Data Segmentation and Clustering

Alankrita Khadtare
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x17126339@student.ncirl.ie

Abstract*— Data clustering and segmentation plays an important role in marketing and is quite new. In today's competitive markets, organizations need to have a complete view of their customers to gain a competitive advantage. They need to focus on their customers', needs, wants, attitudes, behaviors, preferences, and perceptions, and to analyse relevant data to identify the underlying segments. The identification of groups with unique characteristics will enable the organization to manage and target them more effectively with, among other things, customized product offerings and promotions. The goal of this research paper is to provide a comprehensive review about cluster analysis and segmentation, relation between customer segmentation and clustering and challenges in marketing.

Keywords*— Cluster Analysis, Segmentation, Customer Segmentation and Clustering, Marketing

I. INTRODUCTION

In Data Analytics, we often have very large data which are similar to each other. We may want to organize this data in a few clusters with similar observations within each cluster. For example, in the case of a customer data, not every customer will respond to every product offering nor will every product be right for all customers. This provides a challenge for the development and marketing of profitable products and services. Segmentation is a way of organizing customers into groups with similar perceptions, product preferences or expectations. Marketing strategies and products can be customized once segments are identified. The better the segment chosen for targeting by an organization, the more successful the organization is assumed to be in the marketplace.

Depending on "similarities" and "differences" between data observations, segments are constructed based on customers' (1) socio-demographic characteristics, (2) need/attitudinal (3) benefits from products/services, and (4) behavioral value. These days most firms possess rich information about all customers unlike the traditional survey-based market research about few customers which provide information about transactions and billing histories. Customer information can be surveyed as the basis for segmentation in case where firms do not have access to detailed information.

II. WHAT IS CLUSTERING?

Clustering is a division of data into groups of similar objects. Each of these groups, contain objects that are similar or dissimilar to objects of other groups called cluster. Although clustering data loses certain fine details due to lossy compression, but achieves simplification. Data objects are represented by a few clusters, and hence, could be referred to as data modeling. Data modeling emphasizes clustering in a historical perspective rooted in mathematics and statistical analysis. Clustering is an unsupervised learning that searches for clusters where they correspond to hidden patterns and the resulting system represents a data concept. Therefore, clustering is referred to as an unsupervised learning of a hidden data concept, from a machine learning perspective. Clustering can be performed on different types of data like (1) image data [8], (2) time series data [9] and (3) relational data [10].

III. CLUSTER ANALYSIS

Cluster analysis divides data into groups that are meaningful, useful or both [5]. Whether for understanding or utility, cluster analysis plays an important role in a wide variety of fields like psychology, statistics, pattern recognition, machine learning and data mining. Cluster analysis makes no distinction between dependent and independent variables. Interdependent relationships are examined on the entire set. For example, cluster analysis sorts through the raw data on customers using segmentation and groups them into clusters. A cluster is a group of relatively homogeneous customers. Customers who belong to the same cluster are similar to each other. They are dissimilar to customers outside the cluster, particularly customers in other clusters. The primary input for cluster analysis is a measure of similarity between customers, such as (a) distance measures, (b) correlation coefficients, and (c) association coefficients.

The following are the basic steps involved in cluster analysis:

1. Formulate the problem- select the segmentation variables that you wish to use as the basis for clustering.
2. Compute distance between customers along the selected segmented variables and visualize pairwise distances.
3. Apply the clustering procedure to the distance measures.
4. Apply appropriate method and manipulate number of segments.
5. Profile and interpret the segments.
6. Robustness analysis.

IV. CUSTOMER SEGMENTATION AND CLUSTERING

Customer segmentation is the process of dividing customers into distinct, meaningful, and homogeneous subgroups based on various attributes and characteristics. It enables organizations to understand their customers and build differentiated strategies [2]. The identification of groups with unique characteristics such as customer's needs, wants, attitudes, behaviours, preferences and perceptions will enable the organisation to manage and target them more effectively on product offerings and promotions. There are many different segmentation types based on the specific criteria or attributes used for segmentation. The type of segmentation used depends on the specific business objective [1].

Customers can be segmented according to their value, socio-demographic and life-stage information, and their behavioural, need/attitudinal, and loyalty characteristics. In behavioural segmentation, customers are grouped by behavioural and usage characteristics. One way to find behavioural segments is to use the clustering techniques described in [3]. These methods lead to clusters of similar customers but it may be hard to understand how these clusters relate to the business. Segments are built with respect to a marketing goal such as subscription renewal or high spending levels. Decision tree techniques described in [4] are ideal for this sort of segmentation.

V. CLUSTERING TECHNIQUES FOR SEGMENTATION

There are many statistical methods for clustering and segmentation [6][7]. However, in our research paper, we will use two widely used methods: the K-means Clustering Method, and the Hierarchical Clustering Method. Both methods are based on how we measure the distance or similarity between our observations. K-means differs from Hierarchical clustering in a way that the former requires the user to define how many segments need to be created, while Hierarchical clustering does not.

K-means Clustering

K-means clustering algorithm is an iterative clustering approach which aims to find local data points in each iteration. This algorithm works in 5 steps:

1. Choose the desired number of clusters k. For example, k=2 for 5 beads.
2. Assign each of these beads to a single cluster. Let's assign three beads in cluster 1 shown using red colour and two beads in cluster 2 shown using grey.
3. Compute cluster centroids. The centroid of beads in the red cluster is shown using red cross and those in grey cluster using grey cross.

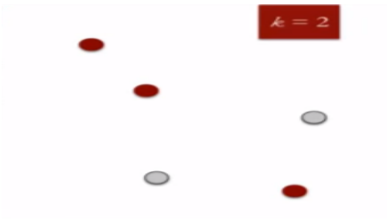

Fig.1

4. Re-assign each bead to the closest cluster centroid.
5. Re-compute cluster centroids.

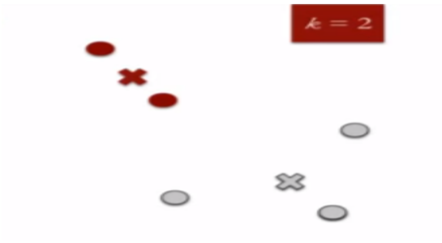

Fig.2

6. Repeat steps 4 and 5 until no clustering is possible.

Hierarchical Clustering

Hierarchical clustering is nothing but building hierarchy of clusters. This algorithm starts with assigning each of these beads to a cluster of their own. Clusters which are nearest are then merged into the same cluster. The algorithm terminates with only a single cluster left at the end.

The results of hierarchical clustering using dendrogram can be illustrated as follows:

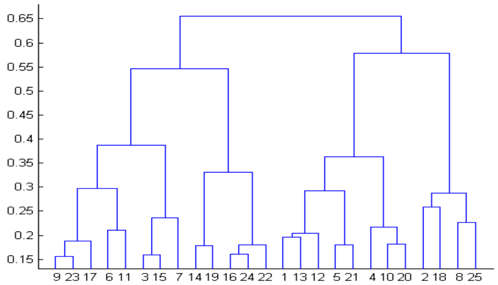

Fig.3

Two important things for hierarchical clustering are as follows:

1. This algorithm can be implemented using two approaches— top down approach and bottom up approach. The above example is implemented using bottom-up approach. Top-down approach starts with all beads assigned in the same cluster and splitting each cluster recursively.
2. The closeness of clusters can be used to decide the merging of two clusters. The closeness of clusters can be calculated by using the following metrics.
   - Euclidean distance: $|a\text{-}b|_2 = \sqrt{}(\Sigma(a_i\text{-}b_i))$
   - Squared Euclidean Distance: $|a\text{-}b|_2{}^2 = \Sigma((a_i\text{-}bi)^2)$
   - Manhattan distance: $|a\text{-}b|_1 = \Sigma\,|a_i\text{-}b_i|$

VI. CHALLENGES IN MARKETING SURVEY

Customers should understand identification of new marketing strategies. The design and development of new products/ services should cater to each segment's characteristics. One of the biggest challenge is to design product offering strategies to existing customers according to identified wants and needs. New strategies should offer tailored rewards and incentives. One must know appropriate advertising and communication channel. There should be more effective resource allocation according to the potential return from each segment.

TABLE I. COMPARISON FOR SEGMENTATION USING CLUSTERING TECHNIQUES

Author/ Year of Publication	Paper Title	Implementation Method/ Algorithm	Application
Anestis Fachantidis, Athanasios Tsiaras, Grigorios Tsoumakas, Ioannis Vlahavas, 2016	"Segmento: An R-based Visualization-rich System for Customer Segmenta	Segmento is implemented as a web application hosted in a typical web server running the Shiny Server developed using R, SQL and R Shiny Web framework	Marketing analytics

	tion and Targeting”		
YoungSung Cho, Seon-Phil Jeong, 2015	“A Recommender System in u-Commerce based on a Segmentation Method”	Cluster Analysis of Segmented merchandise to have different weights based on FRAT (Frequency, Recency, Amount and Type) method	U-Commerce application
Kishana R. Kashwan, 2013	“Customer Segmentation Using Clustering and Data Mining Techniques”	A k-means clustering techniques was used to choose initial cluster centers and then final stable clusters were computed by continuing number of iterations	Marketing Survey
M. Kumar, 2002	“Clustering seasonality patterns in the presence of errors”	Based on the assumed independent Gaussian model of data errors. Agglomerative Hierarchical	Seasonality pattern in retails
Mengfan Zhang, 2010	“Application of computational verb theory to analysis of stock market data”	CVT (Computational Verb Theory) K-means	Stock Market data
Jian Xin Wu, 2007	“Combining ICA with SVR for prediction of time series”	Independent component analysis. Support Vector Regression	Financial Time Series

CONCLUSION

Clustering lies at the heart of data analysis and data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and interrelationships change. Clustering is a most popular way to identify segments not known in advance and split customers into groups that are not previously defined. The main objective of customer segmentation is to understand the customer base and gain customer insight that will enable the design and development of differentiated marketing strategies. The identification of the segments is followed by profiling the revealed customer groupings based on common characteristics. The criteria used to divide customers (behavioral, demographic, value or loyalty information, needs/attitudinal data) define the segmentation type. Like any data analysis, segmentation undergoes an iterative process with many variations of data, methods, number of clusters and profiles generated until a satisfying solution is reached. The success of any segmentation process requires managerial intuition and careful judgement to satisfy necessary needs with the goals and core competencies of the firm.

Computing based system developed these days is an intelligent approach and it automatically presents results to the managers to infer for quick and fast decision-making process. The future work will involve more trial and automation of the market forecasting and planning.

REFERENCES

[1] Tsiptsis, K.K. and Chorianopoulos, A., 2011. Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.

[2] Armstrong, G. and Kotler, P., 2005. Marketing: an introduction. Prentice Hall.

[3] Tan, P.N., Steinbach, M. and Kumar, V., 2005. Introduction to data mining. 1st.

[4] Rokach, L. and Maimon, O., 2014. Data mining with decision trees: theory and applications. World scientific.

[5] Rai, P. and Singh, S., 2010. A survey of clustering techniques. International Journal of Computer Applications, 7(12), pp.1-5.

[6] Fachantidis, A., Tsiaras, A., Tsoumakas, G. and Vlahavas, I., 2016, May. Segmento: An R-based Visualization-rich System for Customer Segmentation and Targeting. In Proceedings of the 9th Hellenic Conference on Artificial Intelligence (p. 23). ACM.

[7] Cho, Y. and Jeong, S.P., 2015, October. A Recommender System in u-Commerce based on a Segmentation Method. In Proceedings of the 2015 International Conference on Big Data Applications and Services (pp. 148-150). ACM.

[8] Bhatia, S.K., 2005, May. Hierarchical clustering for image databases. In Electro Information Technology, 2005 IEEE International Conference on (pp. 6-pp). IEEE.

[9] Rani, S. and Sikka, G., 2012. Recent techniques of clustering of time series data: a survey. International Journal of Computer Applications, 52(15).

[10] Rai, P. and Singh, S., 2010. A survey of clustering techniques. International Journal of Computer Applications, 7(12), pp.1-5.