Methods for Market Segmentation

Shiqi Wang

University of Florida

shiqi.wang@ufl.edu

Abstract. In business analysis, market segmentation is a technique used to classify existing and potential customers based on their similarities, which lays the foundation for the company to maximize their profits. The market can be divided into five subgroups, including demographic factors, geographic factors, psychographic factors, customers' benefit, and behavioral factors. In order to build appropriate market segmentation, several approaches could be applied in the marketing process. Specifically in this article, we will cover K-means clustering through partitions and Principal Component Analysis through color differentiation. For these two methods, each has its own strengths and weaknesses. The benefits of K-means clustering can be reflected by its simplicity, flexibility, and adaptivity in a large dataset, but its defects are also straightforward- manually choosing k values, dependent on the initial values and the chosen k value, and sensitive to outliers. PCA enables people to better visualize and reduce overfitting while it makes independent variables become less interpretable and loss some data.

Keywords: Market segmentation, K-means clustering, Principal Component Analysis.

1 Introduction

Market segmentation is a broad subject that many scholars have studies and researches on this. Some scholars focus on the desirability of potential market segments, specifically the measurability, accessibility, substantiality and actionability and how to satisfy the market's needs in terms of these four aspects [1]. There are also studies on realistic problems after performing market segmentation, including whether it is a wise idea to segment the market or how people could do to lower the chance of failure [2]. In contrast to previous scholars' studies, this article emphasizes the methods of finding groups which have not been explicitly labeled in the data through K-means clustering and simplifying the complex, high-dimensional data with the intrinsic trends and pattems through Principal Component Analysis and their wide applications in the practical life.

H. Mallick et al. (Eds.): ICEMCI 2022, AEBMR 231, pp. 1614–1621, 2023.
https://doi.org/10.2991/978-94-6463-098-5_182

2 The Brief Information

There is no company that could meet the needs of the entire market with its own human resources, material resources, and financial resources. Then, the importance of customer market segmentation could be reflected. The process of market segmentation categorizes the customers according to certain organizational standards, in order to differentiate advertisements for different groups, which reduces corporate risks to a certain extent.

Generally, there are three types of categories that distinguish the customers. The first type is to classify the customers based on their characteristics, specifically locations, income, professions, education level, religion, personality, attitudes, and so forth. The fundamental of this category is the customers' demands based on the society and their economic backgrounds. The other type of category is customer value, namely the values and profits created for the company. This classification method usually covers customers' influence, loyalty, and profit margin. The third type of classification focuses on the customers' shared needs. Different types of customers have different demands; it would be the most efficient and effective if the company provides targeted products and services. The purpose of the market segmentation is the save time and effort in the process of marketing, and thus simplify the organizational management. Also, by categorizing customers, the organization will be able to study the needs of potential customers and then make them the new customers.

The process of the market segmentation would be relatively simple. First, the company researches the size of the target market. In this step, the company should identify the types of customers' needs and their expectations, which will then help the company with their marketing strategies. The next step is to set up the expectations. This step is crucial since it builds up a connection between the customers and the company so that they are all on the same side—making or finding the most suitable products. Then the company should distinguish categories and subcategories. During this process, the company will further collect the information of customers and the market on the basis of customers' diverse needs and make some general market segmentations and the subsegmentations. Next, the company researches the consumption customs and preferences of customers in different social levels. According to the effective market segmentation, the company should conduct research on all market segmentations and eliminate those useless and unsatisfactory categories, and study the consumption habits based on their previous purchases. The final step would then be the plan-making step. After completing all the process of building up market segmentation, it is important to finalize the marketing strategies according to different categories of customers and their characteristics.

3 Methods

3.1 K-means clustering

K-means clustering algorithm is an iterative algorithm of clustering analysis, aiming at approximating the target or desired results. The iterative algorithm refers to the process of repetitively composing with itself, through which each object will be recalculated several times until no object will be reassigned to a new cluster and reach the minimal sum of squared errors. Specifically, people only need several steps. First of all, select the initial k samples as the initial cluster centers. Based on these centers, we calculate the distances from all points to the centers and cluster them in order that each point is assigned to the clusters with minimal distances. Then we need to recalculate the new centroid for all sample points and repeat the above steps generally until the objective function reaches the optimal or reaches the maximum number of iterations to terminate. Clearly, K-means clustering is easy to understand and can reach the local optimal, which would be sufficient for daily use. Especially for processing large datasets, the scalability of K-means clustering can be well reflected; when the clustering is approximately Gaussian distribution, the effect is significant [3]. However, manually set K values lead to different results, and the selection method matters. Also, this method is sensitive to outliers and the initial cluster centers, meaning that different people would probably obtain diverse results.

As shown in the figure below, this is the data for a sample of 200 customers and we are interested in customers' spending scores with respect to certain characteristics. Data in the Figure 1 shows that the age of the participants ranges from 18 to 70, forming a positively skewed distribution [3]. Under such distribution, the mean is greater than the median, meaning that the overall age of the participants is young; the spending score is inclined to be a normal distribution, with approximately the same the mean and the median value

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Fig. 1. Data Collection for Customers with Respect to Age, Annual Income, and Spending Score (Source from https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis)

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Fig. 2. Data Collection for Customers with Respect to Gender, Age, Annual Income, and Spending Score (Source from https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis)

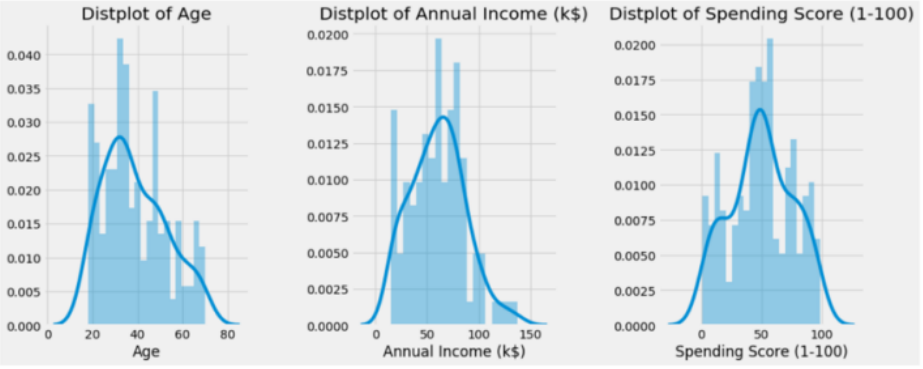

Fig. 3. Inertia for Age, Annual Income, and Spending Score (Source from https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis)

Now, we need to identify the spending scores with K-means clustering. We start by introducing a new concept: inertia. Inertia measures the distance of the points to their centers; thus, the smaller the inertia it is, the better the sample data we have.

With the data, we could obtain a graph of inertia and the number of clusters.

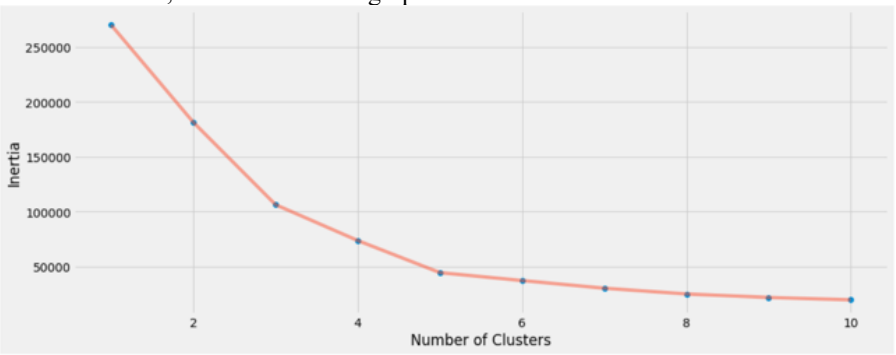

Fig. 4. The Curve Diagram between Inertia and the Number of Cluster (Source from https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis)

Clearly, we see that the value of inertia decreases as the number of clusters increases. Hence, increases in clustering centers could reflect a precision in the classification.

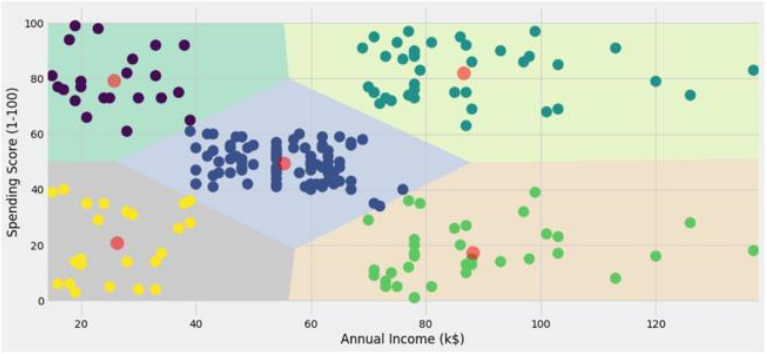

Fig. 5. The Result of K-Clustering (Source from https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis)

Then, through the repetitive processes of K-means clustering, we will obtain a final result with 5 cluster centers and points surrounding the center. It is believed that in the five clusters, the closer the distance between the two targets, the greater the similarity. The graph may change if we choose a different K value.

3.2 Principal Component Analysis

Principal component analysis is one of the most common unsupervised machine learning algorithms, which contains the use of artificial intelligence to identify and react to the generalities in the dataset. This method converts the data from the original space to a new feature space through a certain linear projection, in order to reduce the data dimensions while retaining the most characteristics of the original dataset. Simply speaking, we need the variance to be as large as possible and the dataset to be more scattered,

resulting in a larger information base. In order to perform a principal component analysis, we need first to calculate the mean values and find the covariance matrix. The goal of this step is to obtain the eigenvalue and the eigenvector, which is crucial in finding out the largest eigenvalue and thus realizing the dimensionality reduction by transforming the data matrix into a brand-new space.

To better illustrate the method of principal component analysis, we will look at an example about automobile customers. We collected some information with respect to "Gender", "Married", "Graduated", "Profession", "WorkExperience", "SpendingScore", "FamilySize", "Category", and "Segmentation". The below bar graph visually illustrates the numerical relationships.

Since there are nine characteristics in this example, it would be difficult to present with only one picture; thus, we use colors to distinguish them.

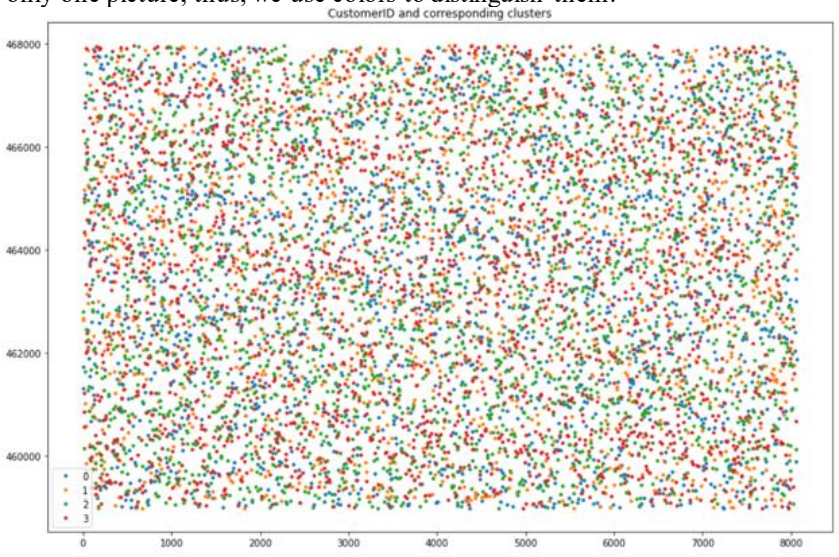

Fig. 6. Colored Graph with Various Customer IDs (Source from https://www.kaggle.com/code/maricinnamon/automobile-customer-clustering-k-means-pca)

This is the graph with customer IDs and the corresponding clusters. We can see that these points are messy and we have no way to extract useful information from this picture. Hence, we could try to reduce dimensions by applying principal component analysis in order to visualize the data.

After constructing a covariance matrix and calculating the eigenvalues and eigenvectors, we obtain the graph as shown below. Now, the graph information is obvious when we reduce the dimension to two.

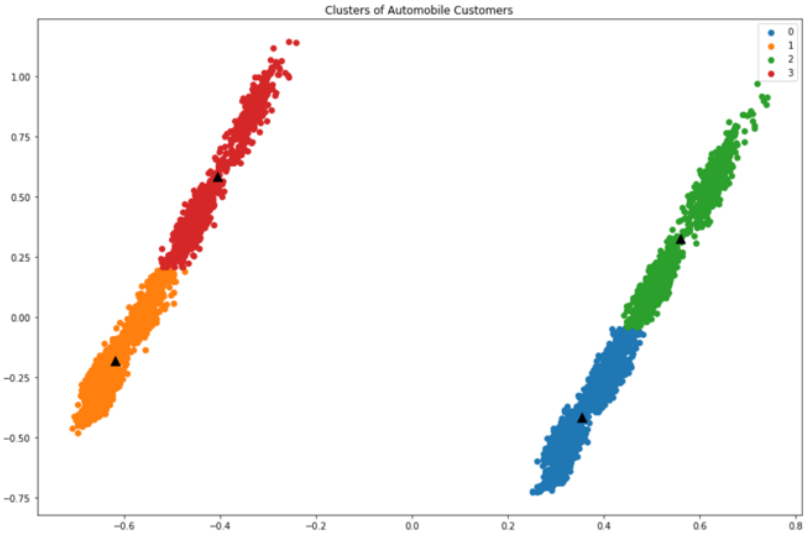

Fig. 7. The Graph After Segmentation (Source from https://www.kaggle.com/code/maricinnamon/automobile-customer-clustering-k-means-pca)

4 Application

As we have described above, K-means clustering is a method of classifying data into clusters based on data similarity in modelling; therefore, we could discover its applications everywhere in our daily lives. Some scholars have found that K-means clustering can be used to monitor students' academic performances. Based on K-means clustering algorithm, researchers are able to arrange students' scores using standard statistical algorithm. This method enables teachers and academic institutions to learn about students' learning behaviors and academic results for making better and more effective decisions [4]. In addition, in order to promote and popularize products into new market while minimizing the cost, data mining could be used during this marketing process. Especially through the application of K-means clustering, people with similar characteristics will then be classified into the same cluster; thus, it is expected that the marketing department could market with the right promotional strategy to gain the prospective target [5]. Similarly, principal component analysis also has a wide variety of applications. Some researchers have presented a novel PCA classifier, which determines the principal components of each class and builds the classifier by projecting the data onto the subspaces spanned by these principal components that correspond to the various classes. In the practical life, this PCA classifier can be applied in the vehicle recognition and detection [6]. Apart from the PCA classifier, PCA can also be used to monitor the air quality monitoring network. A newly developed sensor fault detection and isolation procedure makes some improvements in reconstructing error criteria; a new detection index could perform sensor fault detection among different residual subspaces. This

reconstruction techniques enables both the isolation of defective sensors and the estimation of the default amplitude [7].

5 Conclusion

In this article, we explain two methods that could be used in customer segmentations in details. The results of K-means clustering depend significantly on the value of K and outliers; also, the result may only be a local optimum instead of a global optimum. However, it is also its advantage. When we tune parameters, only one parameter needs to be changed. That is one of the reasons K-means clustering is widely used in the industry. As for the principal component analysis, there are some limitations. The transformation matrix must be a square matrix in order to obtain the eigenvalues and eigenvectors; in the case of the non-Gaussian distribution, the result may also be not the optimal. Compared with the K-means clustering, PCA does not depend on parameters, which in other words, makes the dataset easier to use due to its uniqueness [8].

References

1. Tynan, A. C., & Drayton, J. (n.d.). (2022, May 7) Market segmentation. Taylor & Francis.
2. McDonald, M., Christopher, M., & Bass, M. (1970, January 1). Market segmentation. SpringerLink.
3. Wu Junjie. 2012. Advances in K-means Clustering. Springer Nature Book.
4. Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010, February 11). Application of K means clustering algorithm for prediction of students academic performance. arXiv.org.
5. Application model of K-means clustering: Insights into promotion strategy of vocational high school. (2018, February 27).
6. Junwen Wu, Xuegong Zhang (2002, August 7). A PCA classifier and its application in vehicle detection. IEEE Xplore. (n.d.).
7. Harkat, M.-F., Mourot, G., & Ragot, J. (2006, July). An improved PCA scheme for sensor FDI: Application to an air quality monitoring network.
8. Tsai, J., Ng, L. 2017. Computational Methods with Applications in Bioinformatics Analysis. World Scientific.