Kmeans Clustering to Cluster the Clients into different Segments Based on Spending using Python

Dr.Imran Qureshi[1], Mr.Burhanuddin Mohammad[2] and
Mohammed Abdul Habeeb [3]
Information Technology
University of Technology and Applied Sciences -UTASA
Al Musannah, Oman
[1] imran@act.edu.om, [2] Burhanuddin@act.edu.om and [3] habeeb@act.edu.om

Abstract

In today's world information is considered as vital source, may it be in banking sector, medical sector, retail sector, and many other. One challenge would be how do we get the required information from vast amount of data, sometimes referred as data tombs. The answer to this question is not easy. But as technology advanced, we have got opportunity to discuss regarding such type of real time problems such as information gathering from innumerous amount of data. Machine learning is a branch of science used to study of algorithms and statistical models that will perform a specific task. Machine learning is called as subset of artificial intelligence. Data Mining is a field of study within machine learning. Data mining is used for exploratory data analysis using unsupervised learning, it is also referred as predictive analysis [1].

Keywords*-Data Mining, K-Means, Clustering, Machine Learning, Python*

.

1. Introduction

This KMeans Clustering algorithm is an unsupervised algorithm. KMeans clustering algorithm is useful to find the clusters of data/clients to which they rightly belong. KMeans clustering is used for partitioning the data into 'k' clusters based on the nearest centroid. This is the basic idea behind the KMeans Clustering[1]. The following steps serve as basic porotype for KMeans clustering. The aim of this research is to propose how to group the customers based on their spending using KMeans clustering algorithm. Python programming is used to implement the KMeans clustering using a real time data sets from retail market. The objective of the research is to improve the business based on the customer buying pattern.

The organization of the paper is as follows, In section I contains introduction of the research with aims and objectives defined clearly, Section II defines about the K-Mean clustering algorithm which is used to group the customers based on the buying patterns, Section III defines the problem statement, section IV explains step by step procedure of implementation and finally section V discusses about the results achieved.

2. Algorithm for k-means clustering algorithm

Step1: Choose the number of clusters.
Step2: Select at random k points for the centroids.
Step3: Assign each data point to the closest centroid (This form k clusters).
Step4: Compute and place the new centroid of each cluster.
Step5: Reassign each data point to the new cluster centroid.
Step6: If any reassignment took place, go to step4, otherwise stop (Convergence attained).
Step7: Finally, your model is ready.

The above steps are used in KMeans clustering to find clusters of data. But it is essential to take the following points into consideration.

K-Means suffers from ramdom initialization trap

In order to avoid this problem, we have KMeans++ algorithm. Good news is that KMeans++ actually happens in background. We have to make sure the tools we are using should address this issue.

Choosing the right number of clusters

As seen in the above algorithm(Step2), we used predetermined number of clusters. The following question arises, Are the number of clusters are optimal? If no, then how can we find the optimal number of clusters? The answer to this question is using a metric called WCSS (With in cluster sums of square). Using this metric, we can find the optimal number of clusters.

The formula for WCSS is given below

$$\text{WCSS}=\sum_{pi\ in\ cluster1}^{n}(pi,c1)^2 + \sum_{pi\ in\ cluster2}^{n}(pi,c2)^2 + \cdots + \sum_{pi\ in\ clusterN}^{n}(pi,cN)^2$$

Where,
N=Number of clusters
Pi=Each data point in the cluster

WCSS is used to find the goodness or fit of each cluster. Now the question arises, what is the maximum limit of WCSS. In other words, how many clusters we can have? One answer to this question will be, the number clusters is equal to the number of data points, that makes the distance between each point to be zero. So lesser the value of WCSS the better the goodness of cluster. Another question will be what will be the optimal number of clusters? The answer to this question will be, the optimal number of clusters can be found using 'The Elbow Method'. Elbow method is to draw a graph between the number of clusters and WCSS. The optimal number of clusters will be found when the curve takes the elbow path. In other words, the changes between the number of clusters and WCSS will be substantial after a certain point (Elbow curve). This point is considered as optimal. For example, see below the following graph plotted between Number of clusters and WCSS[3,5].

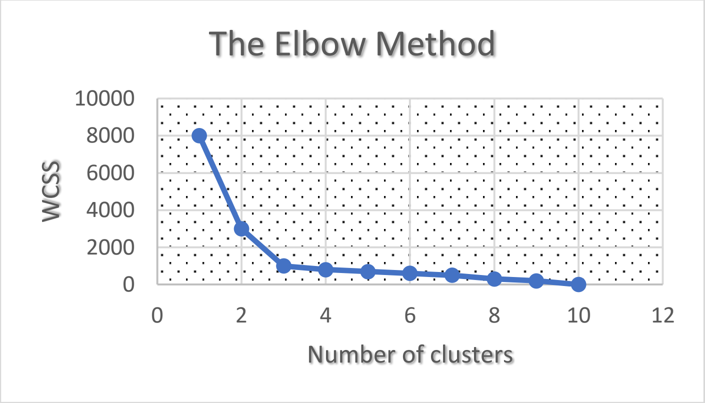

Figure I

From the scattered graph it is clear that the optimal number of clusters for this data will be 3. As you can see in the above figure that the change is substantial after 3 clusters. At 3rd clusters the shape looks like an elbow, hence elbow method. The Elbow method is very effective in finding the number of clusters.

Using KMeans algorithm let us try to solve a real-time solution to the problem

Problem Statement

Market basket analysis is an effective approach to know about the customers in any shopping mall. One such real time problem is to group the customers into different segments based on their spending, such as careful spenders, high spenders, standard spenders, careless spenders etc. Once the groups are identified then it will be very easy for the organizations to build effective strategies for target groups. Let us assume a scenario where you have been hired as a data scientist in some organization and the task given to you is to identify and group the customers into different segments based on their spending. This type of problems is categorized under unsupervised learning, that is you do not have any previous models. How do you solve this real time problem? There may be many solutions based on the tools and strategy you are using. But one thing in common is to solve the problem.

In this research paper I am proposing one such approach to solve this real time problem using KMeans clustering algorithm using python, the reason for choosing python is because it is widely used in machine learning by researchers as it is found to be very effective[2].

Implementation Steps

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

1) Set the working directory. In other words, import the data file on which you are applying the KMeans clustering algorithm
2) Import the following libraries
3) import numpy (Used to do mathematics)
4) import matplotlib. pyplot (Used to plot charts)
5) import pandas (Used to import and manage datasets more easily)
6) Import mall data set using pandas
7) Use elbow method to get right number of clusters
8) Plot the elbow graph
9) Apply KMeans to the mall dataset.
10) Visualize the results.

Results

As we can see from below figure 2, the scattered plot graph that using WCSS and elbow method, the optimal number of clusters are five. The data set used is a real time mall data. The data is preprocessed before using as per requirement. The mall data is a d-dimensional vector. In order to avoid the complexity, we used two dimensions namely annual income and spending (in a range of 1-100). The data is then applied for clustering using KMeans algorithm.

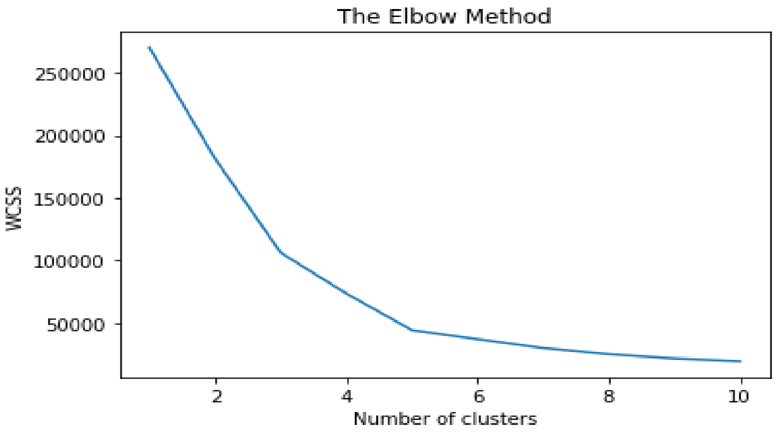

Figure 2

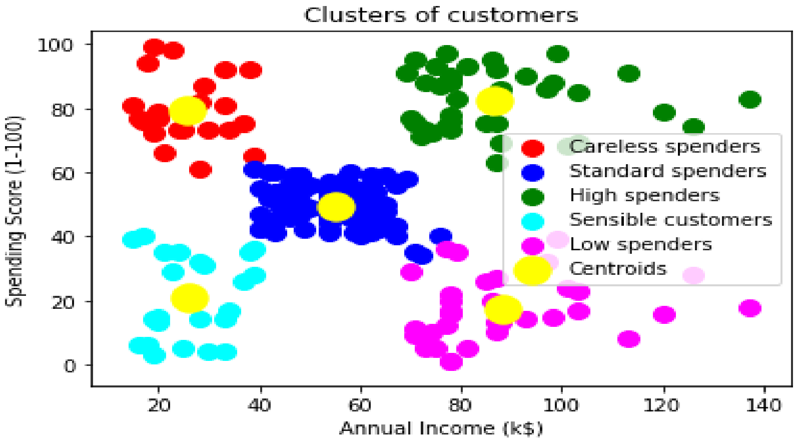

Figure 3

From figure 3 We can observe the following things

i. Red colored dots represent Careless customers, as we can see their earning is less but spend more

ii. Blue colored dots represent standard customers, their income is average and so as the spending

iii. Green colored dots represent high spenders, as their income is high and so as their spending, this group can be considered as our target group.

iv. Cyan colored dots represent sensible customers, as their income is low and so as their spending

v. Magenta colored dots represent low spenders, as their income is high, but spending is low

Conclusion and Future Scope

The data used in the research were taken from a well-known retail store. From above results we can conclude that KMeans algorithm is very useful for categorize the data into different segments-based requirement. In this research paper the criteria were to group the customers based on their spending. Similarly, we can choose the attributes based on the real time problem. Finally, the research can be further improved by using a d-dimensional vector. We can make very small changes to the existing code to achieve the results using d-dimension. We can also consider dimensionality reduction to avoid unnecessary attributes.

References

[1] Tripathi, Shreya & Bhardwaj, Aditya & E, Poovammal. (2018). Approaches to Clustering in Customer Segmentation. International Journal of Engineering & Technology. 7. 802. 10.14419/ijet.v7i3.12.16505.

[2] Kansal, Tushar & Bahuguna, Suraj & Singh, Vishal & Choudhury, Tanupriya. (2018). Customer Segmentation using K-means Clustering. 135-139. 10.1109/CTEMS.2018.8769171.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[4] https://en.wikipedia.org/wiki/Machine_learning

[5] https://scikit-learn.org/stable/modules/clustering.html#k-means