (p-value = 0.524) and number of variables (p-value = 0.135) do not influence the choice of algorithm. The average data size for hierarchical studies in 530 and for partitioning studies (927). Specifically, the clustering algorithm should be chosen with the particular data and purpose of analysis in mind.

Number Of Clusters

The number of clusters problem is as old as clustering itself. Clearly the number of clusters chosen a prior most strongly influences the solution, different approaches have been suggested to tackle the problem but no single superior solution has emerged. Nearly one fifth of all the studies do not explain choice of the number of cluster. Half of them used heuristics (like graphs, dendogramms, indices etc.) and approximately one quarter combined subjective opinions with heuristics. Purely subjective assessment accounts for a small proportion only (7%). As far as the number of cluster chosen for the final solution is concerned, descriptive analysis shows a concentration at three (23%), four (22%) and five clusters (19%). Except for the six-cluster-solution, all remaining possibilities do not reach more than 10% (ranged ranging from 2 to 37).

No interrelation with any data attribute is detected. There is no one solution for this problem. Basically, two approaches can be recommended:

1. Repetition of calculations with varying numbers of clusters and evaluation of the results with regard to relevant criterione.g stability.
2. Calculation of solutions with different numbers of clusters and interactive selection with management according to corporate criteria.

Stability/Internal Validity

Assuming that clearly separated clusters exist in the data, stability is no necessary criterion for the quality of the solution; it is most natural by-product with criteria like classification rate (if the true memberships are known) being the target, but typically such density clusters do not exist in empirical data.

Stability thus becomes a major issue in data-driven market segmentation as compared to the prior approach (Myers and Tauber, 1977). Stability has not been examined by 67% of the studies under investigation. Among the studies which did, the split-half-method (15%), analysis of hold-out-samples (4%) replication of clustering using other techniques (5%) were applied most often.

The recommendation is to validate results in as many ways as possible (e.g. by discriminant analysis on background variables and by multiple repetition of the actual clustering procedure with different numbers of clusters and different algorithms).

Conclusion

The assumptions about the use of cluster analysis for the purpose of market segmentation that motivated this review are supported to a high extent. A number of observations advocate the assumptions: (1) the typically non- explorative use of the explorative cluster analysis is mirrored by the fact that single runs of calculations are conducted and interpreted.

In only 5% of the studies, analytic procedures were repeated. (2) Indicators of the use of cluster analysis in a black-box manner include the fact that characteristics of the algorithm are not studied, the number of variables as related to sample size is not questioned critically and data format is ignored when applying measures of association as well as in data pre-processing. (3) Most applications ignored parameters that define any tool within the family of cluster analytic techniques. Using default settings leads to what was addressed as "lack of dependence of data requirements" in the introduction.

The algorithm chosen should depend on data size, the measure of association on data format, and the number of variables included on sample size etc. instead of critically choosing the building components of the cluster analytic tool applied, most studies are based on ward's hierarchical clustering or the k-means partitioning algorithm both using Euclidean Sdistance. Implications for data-based marketing research are obvious: the application of cluster analytic procedures for the purpose of data-driven segmentation studies should become more careful in the setting of parameters in order to substantially improve the quality of clustering outcome and reduce the proportion of "random results" which are interpreted in detail and misunderstood as best representation of the data in reduced space.

Researchers have to be aware of the fact, that cluster analytic techniques always render a result. This neither means that it is the only possible way of splitting customers into groups nor that the result is of any practical use to a company. Thus, thorough understanding of the procedures, Careful harmonization of algorithm and the data at hand and finally transparent reporting on the application of cluster analysis for segmentation is required to improve the quality of the application of this technique for the purpose of data-driven market segmentation.

Reccommendation For Further Research

Future contributions to the field of market segmentation by means of cluster analysis embrace all improvements in the methodology that supports researchers in optimizing the crucial decisions: choice of algorithm, number of clusters, algorithm parameters, optimal ratio of variables to sample size etc, for the time being the best way of dealing with these issues is to critically question each step and transparently report on the results to ease the interpretation of the value of a particular segmentation solution. Future contributions to the field of market segmentation by means of cluster analysis embrace all improvements in the methodology that supports researchers in optimizing the crucial decisions: choice of algorithm, number of clusters, algorithm parametersand optimal ratio of variables to sample size etc.

References

Aldenderfer MS, Blashfield RK (1984). Cluster analysis series on quantitative applications in the social sciences. Beverly hills: Sage Publications.

Arabie P, Hubert LJ (1994). Cluster analysis in marketing research. In advanced methods in marketing research. Ed. R. P. Bagozzi. Blackwell: Oxford, pp.160 – 189.

Dolni-ar, S. and Leisch, F. (2000). Behavioral Market Segmentation Using the Bagged Clustering Approach Based on Binary Guest Survey Data: Exploring and Visualizing Unobserved Heterogeneity.*Tourism Analysis*, 5(2-4), 163-170.

Dolni-ar, S. and Leisch, F. (2001). Knowing What You Get - a Conceptual Clustering Framework for Increased Transparency of Market Segmentation Studies. Paper presented at the Marketing Science, Edmonton, Canada.

Dolni-ar, S. and Leisch, F. (2003). Winter Tourist Segments in Austria – Identifying Stable Vacation Styles for Target Marketing Action.*Journal of Travel Research*, 41(3), 281-193.

Formann, A.K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung*. Weinheim: Beltz.

Frank, R. E., Massy, W. F. and Wind, Y. (1972).*Market Segmentation*. Englewood Cliffs: Prentice-Hall.

Haley, R. J. (1968). Benefit Segmentation: A Decision-Oriented Research Tool. *Journal of Marketing*, 32, 30-35.

Ketchen DJ jr, Shook CL (1996). The application of cluster analysis in strategic management research: an analysis and critique. Strategic management journal17(6):441-458.

Kohonen T (1997). Self-organizing Maps, 2nd Edition. Berlin: Springer. 26-33

Ketchen D.J. jr. and Shook, C.L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal,* 17(6), 441-458.

Leisch, F. (1998).*Ensemble methods for neural clustering and classification*. Dissertation.Technical University of Vienna.

Leisch, F. (1999).Bagged Clustering. Working Paper # 51, SFB ``Adaptive Information Systems and Modeling in Economics and Management Science'', http://www.wuwien.ac.at/am.

Mazanec, J. A. (1997). Segmenting city tourists into vacation styles. In K. Grabler, G. Maier, J. Mazanec& K. Wober (Eds.), *International City Tourism: Analysis and Strategy* (pp.114-128). London: Pinter / Cassell.

Mazanec, J. A. (2000). Market Segmentation.In J. Jafari (Ed.), *Encyclopedia of Tourism.* London: Routledge.

Mazanec, J. and Strasser, H. (2000).*A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*. Berlin: Springer.

Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in data sets. *Psychometrika*, 50, 159-179.

Milligan, G.W. (1981). A montecarlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2), 187-199.

Myers, J.H. and Tauber, E. (1977).*Market structure analysis*. Chicago: American Marketing Association.

Leisch F (1998). Ensemble methods for neural clustering and classification.Discertation. Technical University of Vienna

Leisch F (1999). Bagged clustering. Working paper #51, SFB "adaptive information systems and modeling in economics and management science", http://www.wuwien,ac.at/am

Martinetz T, Schulten K (1994). Topology representing networks. Neural Networks 7: 507-522.

Myers JH, Tauber E (1977). Market structure analysis. Chicago: American Marketing Association. 38-49.

Wedel M, Kamakura W (1998). Market segmentation – conceptual and methodological foundations. Boston: Kluwer academic publishers. Vol 5:203-215

Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika*, 18(4), 267-276.

Wedel, M. and Kamakura, W. (1998). *Market Segmentation - Conceptual and*

*Methodological Foundations.*Boston: Kluwer Academic Publishers

Figure 8b: This thermal image shows a thermal bridging of a high-rise building

.