IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation using K- Means in Python on MNIST Dataset

Kapil Nalawade[1], Ritik Agrawal[2], Madhav Gupta[3], Krishna Agrawal[4] and Vipin Wani[5]

[1-4]School of Computer Science and Engineering, Sandip University, India.

[5]Assitant Professor, School of Computer Science and Engineering, Sandip University, India

Abstract

This project is about separation of the customers
The separation is to get a targeting audience and devise a marketing strategy.. Enterprises separate their customers before running any campaign. Their main aim is to get a targeted audience as per their product and their marketing strategy. The main characteristics required in separating the customers are Age, Gender, Marital Status, Location(Rural Areas, Urban Areas, Suburban Areas), Life Stage, their employment and source of income. We are going to use K - Means Algorithm to separate our customer from our Database. Our main aim is to separate our customers into classes that share the most similar human traits. They are most likely to buy a similar product.

Customer segmentation is the support that can be offered to the company both before and after sale and use of any of the company products or services. These help them to communicate in a simple and easy way with their team to increase sales.. For this project we have used a dataset "mall". The dataset consists of parameters like Age, Gender, Annual Income, Spending score. All of this features are used to predict the pattern and generate customer segmentation model

1.INTRODUCTION

Customer Segmentation (which is also known as Market Segmentation) is the procedure to Divide customers into groups which share relative Human behaviour and traits. They are most likely to share their preferences for the same range of products.

Let us take an example of a pet store: The customers for them are very discrete, with each in different needs and preferences. Their requirements are highly dependent on the type of pets they have, their pets' breed, dietary needs, and their ages. Their owners lifestyle, their income and their relation towards their pet.

From a marketing point of view, it won't be feasible for store suppliers to communicate with their customers in the same way.

But customer segmentation is about more than separating customers with appropriate product offers. It's like changing the way you communicate with your customers based on what you know about them. It's about identifying your most profitable customers and tailoring your products and services to meet their specific needs.

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).

Select the best communication channel for the segment, which might be email, social media posts, radio advertising, or another approach, depending on the segment.

Identify ways to modify products or new product or service opportunities.

- Establish better customer relationships.
- Test pricing options.
- Focus on the most profitable customers.
- Improve customer service.
- Upsell and cross sell other products and services.

2. Literature Review

2.1 Customer Segmentation

For many years, because it's a very, very strong competition in the business world organizations will have up to an increase of the number of employees at the business world Your earnings to comply with the requirements of your customers and attract new customers, according to the to meet their needs. The identification of the customers, and it can satisfy the needs of every customer is a complex and difficult task. This is because the customers can be different depending on their needs, their tastes and preferences, and so on. Instead of a "one-size-fits-all" approach segment customer groups in groups, have the same characteristics, or behaviour. Consumer segmentation is a strategy of dividing the market into homogeneous groups. The data that may be used in the customer segmentation techniques, which distributes it to customers in the the groups will vary depending on a number of factors, such as the by

geographical details, terms, conditions, economic conditions, demographics, circumstances, and behavior problems patterns. Customer segmentation is a method that makes it possible for companies to use it for marketing purposes. Budgets can be more effective and gain an advantage over competing companies with a greater level of understanding the needs of the customer.

2.2 Clustering and K-Means Algorithm

Clustering algorithms generate clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.

The K-means algorithm is one of the most popular centroid based algorithms. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, C1,...,Ck , that is, Ci ⊂ D and Ci ∩Cj = ∅ for ($1 \leq i, j \leq k$). A centroid-based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object p ∈ Ci and ci , the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y.

Algorithm: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. Input: k: the number of clusters, D: a data set containing n objects. Output: A set of k clusters. Method: (1) arbitrarily choose k objects from D as the initial cluster centers; (2) repeat (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

2.3 Keras:

Keras library is the open-source library which provides an interface for artificial neural networks.

Keras library is used to implement neural networks, provides some tools and by using those tools you can work with images, text data, writing code for neural networks etc.

2.4 Pandas:

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with structured and time series. The main aim of using Pandas is for, real world data analysis in Python Programming language.

2.5 NumPy:

NumPy is the array-processing package. The main aim of Numpy is to provide the multidimensional array object with high-performance, and also provides tools to work with these arrays.

In past years, many clustering algorithms for big data have been proposed and accepted from which are based on distributed and parallel computation. Mac Queen in 1967 first estimated this technique.

According to Y. S. Thakare in recent years concluded the performance of k-means algorithm which is estimated with different databases such as Iris, Wine, Vowel, Ionosphere and Crude oil data Set and various distance metrics. The conclusion was achieved that performance of K-means clustering is dependent on the database used and the distance metrics.

This colloquy is on the basis of performance evaluation of the efficiency, flexibility and clustering output applying these algorithms. The conclusion of this comparative study is that FCM estimates a closer result to the K-means but due to its computation time is more than k-means due to involvement of the fuzzy measure calculations.

3. Methodology

3.1.1 Pre-Processing:

Data will be cleaned, transformed, reduced before performing segmentation. Pre-processing gives more - efficiency to our results.

3.1.2 Data Cleaning:

Data will have missing entries, not relevant data. Thus it is important to fill the missing entries and clear noisy data. The missing value are removed by ignoring the tuples that have a large number of parameters missing or filling the missing parameters with a default value.

3.2.2 Data Transformation:

Data transformation will be by normalization, attribute selection, discretization, concept hierarchy generation

3.3.1 K-Means Algorithm:

The algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the Euclidean distance as measurement.

3.3.2 Initialization of K points:

First we initialize k points, called means, randomly. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and at the end, we have our clusters.

3.2 Elbow Point:

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The optimal number of clusters can be defined as follow:

- Compute clustering algorithm (e.g., K -Means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the total within the cluster sum of squares (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

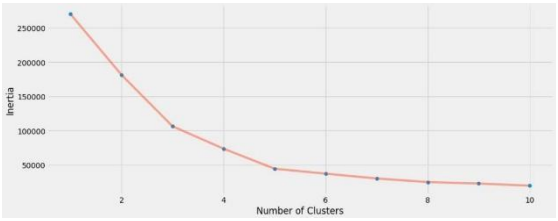

Figure 1-Graph to determine the elbow point

4. Mathematical Model

To determine Elbow point

The Elbow method is the best way to find the number of clusters. The elbow method constitutes running K-Means clustering on the dataset. Next, we use within sum of squares as a measure to find the optimum number of clusters that can be formed for a given data set. Within the sum of squares (WSS) is defined as the sum of the squared distance between each member of the cluster and its centroid..

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where x_i = data point and c_i = closest point to centroid

Figure 2 - Calculation of WSS

5. Visualize the clusters

The main objective of this investigation is to find a pattern among customers and how they like to purchase for maximum profit.

Here we have formed clusters to visualise our dataset. This cluster is our end result that provides us Customer Segmentation output.

```
plt.figure(1 , figsize = (15 , 7) )
plt.clf()
Z2 = Z2.reshape(xx.shape)
plt.imshow(Z2 , interpolation='nearest',
           extent=(xx.min(), xx.max(), yy.min(), yy.max()),
           cmap = plt.cm.Pastel2, aspect = 'auto', origin='lower')

plt.scatter( x = 'Annual Income (k$)' ,y = 'Spending Score (1-100)' , data = df , c = labels2 ,
            s = 200 )
plt.scatter(x = centroids2[: , 0] , y =  centroids2[: , 1] , s = 300 , c = 'red' , alpha = 0.5)
plt.ylabel('Spending Score (1-100)') , plt.xlabel('Annual Income (k$)')
plt.show()
```

Figure 3- Result of K- Means Clustering

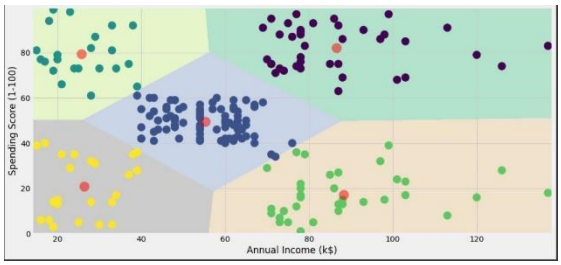

Figure 4- Result of K- Means Clustering

According to Figure 4, the graph of Annual Income Vs Spreading Score is plotted. The dots are the customers. The colours represent the clusters formed. The dot falling in the specific colour shares the most probable characteristics. The dots which fall on the edge of two colours are more likely to change over a time to specifically either of the segments.

6. Conclusion

Customer segmentation can have positive and powerful impact on business if done properly. It is essential to regularly review and rectify your customer segments, since parameter like customer behaviours, perceptions, demographics and other real world entities change over time. To avoid as much as creation of customer segments based on factors that quickly change instead calibrating according to values and parameters which has less frequency of changing. The method of elbow point is best way to find number of cluster which helps in finding and identifying pattern among customers and how they like to purchase for maximum profits. The proposed algorithm is trying to analyse the best possible segments depending upon the parameters provided delivering efficient results with good accuracy and time complexity.

References

[1] 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.

[2] 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.

[3] MacKay and David, "An Example Inference Task: Clustering," Information Theory, Inference and Learning Algorithms, Cambridge University
Press, pp. 284-292, 2003.

[4] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques", Third Edition.

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[6] S. Dasgupta and Y. Freund, "Random Trees for Vector Quantization," IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.

[7] Puwanenthiren Premkanth, —Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2