Prediction of Consumer Trends in Retail Industry

Chandrashekhara K T
Dept of Information Science and Engineering
BMS Institute of Technology and Management

Abstract- *the most commonly used algorithm by practitioners whenever they want to solve clustering problems is K means algorithm. K means is one of the most easy and popular algorithm. K means groups data points into subgroups and these are non-overlapping subgroups. K means algorithm also gives a clear idea about the dataset. K means algorithm's performance is best with spherical shaped clusters.*

I. INTRODUCTION

Machine learning and algorithms includes clustering process which identifies how different types of data are related. It creates new clusters or segments based on the relationship. Clustering also finds data points relation so that they can be clustered.

K means is classified as unsupervised learning algorithm and it is used for unlabeled data. The aim of the algorithm is to find the consumer groups for the dataset. These groups are based on the behavior of the consumers.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

Elbow method is used to find the number of consumer clusters to be formed. Based on the features provided algorithm iteratively works to assign each data point a cluster. These data points have similarity. With the help of clustering we can discover a new segment of users and their behavior and based on the behavior of the users and 3 distinct clusters are formed. Data analytics usually consists of large amount of data but the data has similarities. Therefore we group data into fewer clusters and each of these clusters will have data which is similar.

Even though we have millions of customers in data they belong to just few segments (cluster). It required for us to study each segment separately as they may behave different. Example different market segments have different preferences of products and different behavioral patterns.

We can use statistical techniques in such situation and these are called as clustering techniques. These methods are called as clustering techniques. Distance metrics is one of the mathematical techniques and is used to define similarities and differences between data observations. The exact definition of distance metrics acts as an important key of clustering and segmentation. These cannot be defined using black box mathematical equations but they require contextual knowledge.

There are various applications of cluster analysis. For example it can be used to identify consumer segments, or competitive sets of products, or groups of assets whose prices co-move, or for geo-demographic segmentation, etc.

It is necessary for us to split our dataset into clusters and is used to perform analysis in each cluster so that cluster insights are developed. This is applied even if there are no natural clusters in our dataset.

Highlight Clustering is a statistical technique much similar to clustering. It is used for raw dataset and it gives sensible clusters. The aim is that these clusters should have similar behavior and characteristics.

The main goal is to find the optimum number of clusters. There are two types of clustering techniques

- Hierarchical
- Non hierarchical

Cluster should be performed repetitively and continuously to discover the knowledge from huge quantities of raw and unorganized data.

Particular clustering type should be selected for particular problems to obtain efficient results.

Clustering technique is a data mining tool used in various applications. The fields of application include machine learning, patter recognition, classification. Recently data

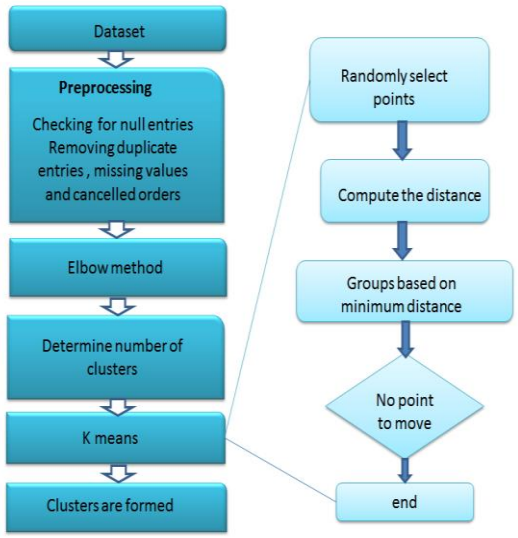
mining has found its application in distributed and grid computing.

II. PROPOSED SYSTEM

A. K means

The diagram 3.1 shows the workflow of the proposed system. First step is to cleanse the dataset. In this step the dataset is checked for null entries, duplicate entries are removed, missing values and cancelled orders are also deleted

The second step is to perform the elbow method to determine the number of cluster for the dataset then the k means algorithm is implemented.

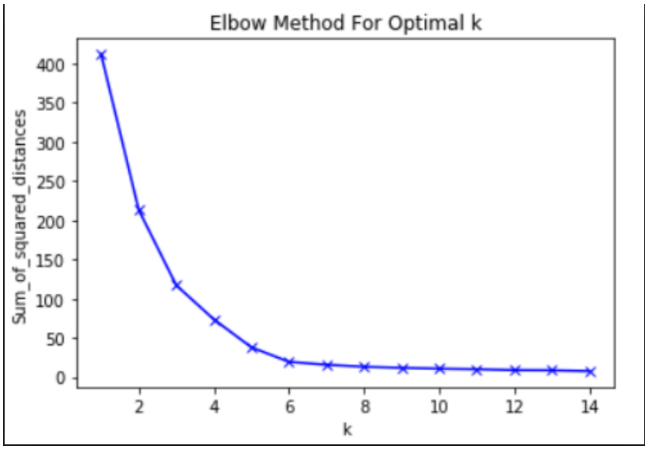


3.1 System diagram

The clustering algorithm k means is used to segment and target the consumer of the wholesale super market. The consumer clustering is based on the annual spending of the consumers. K means algorithm performs exceptionally well with large data set and therefore it works efficiently with the super market dataset used in the system. K means aims to form the K cluster in the given data.

The algorithm described above clusters the data for a pre-chosen K. Number of clusters in data set is obtained by running the algorithm for a range of K values and compares the results. There is no defined method through which we can find the accurate number of clusters for a dataset.

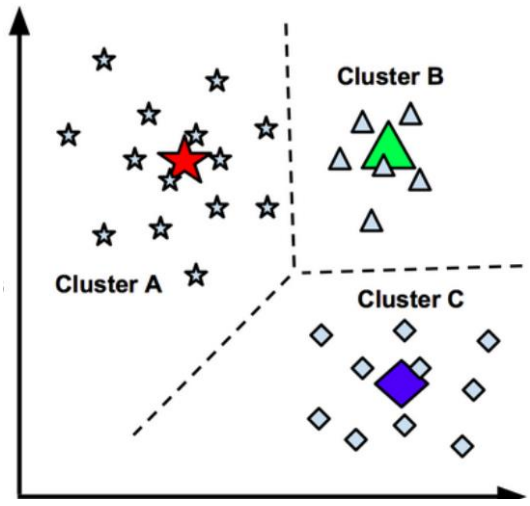
Elbow Method



3.2 Graph for elbow method

There is no defined method through which we can find the accurate number of clusters for a dataset. One of the commonly used techniques is elbow method, results are compared across different values of K the mean distance between the centroids and data points (WCSS - With in Cluster Sum of Squares) is determined. Increasing number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts (where distortion goes rapidly), can be used to roughly determine K. Based on the analysis the Elbow method results the value of k is estimated to be 3 for further analysis.

Cluster Analysis



Consumer clustering is an important marketing strategy which is widely used in businesses. The benefit of consumer clustering gives the better knowledge about the types of users which helps in business and marketing

strategies. This type of marketing technique is a subset of a company's Business Intelligence.

In this algorithm 3 different clusters are formed

Customer Type 1: These customers like to purchase few products with low price

Customer Type 2: These customers like to purchase more products with low price

Customer Type 3: These customers like to purchase few products with high price

Dataset

Data is the most vital part of clustering algorithm. The important aspects of data are the quality and quantity of data. To run any function on the data that finds out some sort of similarity or clusters the data needs to be arranged into feature vectors with a set of feature values. The amount of data is however not a problem, but this algorithm performs well with large size of data. Another important aspect in customer segmentation is to understand the available data.

The dataset used in this proposed system is taken from kaggle; kaggle is a free online website, which has huge deposit of large dataset. The data set used is from a super global store, which is a wholesale store. This store has dealer from all over the world and it sells electronic goods. There are multiple columns in the dataset such as invoice number, consumer id, product id, unit price, country and number of products purchased.

Equations

K-means clustering is a type of unsupervised learning, which is used when unlabelled data (i.e., data without defined categories or groups) is used. The goal of this algorithm is to find customer groups (clusters) in the data based on their behaviour. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \quad (1)$$

where

wik=1 for data point xi if it belongs to cluster *k*;
otherwise, wik=0.
Also, μk is the centroid of xi's cluster.

It's a minimization problem of two parts. We first minimize J w.r.t. wik and treat μk fixed. Then we minimize J w.r.t. μk and treat wik fixed. Technically speaking, we differentiate J w.r.t. wik first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. μk and recomputed the centroids after the cluster assignments from previous step (*M-step*).

Therefore, E-step is:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster's centroid.

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}} \quad (3)$$

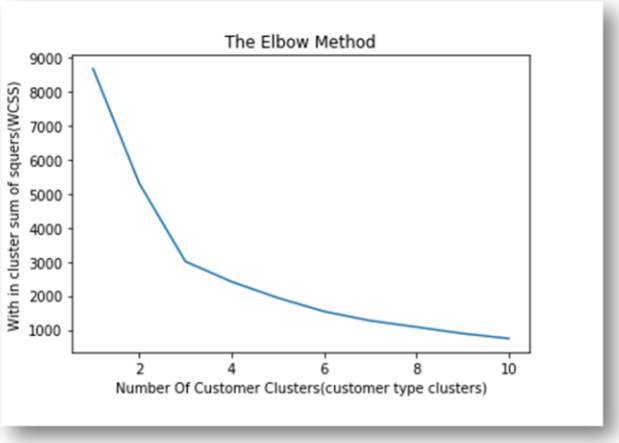
$$\frac{1}{m_k} \sum_{i=1}^{m_k} \| x^i - \mu_{c^k} \|^2$$

The number of groups to be formed is determined by using the Elbow Method. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. Clustering helps to discover a new segment of users and their behaviour and based on the behaviour of the users 3 distinct clusters are created.

Results

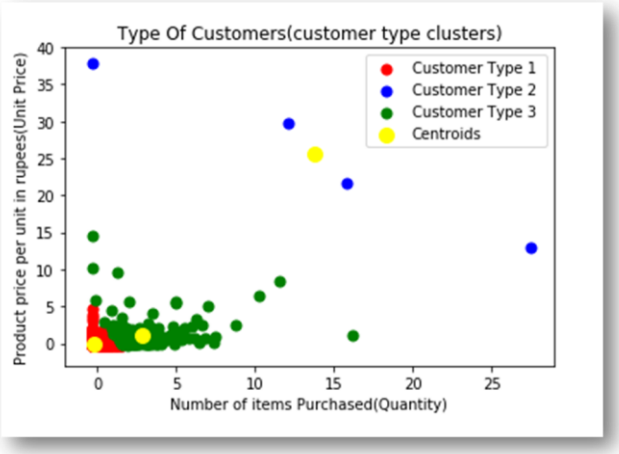
Elbow Method:

The graph 5.1 indicates the number of cluster that has to be formed for the dataset. The x-axis of graph has number of cluster and y-axis has WCSS (within the cluster sum of squares). The sharp bend in the graph indicates the number of clusters to be formed. In the below diagram the sharp bend is observed at 3, therefore 3 unique clusters are formed.



K means

The number of clusters are formed here are the consumers who have similar purchasing habits from the wholesale shop dataset. The final resulting graph obtained with the consumer clusters indicated in 3 distinct colours and centroids are also indicated using separate colour. Each coloured dot indicates the consumers who have similar spending habits.



Customer Type 1 (Red): These customers like to purchase few products with low price.

Customer Type 2 (Blue): These customers like to purchase more products with low price.

Customer Type 3 (Green): These customers like to purchase few products with high price.

III. CONCLUSION

In this paper, I have determined consumer clusters based on their purchasing habits and 3 types of groups are formed for the retail store dataset. Elbow method is used to determine the number of clusters and K means algorithm is used to form clusters consisting of similar data.

IV. FUTURE SCOPE

In the future, we can use the algorithm to form more than three consumer groups and also find the number of consumer in each groups for better market prediction.

REFERENCES

[1] Stephen Haben, Colin Singleton, and Peter Grindrod, 'Analysis and Clustering of Residential consumers Energy Behavioral Demand Using Smart Meter Data' IEEE TRANSACTIONS ON SMART GRID, VOL. 7, NO. 1, JANUARY 2016

[2] Paria Jokar, Nasim Arianpoo, and Victor C. M. Leung 'Electricity Theft Detection in AMI Using Consumers' Consumption Patterns IEEE TRANSACTIONS ON SMART GRID, VOL. 7, NO. 1, JANUARY 2016

[3] Chiung-I Chang and Jui-Chih Ho 'A Two-Layer Clustering Model for Mobile consumer Analysis' May/June 2017

[4] Dhendra Marutho , Sunarna Hendra Handaka , Ekaprana Wijaya , Muljono 'The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News' 2018 International Seminar on Application for Technology of Information and Communication

[5] Mediana Aryuni, Evaristus Didik Madyatmadja, Eka Miranda 'consumer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering' 2018 International Conference on Information Management and Technology.