Customer Segmentation using K-means Clustering

J Madhu[1], Kavita K Revanakar[2], Lavanya[3], Akash[4]
Department of Computer Science and Engineering Srinivas Institute of Technology, Mangalore, Karnataka, India

Submitted: 01-07-2021 Revised: 13-07-2021 Accepted: 16-07-2021

ABSTRACT - Analyzing customer related data is the most widely used applications of machine learning. In the clustering, similar objects are grouped together. This type of grouping mechanism is helps marketers to divide the customers depending on their purchasing characteristics this paper, we analyzed online retail customer data which contains all the transactions of UK-based store online retail between 2009 and 2011.This store mainly sells unique all-occasion gift-ware products. Most of the customers of this company are wholesale customers. We proposed a model for analyzing customers using K-means, Hierarchical clustering algorithms. The dataset is collected from UCI machine learning repository. Dataset contains features. For these features, we build a machine learning model for dividing wholesale customers into different clusters. We used python programming language for implementing clustering algorithms.

Keywords: Wholesale customer-data, Clustering, UCI repository, Machine Learning, Python.

I. INTRODUCTION

Advertising methodologies are continually changing dependent on client division. The marketing offices depends on a lot of information analytics to improve their growth[1]. Customer needs are consistently significant for marketing. So, finding the best approach to improve the deals is just by breaking down client needs. This paper investigates the dataset of wholesale customers. Exploration of information uncovers different realities about the informational collection which can supportive for enhancement of marketing strategies. Clustering is one of the important learning procedure which attempts to aggregate the information dependent on their similarities. Clustering is a grouping methodology in which we can easily identify the dissimilar items. Machine Learning calculations assists with investigating the client information in various ways. Clustering bunches the dataitems for simple analyzation. Selection of a clustering algorithm is additionally considered to be important factor for better analyzation [2]. Based on this grouping, promoting systems updated. Hence, data experts are doing part of research in clusterings ideas. ML clustering analysis is helpful for making any sort of marketing strategies based on datasets.

II. LITERATURESURVEY

[1] Customer Segmentation based on Behavioural Data in E-marketplace

An important marketing strategy that is widely used by businesses is customer segmentation. The point of customer segmentation is to split the user-base into smaller groups that can be targeted with specialized content and offers. The produced customer groups are drawn from user behaviour data which gives the business a deeper understanding of the types of users that exists in the system. The benefit of customer segmentation is twofold. Firstly, a better knowledge about the types of users in a system can lead to better business and marketing strategies. Secondly, a user is likely to use an application more often if he/she always receives relevant content. To be able to create a set of similar customer groups, an extensive analysis of the available data combined with research and evaluation of clustering algorithms is needed. The available data is the most vital part of any clustering algorithm. The most important aspects are the quality and amount of the available data. In order to run some sort of similarity function to cluster items or users in a system, the data needs to be arranged into feature vectors with a set of feature values. To achieve the best results, a large amount of data is needed and more importantly the absence of data points needs to be minimal. Another important aspect in customer segmentation is to understand the available data. In a system where items are rated using some sort of scale, e.g. a rating from zero to five, it is fairly easy to interpret a user’s preferences. However, in systems where the set of items is not predefined, as in a E-marketplace where users upload items which are removed when sold, it is much harder to determine a user's preference.

[2] Customer Segmentation Using Clustering and Data Mining Techniques

The market segmentation is a process to divide customers into homogeneous groups which have similar characteristics such as buying habits, life style, food preferences etc. Market segmentation is one of the most fundamental strategic planning and marketing concepts wherein grouping of people is done under different categories such as the keenness, purchasing capability and the interest to buy. The segmentation operation is performed according to similarity in people in several dimensions related to a product under consideration. The main objective of market segmentation is accurately predicting the needs of customers and thereby intern improving the profitability by procuring or manufacturing products in right quantity at time for the right customer at optimum cost. To meet these stringent requirements k-means clustering technique may be applied for market segmentation to arrive at an appropriate forecasting and planning decisions. It is possible to classify objects such as brands, products, utility, durability, ease of use etc. with cluster analysis. For example, which brands are clustered together in terms of consumer perceptions for a positioning exercise or which cities are clustered together in terms of income, qualification etc.

[3] Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services

The thrust of this paper is to identify customer segments in a retail business using a data mining approach. The customer segmentation process consists of 3 stages, namely customer segmentation, big data, clustering and k-Means algorithm. Each of stage uses different methods. Customer segmentation is the subdivision of a business customer base into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. This segmentation is based on factors that can directly or indirectly influence market or business such as products preferences or expectations, locations, behaviours and so on. Big data as "the word describing the large volume of both structured and unstructured data, which cannot be analyzed using traditional techniques and algorithm." Companies capture trillions of bytes of information about their customers, suppliers, and operations, and millions of networked sensors are being embedded in the physical world in devices such as mobile phones and automobiles, sensing, creating, and communicating data. Clustering algorithms include k-Means algorithm, k-Nearest Neighbor algorithm, Self-Organizing Map (SOM) and so on. According to, clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). opined that clustering algorithms generate clusters having similarity between data objects based on some characteristics. Clustering is extensively used in many areas such as pattern recognition, computer science, medical, machine learning. In this paper, the k-Means clustering algorithm has been applied in customer segmentation. A MATLAB program (Appendix) of the k-Means algorithm was developed, and the training was realized using z-score normalized two feature dataset of 100 training patterns acquired from a retail business. After several iterations, four stable clusters or customer segments were identified. The two features considered in the clustering are the average amount of goods purchased by customer per month and the average number of customer visits per month. From the dataset, four customer clusters or segments were identified and labeled thus: High-Buyers-Regular Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular Visitors (LBIV). Furthermore, for any input pattern that was not in the training set, its cluster can be correctly extrapolated by normalizing it and computing its similarities from the cluster centroids associated with each of the clusters.

[4] Marketing Segmentation Through Machine Learning Models

Customer relationship management (CRM) aims to build relations with the most profitable clients by performing customer segmentation and designing appropriate marketing tools. Statistical techniques, such as cluster and principal component analysis (PCA) combined with discriminant analysis (DA) or logistic regression, have been traditionally used for building segmentation models, but the existence of large volumes of data linked to multiple-correlated features reduce the real fit, robustness, and interpretability of these models. To develop a complete CRM program, several stages need to be completed. From there, the initial segmentation of clients and the design of metrics for measuring the success of the CRM program are particularly critical, despite several limitations about their development. In addition, customer profitability accounting (CPA) recommends evaluating the CRM program through the combination of partial measures in a global cost–benefit function. Several statistical techniques have been applied for market segmentations although the existence of large data sets reduces their effectiveness. As an alternative, decision trees are machine learning models that do not consider a priori hypotheses, achieve a high performance, and generate logical rules clearly understood by managers. The development of a CRM program needs to fulfill several successive stages, customer segmentation and the measurement of the success of the CRM program being two of the most critical steps.

III. IMPLEMENTATION

System Implementation is the stage where the theoretical design is converted into a working system, the new system may be totally new, replacing an existing manual, or automated system or it may be a major modification to an existing system. The system is implemented using Anaconda a standard platform and coded using jupyter notebook.

Anaconda

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system conda The Anaconda distribution includes data-science packages suitable for Windows, Linux, and MacOS. Anaconda distribution comes with 1,500 packages selected from PyPI as well as the condapackage and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command line interface (CLI). The big difference between conda and the pip package manager is in how package dependencies are managed, which is a significant challenge for Python data science and the reason conda exists.

Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glue
- Orange
- RStudio
- Visual Studio Code

In our Project we use Jupyter Notebook as an Anaconda Navigator.

Jupyter Notebook

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that we can also use. A code cell allows you to edit and write new code, with full syntax highlighting and tab completion. The programming language you use depends on the kernel, and the default kernel (IPython) runs Python code. When a code cell is executed, code that it contains is sent to the kernel associated with the notebook. The results that are returned from this computation are then displayed in the notebook as the cell's output. The output is not limited to text, with many other possible forms of output are also possible, including matplotlib figures and HTML tables (as used, for example, in the pandas data analysis package). This is known as IPython's rich display capability.

Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Van Rossum led the language community until stepping down as leader in July 2018. Python features a dynamic type system and automatic memory management. It supports multiple programming, including object-oriented, imperative, functional and procedural. It also has a comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit

Python Software Foundation.

Methods for Customer Segmentation

A. Collect data

The dataset is collected from UCI machine learning repository. The dataset contains 8 features Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen, Channel, Region. Divide this wholesale customer data based on machine learning clustering algorithms. Python programming language is used for implementing clustering algorithms.

B. Methods of customer classification

There are many ways to partition, which vary in severity, data requirements, and purpose. The following are some of the most commonly used methods, but this is not an incomplete list. There are papers that discuss artificial neural networks, particle determination and complex types of ensemble, but are not included due to limited exposure. In future articles, I may go into some of these options, but for now, these general methods should suffice. Each subsequent section of this article will include a basic description of the method, as well as a code example for the method used. If you do not have the expertise, well, just skip the code and you have to get a good handle on each of the 4 sub-sections included in this article.

C. Group analysis

Group analysis is an integration or unification, approach to consumers based on their similarity. There are 2 main types of categorical group analysis in market policy: hierarchical group analysis, and classification (Miller, 2015). In the meantime, we will discuss how to classify groups, called k-methods. D. K. Means encounter The K-means clustering algorithm is an algorithm often used to draw insights into formats and differences within a database. In marketing, it is often used to build customer segments and understand the behavior of these unique segments. Let's try to build an assembly model in Python's environment.

D. Centroids initiation

Selected cents or initials were selected. Figure 1 introduces the beginning of graduate centers. The four selected centers, shown in different sizes, were selected using the Forgi method. In Forgy's method, data points are randomly selected as cluster centroids using k (k = 3 in this case)

Procedure for Customer Segmentation

Step 1: Start
Step 2: Input dataset
Step 3: Data pre-processing
Step 4: Principal component analysis
Step 5: Elbow Method
Step 6: K-means
Step 7: Visualizing customer Segments
Step 8: Stop

K-means Algorithm

Step 1: Start
Step 2: Select the number K to decide the number of clusters.
Step 3: Select random K points or centroids. (It can be other from the input dataset).
Step 4: Assign each data point to their closest centroid, which will form the predefined K clusters.
Step 5: Calculate the variance and place a new centroid of each cluster.
Step 6: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
Step 7: If any reassignment occurs, then go to step-4 else go to FINISH.
Step 8: The model is ready.
Step 9: Stop.

Flow Diagram

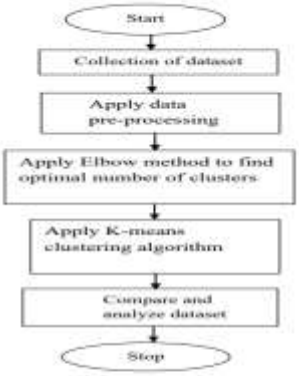
Start
Collection of dataset
Apply data pre-processing
Apply Elbow method to find optimal number of clusters
Apply K-means clustering algorithm
Compare and analyze dataset
Stop

Figure 1: Flowchart for Customer Segmentation using Machine Learning

Figure Shows the flowchart of Customer Segmentation. A system flowchart symbolically shows how data flows throughout a system. Initially the dataset is input, and the pre-processing of the input dataset takes place followed by elbow method to find optimal number of clusters. Later the dataset is classified according to kmeans algorithm. Finally, the different type of customer is detected.

IV. EXPERIMENTATION AND RESULT

Implemented clustering algorithms in python. Machine Learning models can be implemented in python and R programming. Python has selected because python provides various number of packages for implementing machine learning algorithms. The dataset is taken form UCI data repository. The dataset contains 440 instances with eight features. As excluded two features namely “Channel” and “Region” from analysis as they are not useful for clustering.

Implementation of K-Means The major decision in any clustering algorithm is choosing number of clusters. For finding optimal number of clusters need to use a separate process. So, first need to find the optimal number of clusters before proceeding for K-means clustering. For this, used Elbow method. But for implementing elbow method, there is no package in python. Python provides K-means class for K-means implementation. As used K-means class and manually written code for Elbow method-means class has three parameters namely number of clusters, number of iterations, random state. The first parameter is the optimal number of clusters. This value can be obtained from elbow method.
The plot of elbow method was given below.

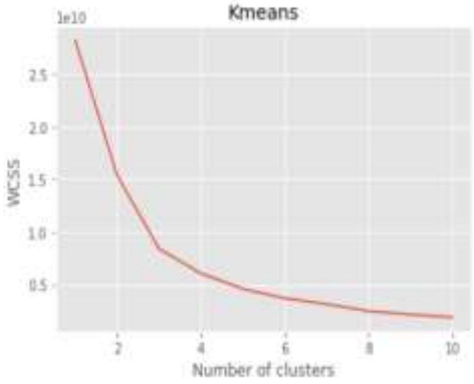

Figure 2: Result of elbow method

Since Customer Segmentation is unsupervised learning algorithm, it will not have accuracy score instead it has WCSS(Within-Cluster-Sum-of-Squares).

From the Figure 2, it is observed that, there is less variation from 2 or 3 to 10.So,This can select optimal number of clusters as 3. Next step is applying K-means clustering algorithm. As applied K-means algorithm and the resulted are plotted using python matplotlib library.

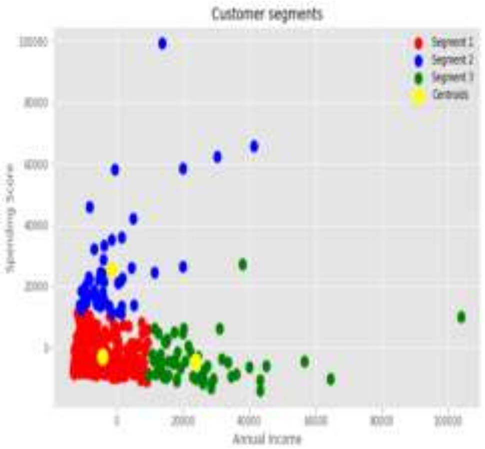

Figure 3: Clusters of data points using K-means

In the figure 3, yellow color points are centroids. Divide the data points into 3 clusters showing with three different colors red, blue, green. Segment 1 contains medium spenders, they tend to spend on Grocery, Milk, Detergents_Paper. Segment 2 contains high spenders, who spending on Fresh, product category. Segment 3 is low spenders.

V. CONCLUSION AND FUTURE WORK

In this paper, the dataset of Wholesale customers is analysed. The dataset is grouped into three clusters by both K-means algorithms. Segment 1 contains medium spenders, they tend to spend on Grocery, Milk, Detergents_Paper. Segment 2 contains high spenders, who spending on Fresh, product category. Segment 3 is low spenders. K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the data set. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

The result of any machine learning project is directly linked to the model and the data at hand. As shown in the result; having more data and features does not always improve a model, but better data and better features certainly do. Since the sample size was limited in this project, more data to train the model could in fact improve the performance; especially if more features are added. Adding more features, or at least further improving the current features could increase the performance of the model. Therefore it is of interest to investigate if more suitable features, describing user behaviour can be aggregated from the datasource. It could also be of interest to explore if unsupervised machine learning methods could be useful for this case, as several other studies on customer segmentation have gotten great results with clustering methods.

REFERENCES

[1]. Jean-Patrick Baudry, Margarida Cardoso, Gilles Celeux, Maria JosÃ© Amorim, Ana Sousa Ferreira (2012). Enhancing the selection of a model-based clustering with external qualitative variables. RESEARCH REPORT NÂ° 8124, October 2012, Project Team SELECT. INRIA Saclay - ÃŽle-de-France, Projet select, UniversitÃ© Paris-Sud 11

[2]. Laha Ale, Ning Zhang, Huici Wu, Dajiang Chen, and Tao Han, Online Proactive Caching in Mobile Edge Computing Using Bidirectional Deep Recurrent Neural Network, IEEE Internet of Things Journal, Vol. 6, Issue 3, pp. 5520-5530, 2019.

[3]. Andrew Aziz, "Customer Segmentation based on Behavioural Data in E-marketplace", Examensarbete 30 hp, August 2017, https://uu.diva portal.org/smash/get/diva2:1145508/FULLTEXT01.pdf.

[4]. K.Biruntha1, R.Porkodi, "Use of Data Mining Techniques to Improve the Effectiveness of Sales and Marketing" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653, Volume 6 Issue I, January 2018.

[5]. Weblink:https://archive.ics.uci.edu/ml/datasets/Wholesale+customer