International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

TWO-STAGE CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING AND ARTIFICIAL NEURAL NETWORK

Harshit Muhal[1], Harsh Jain[2]

[1]Student, Dept. of Computer Engineering, Delhi Technological University, Delhi, India
[2]Student, Dept. of Computer Engineering, Delhi Technological University, Delhi, India

---***---

Abstract - *Customer segmentation is a very important technique used by companies to target their product features and prices and better serve their customers providing them the best of both worlds. In this project, we use the K-means algorithm to perform customer segmentation on data and try to analyze the results with regards to how many clusters give us the best results and use clustering. Then we train a neural network on the data obtained from clustering to classify new samples so as to present the problem in a supervised way instead of finding clusters using K-means in an unsupervised fashion. Using Neural networks we have not only segmented data into various clusters to perform customer segmentation but we also managed to capture the complex relationships that exist between the data attributes and the different clusters the observation points belong to without sacrificing the generalization used for the formation of initial clusters. We received an accuracy of 90.62% on training and 92.5% on the testing data.*

Key Words: Customer segmentation, clustering, Neural Networks, K-Means, Two-stage approach.

1. INTRODUCTION

In today's highly competitive world, companies need a way to efficiently make and manage products, in regard to their pricing, features, as well as other add-ons. Machine learning is one of the latest fields in innovation and engineering with applications in a vast variety of topics, subjects, and ways. Since machine learning is based around data, it is perfect for a data-centric world such as ours. By using this technology, we can efficiently perform customer segmentation, which is the task of dividing customers into classes that will group customers with similar purchasing behavior, feature demands, or other factors such as geographical variance, add-on, and extra services demanded, etc. This makes companies make better products as they can model their product around the classes of customers they want to cater to and thus balance both the features and price of the products giving the customers the best of both worlds.

Past work done in the field of customer segmentation includes the use of the Apriori algorithm [1], which is one of the fastest and earliest tools for association mining for segmentation. The use of Customer Value and Customer Behavior is made in customer segmentation using data mining techniques of simple CRM method for mobile VIP customer segmentation in [2]. In [3] Customer segmentation is done using an integrated approach with the Apriori algorithm as well as CRM method with associated mining, bringing the benefits of both these methods to solve this problem. In [4] The customer segmentation was done based on the e-commerce data . In [5], two approaches using LRFM (Length, Recency, Frequency, Monetary) model and extended model called LRFM -Average Item (AI) model were used for customer segmentation, in this, they found that adding simple parameter and averages did not produce better segmentation and did not show a significant difference in results indicating that complex parameters are required to get better results for customer segmentation.[6] showed an analysis of silent customers were performed and as silent customers are part of customers that a company can very easy to lose. Thus, it is necessary to analyze the features of such customers and make appropriate market decisions. To improve the enterprise's revenue in the telecom industry, that paper proposed a K-means++ method for customer segmentation based on silent customers. In [7], a segmentation algorithm based on density-based spatial clustering of applications with noise(DBSCAN) along with the K-means method was designed to satisfy the requirement of the Yunnan Electricity Market. In [8] K-means clustering technique and SPSS Tool were used to develop a real-time and online system to predict seasonal sales on annual cycles, which incorporated an important complex parameter of temporal spikes in sales of certain items. [9] chronicles the use of unsupervised machine learning to solve the problem of customer segmentation using the data for credit card based purchases for users in Africa. In [10]

Customer segmentation using a multi-layer perceptron (MLP) neural network which classifies customers into different sets according to attributes.

1.1 Advantage Over Other Methods

These methods use either a single complex technique or merge a few simpler ones to solve the task of customer segmentation and thus are mostly able to generalize well and some approaches are also able to capture a few complex parameters such as the inclusion of silent customers, but still lack the ability to deal extensively with the complex relationships which are required to be processed for better segmentation. These complex relationships give rise to some invisible parameters which are not directly obtained from the customers but form an important part of the segmentation task. The first stage of our approach helps in generalizing the classes with the initial formation of clusters. The second stage with the neural network helps capture the intricacies of these relationships between customers by selecting which attributes to use and combining them to form newer complex attributes that are present in the hidden layers of the network.

1.2 Contribution Of This Work

This paper introduces the idea of using a two-stage approach for customer segmentation, unlike any which has been proposed earlier. This method uses both unsupervised and supervised learning methods to achieve a meaningful segmentation and continuously update the model with new data without having to do the whole process again thus reducing the resources required for the task. This method also discusses the need to view and process complex relationships between the customers as part of the segmentation process which has not been discussed in earlier works.

2. THE DATA

The data used for this project has some basic information like Customer ID, annual income, spending score, gender, and age of the customer. The Spending Score is a variable that is calculated based on customer behavior and purchasing information. In the dataset, it is already calculated and available to us. The data is available at LINK.

3. METHODOLOGY

1. **Data Collection:** This step involves the collection of customer data which includes attributes like Age, Gender, income, and spendings.
2. **Data Preprocessing:** It involves Preprocessing of data by removing features that bring redundancy to the data and keeping only necessary features.
3. **Training:** Use the K-means algorithm to cluster the customers into different segments.
4. **Testing:** clusters will be used for extracting the associative buying pattern of the segmented customer to benefit the organization.
5. **Labeling:** Each observation in the dataset is labeled with the cluster it belongs to. This data will act as a base for training the neural networks.
6. **Model Architecture:** Feed-forward neural network is made for the prediction of the cluster to which a given observation belongs.
7. **Training Neural Network:** Training the neural network by using appropriate optimizers like Adam and loss functions like categorical cross-entropy.
8. **Predictions:** Neural Network is used to predict the cluster to which the given observation belongs to.

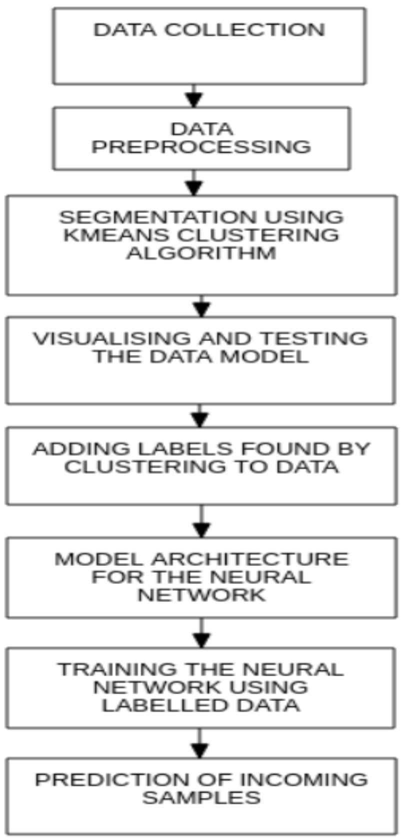

Fig -1: Methodology

4. K-Means Technique

K-Means is an algorithm that works iteratively and attempts to isolate the dataset into K distinct non-overlapping subgroups. The K-means algorithm assigns each observation in data to the cluster such that the squared distance of that point from the centroid of the cluster is minimum. Kmeans algorithm tries to attain a configuration in which clusters are far from each other but points in each cluster are similar.

K-Means algorithm follows the principle of expectation and maximization which consists of the following procedure:

1. Initialize all cluster centers randomly.
2. Repeat the steps given below until convergence.
3. E-Step: Points are assigned to the cluster center with a minimum distance from the point.
4. M-Step: New Cluster center will be updated to the mean.

The Expectation step is responsible for updating the expectation of which cluster each point belongs to. To do it the distance of the points is calculated from each cluster. The Maximization step is responsible for maximizing a cost function that represents the location of the centroid of the clusters. In K-Means maximization is done by taking a mean of the data in each cluster.

If we have (x1, x2,....., xn) as input, where each observation in the input is a p-dimensional vector with real values. K-Means will divide the input into k clusters (k<=n).

4.1 K-means Algorithm

1. **Begin** with input as X=(x1,x2,...xn)
2. **Initial Centroids** are taken as C1, C2,...Ck.
3. **Assign points** to the cluster which has a minimum distance from the point.
4. **Determine:** Update all Centroids -C1, C2,..Ck
5. **Repeat:** Until the significant change is not seen in centroids.
6. **Output:** Final Centroids - C1, C2,...Ck
7. **End**

5. DEEP NEURAL NETWORKS

Deep Neural Networks are used in modern computing in a variety of applications such as computer vision and NLP. This is due to their merit of being able to identify complex relationships by breaking them into smaller parts. We have used this deep neural network to classify new customers into the clusters by training on the labeled data we obtained by running k means on the unlabeled source data.

5.1 FULLY CONNECTED DENSE LAYER

A fully connected dense layer is a layer of neurons, each of which is connected to all the neurons of the previous layer, this helps us as we can weigh in all the possible combinations pertaining to the relationships.

5.2 DROPOUT LAYER

Dropout helps us reduce overfitting in our model so that even though the relationships are complex, they can still be generalized over a large audience. This layer sets weights of few edges passing through it to 0 during the forward training process. Since this happens randomly and variably, this reduces the chances of us missing out on better relationships.

5.3 BATCH NORMALIZATION

Batch normalization helps us to reduce training time as it normalizes the data using the mean. This helps with reducing computation resources required as everything is scaled using the mean and thus elementary operations are faster to perform(numbers are smaller). This adds the batch mean and standard deviation as parameters for our model.

5.4 PROPOSED NEURAL NETWORK :

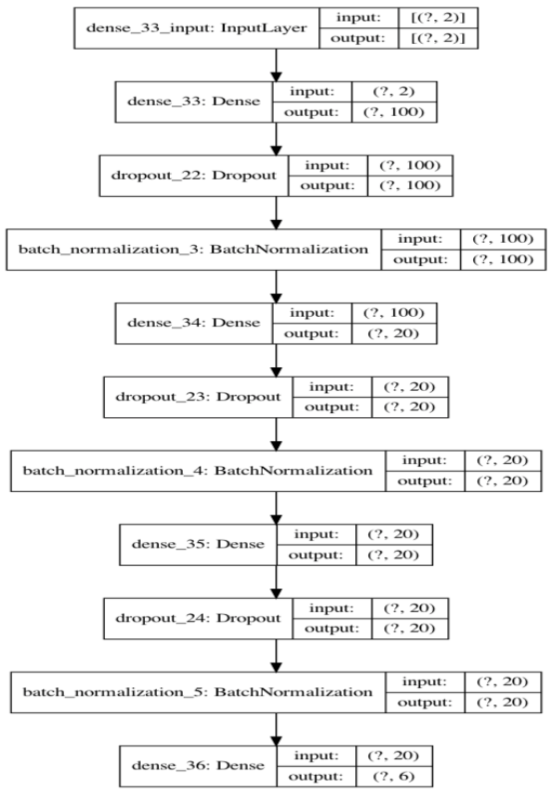

Fig -2: Model Architecture

In our experiment, we used a sequential deep neural network of fully connected dense layers. There are 4 fully connected dense layers in the model with 100, 20, 6, 6 numbers of neurons in each layer respectively. Each dense layer is followed by one Dropout layer and one Batch Normalisation Layer. All dropout layers have a frequency of rate 0.5 while the Batch Normalization layer doesn't have any parameters.

6. EXPERIMENTATION:

Clusters were obtained for different numbers of cluster centers as 4,5,6,7.

For training neural network RMSprop and adam optimizers were experimented and the best results were provided by adam optimizer. For our dataset optimal no of epochs was around 200 and no of dense layers were 4. Training is done in 2 parts. Different learning rates with values 0.005, 0.008,0.01 were experimented with. The model was trained with a learning rate of 0.001 for 200 epochs then the model is trained further trained but with a different learning rate that is 0.0001. The best model was saved by creating checkpoints by monitoring the accuracy.

For both stages training various parts of the model was done using google colab (collaboratory). Python 3.7.6 is used to write code for the project. Different Libraries like Numpy, Pandas, random, matplotlib, Keras, scikit-learn, and seaborn were used.

Chart -1: Training

7. Accuracy:

Accuracy = No of Correct predictions/ Sample Size

The model was able to get 90.625% accuracy on training data and 92.5% on the testing data.

8. RESULTS AND ANALYSIS:

8.1 Clustering With Different Values Of Cluster Centers

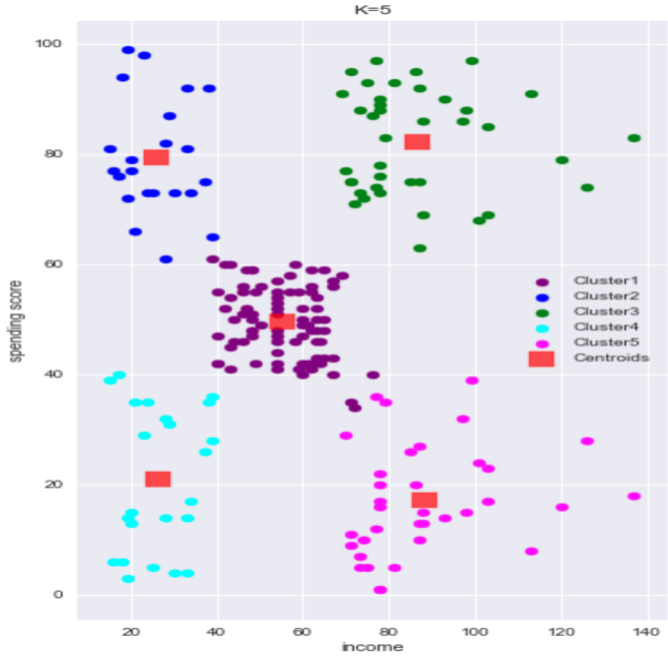

Chart - 2.1: Clustering with k=5

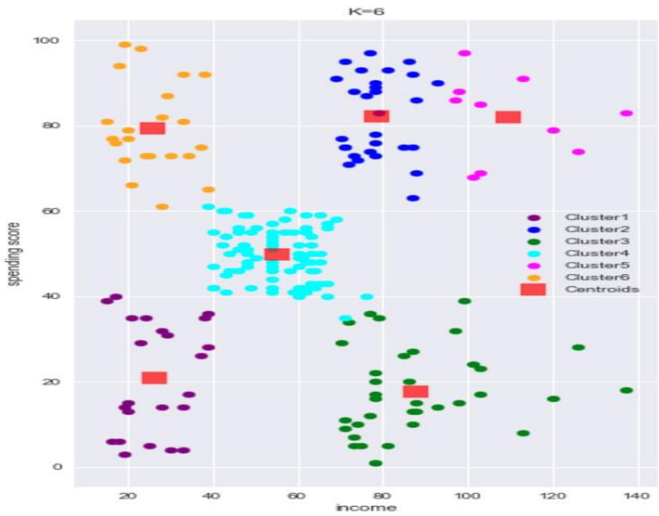

Chart - 2.2: Clustering with k=6

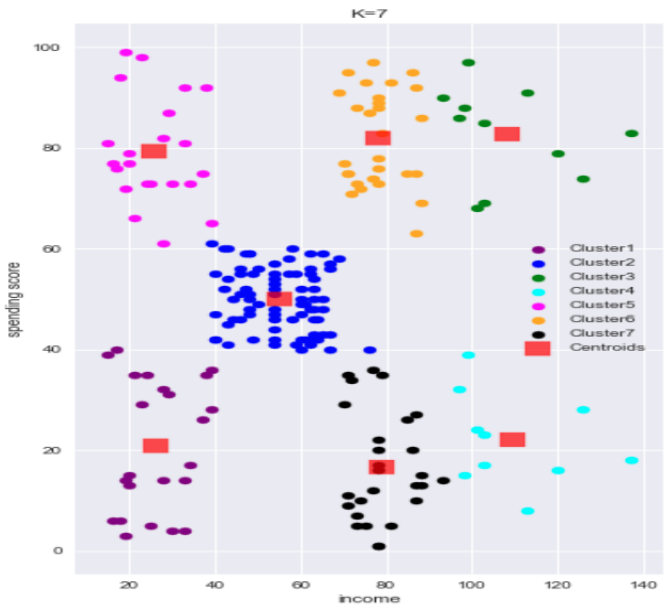

Chart - 2.3: Clustering with k=7

For the first clustering with k=5 as shown in chart 2.1, we see that we get some meaningful clusters but cluster 3 seems to try to integrate varied data into a single cluster and is thus constricted with this value.

For clustering with k=6 in chart 2.2, we see cluster 3 is broken into two clusters which have a better representation of the target audience and a meaningful split in the clusters.

For clustering with k=7 shown in chart 2.3, we see that the clusters are now being forced to split meaningful clusters into even more groups that incorporate noisy data into these clusters.

We see that for the value of k=6 we get the best results for our data with meaningful clusters whereas other values either miss clusters or try to integrate noise into the clusters. This value of k can be decided based on the data.

8.2 Cluster Analysis

Here are our findings from the clusters obtained :

Cluster 1 has medium income and low expenditure annually

Cluster 2 has low income and low annual spend

Cluster 3 has both income and annual spend high.

Cluster 4 has low income but has high annual spending.

Cluster 5 has both income and annual expenditure as a medium.

Cluster 6 has even higher income and high annual spending.

8.3 Key Strategies For These Clusters

Both clusters 3 and 6 have higher income and high expenditure, this might lead to strategizing the most feature-full and top quality products and services for them.

Cluster 4 has a low income and high expenditure, this reveals that they show some aspect of branding impact so they can be targeted with a discount or promotional offers on higher-end brand valued products.

Clusters 1 and 2 have lower expenditure so the focus for them would be basic level features or products which bring the best features for the price.

Cluster 5 has some flexibility as to what strategies can be applied to them, this brings some flexibility to the strategies that can be applied to them with a mixture of promotional offers for higher-end products to improve the best-valued products and services by a small margin to accommodate for the increased demand for features.

The reaction of clusters 1,2 and 5 will provide important information as to how the market will be progressing in terms of feature importance at a basic stage and feature qualities, i.e the quality of features at the upper level of expenditure.

9. CONCLUSION:

In this paper, various solutions to customer segmentation proposed by various authors have been explored. Using the K-Means technique we were able to cluster the data into meaningful clusters and provided key strategies based on associative customer behavior as observed from the obtained clusters. A Feed-Forward neural network was trained to present the customer segmentation as a supervised problem.

10. FUTURE SCOPE:

Algorithms like DBSCAN, CRM and LRFM, Fuzzy clustering, and Hierarchical clustering may be tried for the primary stage of clustering. These may prove to be advantageous in different scenarios based on different datasets. For the second stage, to capture relationships between data attributes and clusters more complex layers can be added in the neural network like convolutional layers. Different values of the learning rate, batch size, and other hyperparameters can also be experimented with.

ACKNOWLEDGEMENT

We would like to thank **Sanjay Patidar**, Assistant professor in the Department of Computer Engineering, Delhi Technological University (Formerly Delhi College of Engineering), Delhi, India for his continuous support and encouragement throughout the process and for guiding us while writing the research paper. We offer him our sincere appreciation for his excellent supervision.

REFERENCES

[1] S. S. Güllüoğlu, "Segmenting customers with data mining techniques," in International Conference on Digital Information, Networking, and Wireless Communications, Moscow, Russia, Feb. 2015, pp. 154-159.

[2] Z. Yihua, "Vip customer segmentation based on data mining in mobile-communications industry", Computer Science and Education (ICCSE) 2010 5th International Conference on, pp. 156-159, 2010.

[3] Balmeet Kaur, Pankaj Kumar Sharma, “Implementation of Customer Segmentation using Integrated Approach”, IJITEE, ISSN: 2278-3075, Volume-8 Issue-6S, April 2019.

[4] Juni Nurma Sari1,2, Lukito Edi Nugroho ,Ridi Ferdiana ,P. Insap Santosa, “Review on Customer Segmentation Technique on Ecommerce”,American Scientific Publishers, Vol. 4, 400–407, 2011.

[5] Pradnya Paramita Pramono, Isti Surjandari, Enrico Laoh , “Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method”, 2019 16th International Conference on Service Systems and Service Management (ICSSSM).

[6] Yuhang Qiu , Pingping Chen , Zhijian Lin , Yongcheng Yang , Lanning Zeng , Yaqi Fan, “Clustering Analysis for Silent Telecom Customers Based on K-means++”,2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 123-127.

[7] Xuejin Wang *et al*, “Electricity Market Customer Segmentation Based on DBSCAN and K-Means”, 2020 Asia Energy and Electrical Engineering Symposium, pp. 869-874.

[8] Kishana R. Kashwan, Member, IACSIT, and C. M. Velu,”Customer Segmentation Using Clustering and Data Mining Techniques”, IJCTE Journal, Vol. 5, No. 6, December 2013.

[9] Eric Umuhoza, Dominique , Jane Awuah , and Beatrice Birir, “Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa”, Vol.111(3) September 2020.

[10] Şukru Ozan and Leonardo O. Iheme, “Artificial Neural Networks in Customer Segmentation”, 2019 27th Signal Processing and Communications Applications Conference (SIU).