THEORY OF COMPUTATION

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

Submitted to :

Mr. Sanjay Patidar

Asst. Professor

Computer Science department

Delhi Technological University

Submitted by :

Harsh Jain 2K18/CO/140

Harshit Muhal 2K18/CO/145

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

Harshit Muhal , Harsh Jain, student, Delhi Technological University

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. Clustering becomes a difficult task if performed repeatedly with small inclusions in the dataset.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 best results and useful 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. This saves time as we do not have to perform clustering over all the data repeatedly. We also want to analyze the performance of our system when dealing with big data, so we use graphical processing units to boost our algorithm and make it perform faster on bigger and bigger sets of data . We also look at the differences between computation on a CPU and a GPU for our algorithm and compare their advantages and disadvantages for such types of tasks.

Introduction

In today's highly competitive world, companies need a way to efficiently make and manage products ,in regards to it's 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, that is the task of dividing customers into classes which will group customers with similar purchasing behaviour, 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 best of both worlds.

Past work done in the field of customer segmentation use of Apriori Algorithm, which is one of the fastest and earliest tools for Association Mining for segmentation [8] .Unsupervised machine learning to cluster where the results of credit card customers segmentation revealed that customers are grouped into four distinct segments. [7] Customer segmentation using data mining techniques of CRM method which is a simple method of the mobile VIP customer segmentation makes the most use of the Customer Value and Customer Behavior model. [10] Customer segmentation using ecommerce data. [13] Customer segmentation using an integrated approach with Apriori algorithm as well as CRM method with associated mining which brings the benefits of multiple methods to solve a problem. [12] Two approaches using LRFM (Length, Recency, Frequency, Monetary) model and extended model called LRFM -Average Item (AI) variables in

clustering process are compared by validity index to obtain the best result for customer segmentation. The result shows that adding new variable Average Item in LRFM model has no significant difference or better results in clustering . [9] Analyzing silent customers as silent customers are part of customers that company is very easy to lose. 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. This paper proposes a K-means++ method for customer segmentation based on silent customers. [11] A segmentation algorithm based on density-based spatial clustering of applications with noise(DBSCAN) and k-means method is designed to satisfy the requirement in Yunnan Electricity Market.[6] A comprehensive report of using k-means clustering technique and SPSS Tool to develop a real time and online system for a particular super market to predict sales in various annual seasonal cycles.[1] Customer segmentation using a multi layer perceptron (MLP) neural network which classifies customers into different sets according to attributes.[14]

As we get bigger and bigger sets of data, a CPU may not give us fast enough processing ,this is why we look towards using GPUs to process big data. Past work has not explored the idea of clustering the data and labelling it based on those clusters to produce a labelled dataset which can then be used to train a neural network which can better identify the complex relationships between different types of customers which brings them into same category .By using this alternate two stage approach , we can easily save time and resources as we do not have to run large computations repeatedly for new data. The dataset can be clustered initially and then used to train the neural network which can allow classification of new customers without having to run the clustering algorithm again.This process can be repeated after a set interval of time to cluster the data again and retrain the neural network to minimize any error that may occur due to persistent use without retraining .

THE DATA:

The data used for this project has some basic information like Customer ID, age, gender, annual income and spending score where the Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data. The data is available at LINK.

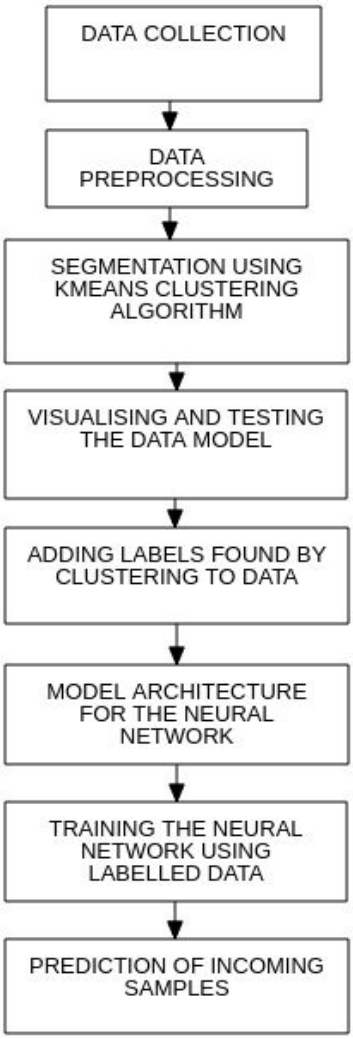
METHODOLOGY :

A. **Data Collection**: This step involves the collection of customer data which includes attributes like Age, Gender, income, and spendings.
B. **Data Preprocessing:** It involves Preprocessing of data by removing features that bring redundancy to the data and keeping only necessary features.
C. **Training:** Use K-means algorithm to cluster the customers into different segments.
D. **Testing:** clusters will be used for extracting the associative buying pattern of the segmented customer to benefit the organization.
E. **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.

F. **Model Architecture:** Feed-forward neural network is made for prediction of the cluster to which a given observation belongs.
G. **Training Neural Network:** Training the neural network by using appropriate optimizers like Adam and loss functions like categorical cross-entropy.
H. **Predictions:** Neural Network is used to predict the cluster to which the given observation belongs to.

BLOCK DIAGRAM OF PROPOSED WORK:



K-Means Technique:

K-Means algorithm is an iterative algorithm that tries to divide the dataset into K distinct non-overlapping subgroups (clusters). It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

K-Means algorithm is based on expectation-maximization which consists of the following procedure: [2]

1. Initialize cluster centers randomly.
2. Repeat until convergence.
   2.1. E-Step: Points are assigned to the nearest cluster center.
   2.2. M-Step: Cluster center is updated to the mean.

The E-step or the Expectation step updates the expectation of which cluster each point belongs to by calculating distance of the point from each cluster. The M-step or Maximization step involves maximizing some cost function that defines the location of the cluster center. In K-Means maximization is accomplished by taking a mean of the data in each cluster. [2]

Given a set of observations (x1, x2,....., xn), where each observation xi is a d-dimensional real vector and k is the no of clusters (k<=n). [1]

K-means Algorithm :

Begin *with input as X=(x1,x2,...xn)*
Initial Centroids *- C1, C2,...Ck.*
Assign *points to the nearest cluster.*
Determine: *Update Centroids -C1, C2,..Ck*

Repeat: *Until Centroids don't change. significantly (specified Threshold value)*
Output: *Final Centroids - C1, C2,...Ck*
End

Approaches to customer segmentation : [4]

1. **A priori segmentation:** It utilizes an order plot dependent on freely accessible attributes, for example, industry and company size — to make unmistakable gatherings of clients inside a market.
2. **Needs-based segmentation:** It depends on separated, approved drivers (needs) that clients express for a particular item or administration being advertised. The necessities are found and confirmed through essential statistical surveying, and fragments are divided dependent on those various needs as opposed to attributes.
3. **Worth based segmentation:** It separates clients by their monetary worth, gathering clients with a similar worth level into singular fragments that can be unmistakably focused on.

COMPUTATION-

Using a graphical processing unit (GPU) for computation - A graphical processing unit or a GPU can be used to accelerate the computation in various tasks such as 3D rendering, simulation, video acceleration, and boosting machine learning algorithms for big data. One important tool in using a graphical processing unit in computational tasks is CUDA.

CUDA:

CUDA is a parallel computing platform and programming model developed by NVIDIA for general computing on graphical processing units (GPUs). With CUDA, developers are able to dramatically speed up computing applications by harnessing the power of GPUs. [5]

CUDA stands for Compute Unified Device Architecture and it refers to two things CUDA architecture and CUDA programming model. The architecture is formed with lots of CUDA cores that are placed on a GPU, the programming refers to CUDA programs written to perform tasks in parallel computing on these GPU's. The programs can be written for just one data element and CUDA can automatically scale it to work parallely on the hundreds of cores in the GPU.

GPU vs CPU:
A GPU differs vastly from a CPU but together they can be used to form a heterogeneous system which brings us the utility and advantages of both these processing units.

A CPU usually works on a few cores, averaging between 4-8 while CPU's with 64 and 128 cores are also commonly used in supercomputers, these usually work with one or two threads per core. In a GPU there are usually a few hundred cores, and each of these have tens or hundreds of threads thus bringing the total to thousands of threads parallely computing and performing tasks .

A GPU works parallely on tasks which if performed on a CPU are done sequentially, that is using a for loop while in GPU vector addition and vector operations may be performed.

CPU's are suited to tasks related with latency or where per-core performance is important while GPU's are suited to tasks which can be divided, and performed parallelly or where high throughput is required.

Advantages-Disadvantages:

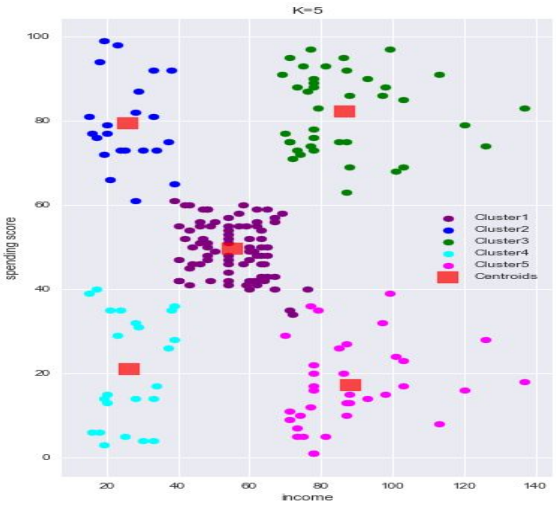
CPU's have more powerful cores as compared to GPU 's thus they can perform better for computationally complex task if per-core performance is considered

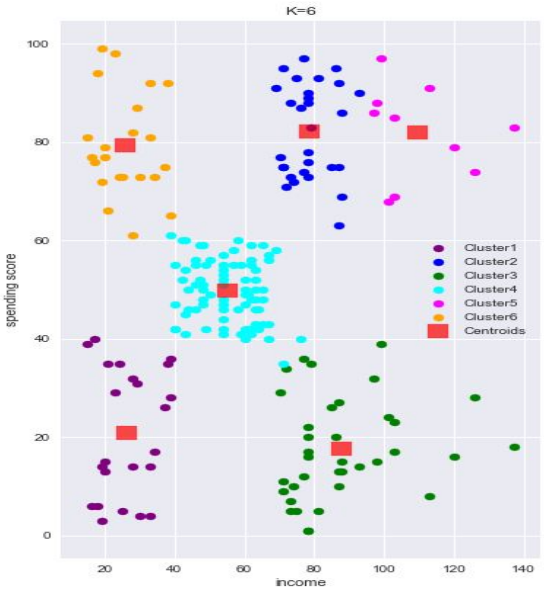
GPU's have more number of weaker cores which can outperform CPU when task can be parallelly processed such as in big data analysis or 3D rendering.

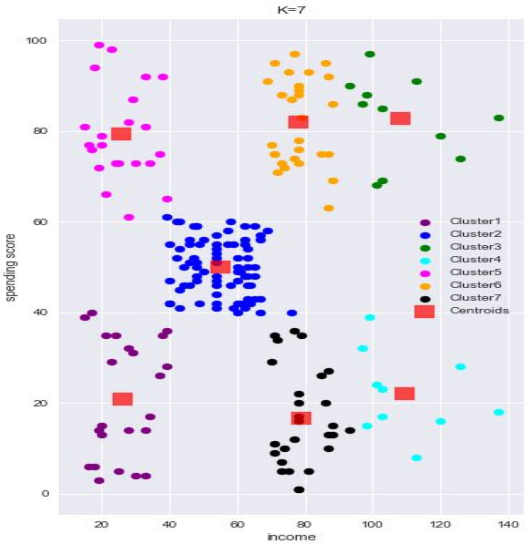
EXPERIMENTATION:

This project was implemented using Jupyter notebook software which is used to write notebooks in python programming language. Python 3.7.6 is used to write code for the project. For training various parts of model google collab was used. Different Libraries like Numpy , Pandas, matplotlib, Keras, scikit-learn and seaborn were used.

STATISTICAL ANALYSIS:







Customers are segmented with k=5,6,7 and setting k as 6 seems to provide more meaningful results.

References :

[1] Kishana R. Kashwan, Member, IACSIT, and C. M. Velu, International Journal of Computer Theory and Engineering, Vol. 5, No. 6, December 2013.

[2]https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html

[3]https://moosend.com/blog/customer-segmentation-ecommerce/
[4]https://openviewpartners.com/blog/customer-segmentation/#.X3avnC8Rqu4
[5] https://developer.nvidia.com/cuda-zone
[6] Xuejin Wang , Chongdong Zhou , Yijing Yang , Yueyong Yang , Tianyao Ji
Jifei Wang, Jingrui Chen, Ying Zheng. “Electricity Market Customer Segmentation Based on DBSCAN and k-Means”,Kunming Power Exchange, Kunming, China.
[7]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
[8] abri Serkan Güllüoğlu, Türkoba Mah. Erguvan Sok. 26-K Tepekent/Büyükçekmece, Segmenting customers with data mining techniques.
[9] Pradnya Paramita Pramono, Isti Surjandari, Enrico Laoh , “Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method”
[10] ZHANG Yihua, “Vip Customer Segmentation Based on Data Mining in Mobile-communications Industry”.
[11] Yuhang Qiu , Pingping Chen , Zhijian Lin , Yongcheng Yang , Lanning Zeng , Yaqi Fan, “Clustering Analysis for Silent Telecom Customers Based on K-means++”, pp. 123-127
[12] Balmeet Kaur, Pankaj Kumar Sharma, Implementation of Customer Segmentation using Integrated Approach IJITEE, ISSN: 2278-3075, Volume-8 Issue-6S, April 2019
[13] Review on Customer Segmentation Technique on Ecommerce, Juni Nurma Sari1,2, Lukito Edi Nugroho ,Ridi Ferdiana ,P. Insap Santosa, American Scientific Publishers, Vol. 4, 400–407, 2011.
[14] Şukru Ozan and Leonardo O. Iheme, “Artificial Neural Networks in Customer Segmentation”.