Travel Review Ratings

Rajaraman Ganesan
Master Engineering
Electrical & Computer Dept.
Western University, London
rganesa@uwo.ca

Vatsal Shah
Master Engineering
Electrical & Computer Dept.
Western University, London
vshah56@uwo.ca

***Abstract*-- Advancement in technology has fundamentally changed how information produced and consumed by all users involved in travel. Travelers can now access different sources of information, and they can generate their own content and share their views and experiences. Reviews shared through online has become a very influential information source that affects travel in terms of both reputation and performance. However, the volume of data on the Internet has reached a level that makes manual processing almost impossible, demanding new analytical approaches. Clustering technique considered as a possible solution to limit the volume of data. In travel review, based on users rating on social media say Google, users clustered to a group of various interest. After pre-processing the dataset, the problem reduced to Clustering problem. Here, we make use of clustering algorithm techniques like the k-mean algorithm. We conclude by evaluating results and compared with custom models and available libraries in python such as Sklearn.**

Keywords-- Clustering models, online review, k-means, Cluster evaluation, tourism, online review, Sklearn, python

I. Introduction

In the age of e-commerce, every industry is involved in online sales, and the hospitality and tourism industry are no exception. The participatory nature of the Internet in recent years has led to an explosive growth of travel-related user-generated content. Travel planning has become one of important commercial use. Sharing on the web has become a major tool in expressing customer thoughts about a product or Service. Many tourists look for some places like fun malls, restaurants or vacation spots, etc. online in recent times [1]. After consumption customers give feedback/rating, online so online reviews have become increasingly important. They are fast, updated and available everywhere and have become the word-of-mouth of the digital age. Thus, online review plays a critical role in the tourism industry, which mainly offers services and focuses on customer satisfaction. This is the main reason people spend time online reading the review/rating backing their decision-making.

In this study, user ratings are captured from Google reviews across the Europe region and average rating ranges from 1 to 5. The dataset contains information on 25 variables, obtained from the UCI Machine Learning Repository. With these reviews, we can make a good decision about the places about to visit, nature of the user.

A. Background & Motivation

The reason for choosing the topic is to find the best places the people can visit. We process the data provided analyze and clusters the range of rating provided by the tourist. The goal of this project is to resolve this problem by building and comparing various techniques using unsupervised learning algorithm. Moreover, to encounter the difference in the process and issues by applying the custom model and pre-built libraries.

B. Aim & Objective

The objective of the problem is to cluster the range of ratings provided by various consumers in various places they have visited. This project is helpful for solving problems using KNN, k-medoids, fuzzy C-means and find the average ratings given by users on different places. We implement by applying various methods, custom model and pre-built libraries in python by understanding the process, compare and evaluate results.

The steps followed to manage these goals:

1. Understand the selected dataset.
2. Display some graphical information and visualize the features.
3. Data pre-processing.
4. Apply clustering algorithms on the selected dataset using a custom model
5. Apply algorithms using pre-built libraries
6. Evaluate the model.
7. Compare the model and find the optimal one.

C. High Level Overview

The major purpose of this is to observe the ratings provided by users in various places. Clustering algorithms implemented to analyze this model. Many methods of clustering including soft clustering been used to develop this model of users review.

The remainder of the paper organized as follows. Section 2 summarizes the basic properties of applied models, section 3 explores the methodology with data preprocessing. Section 4 comprises the evaluation process and section 5 presents a summary.

II. BACKGROUND & LITERATURE SURVEY

The research work of [2] presents a comparison of three different datasets gathered from travel and tourism domain. The first dataset has 249 user records with 6 attributes, seconds dataset has 980 user records with 10 attributes and the third dataset has 5456 user records with 24 rating attributes. They applied various clustering techniques such as k-means, k-medoids, and CLARA and Fuzzy c-means using R packages. In the end, they concluded the k-means algorithm performed better than other clustering algorithms.

There is much research on review rating because every people believe the rating provided online and plan his or her visit to those selected places. Many statistical methods have been applied to develop a travel review rating model, using, K-means, soft clustering algorithms.

A. *Clustering:*

Clustering is a set of observations into subset in the same cluster are similar in some sense. In the world of machine learning, it is an unsupervised approach. Unsupervised learning applied while there is input data, but there are no corresponding output variables associated with it. The objective of clustering is to find different groups within elements in the data. Clustering algorithms find the structure in the data so that elements of the same cluster are more like each other than to those from different clusters. It has manifold usage in many fields such as machine learning, pattern recognition, image analysis, information retrieval, bio-informatics, data compression, and computer graphics [2].

B. *K-means Algorithm:*

K-means is one of the simplest algorithms uses unsupervised learning method to solve clustering issues. It is mainly used when u have unlabeled data. The goal of this algorithm is to find groups in data, with the number of groups represented by the variable.

$$WCV(C_k) = \sum_{x_i \in \mu_k} (x_i - \mu_k)^2 \qquad [3]$$

Where,
x_i = entity belong to cluster C_k
C_k & μ_k = mean of all entities forming the cluster.

The following steps followed to implement the k-means algorithm:

1. Determine the total no: of clusters to be formed.
2. Start by identifying centroids initially and random sampling done within the dataset to find initial points.
3. Calculate the distance between the entity and each of centroids and assign the entity to the cluster close to the centroids.
4. Revise the cluster centroid by calculating mean values of all entities. Repeat this for each cluster.
5. Reduce the total within-cluster variation. In addition, repeat the above step 3 and 4 until max number of iterations reached.

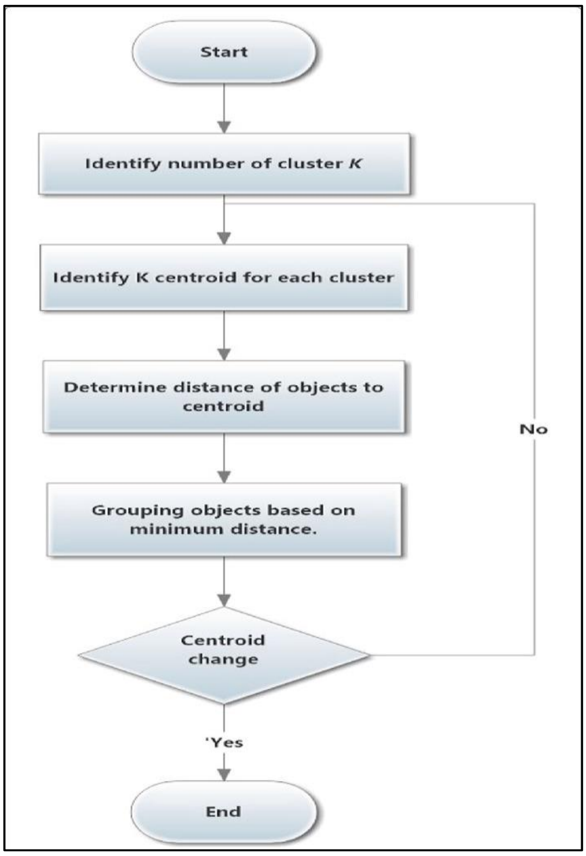
Start
Identify number of cluster K
Identify K centroid for each cluster
Determine distance of objects to centroid
Grouping objects based on minimum distance.
Centroid change
No
'Yes
End

Fig (1.1) Process for k-means algorithm

III. METHODOLOGY

In methodology, data description, independent variable, and dependent variable described with the scale of variables. Moreover, in the process data preprocessing and feature engineering described below.

A. *Data Description & Preparation:*

This dataset consists of 5456 total instances and 25 features including:

TABLE I
ATTRIBUTES OF THE DATASET

Attribute No	Description
1	Unique user ID
2	Average ratings on churches
3	Average ratings on resorts

4	Average ratings on beaches
5	Average ratings on parks
6	Average ratings on theatres
7	Average ratings on museums
8	Average ratings on malls
9	Average ratings on zoo
10	Average ratings on restaurants
11	Average ratings on pubs/bars
12	Average ratings on local services
13	Average ratings on burger/pizza shops
14	Average ratings on hotels/other lodgings
15	Average ratings on juice bars
16	Average ratings on art galleries
17	Average ratings on dance clubs
18	Average ratings on swimming pools
19	Average ratings on gyms
20	Average ratings on bakeries
21	Average ratings on beauty & spas
22	Average ratings on cafes
23	Average ratings on viewpoints
24	Average ratings on monuments
25	Average ratings on gardens

There are 24 categories reviewed by users. It is shown in the below rating ranger from 1 to 5. Each categories has various number of users rating. To get better understanding of the data, visualization with hist and plot diagram as below:

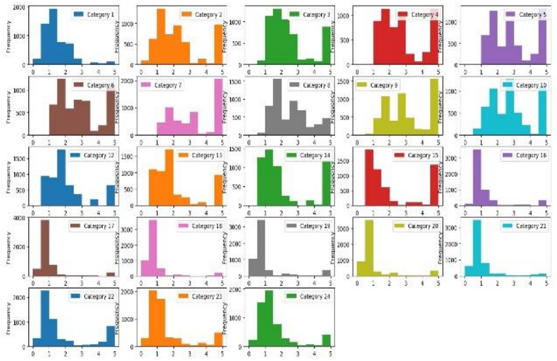

Fig (1.2) Visualization of data with hist diagram

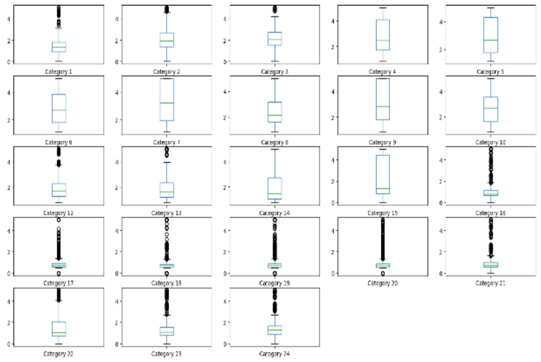

Fig (1.3) Visualization of data with box-plot diagram

B. *Data Clensing & Normalizing:*

Mostly, data gathered from various sources, which have missing values and noises. Because of this, data cleaning is an important step to perform before applying an algorithm. There are various approaches for data cleaning. In Python, isna() used to identify null values of all attributes and isna().sum() gives the summery of all attributes with the number of null records. It can be replaced with mean value. In python, it can be implemented as fillna() function which replaces all missing values with the mean value.

Scaling or normalizing of the attribute is a practice [2] can be performed in clustering problem. By performing normalization all, the attributes are in the same range.

C. *Tools Used:*

The main tool used for the implementation is Python 3.6 programming language, which is free open source. There are multiple Python packages with statistical computing such as NumPy, pandas. For data representation and visualization such as matplotlib and seaborn. Moreover, for standard machine learning algorithms such as Sklearn libraries. It used to evaluate results using pre-built libraries.

IV. EVALUATION

The objective of a project is to identify the optimal number of clusters along with the best approach to apply KMeans algorithms. The objective of a clustering algorithm is to accomplish minimum inter-cluster similarity among clusters and maximum intra-cluster similarity within each cluster [4][5].

A. *Elbow Method:*

A method of interpretation and validation of consistency within cluster analysis designed to help to find the appropriate number of clusters in a dataset [6][7].

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

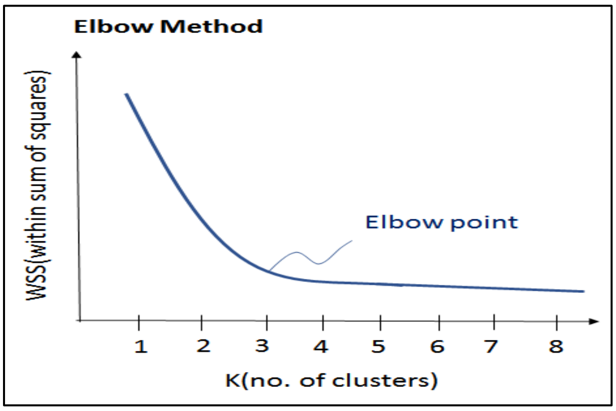

Fig (1.4) Elbow method graph represents WSS vs No. of clusters

B. *Accuracy Measures:*

There are various accuracy measures to finalize the results. We will use accuracy measures as define below to conclude the best accurate model while comparing pre-built libraries and custom models.

*1) **Accuracy:** The accuracy of a model is usually determined after the model parameters are learned and fixed and no learning is taking place. Then the test samples are fed to the model and the number of mistakes (zero-one loss) the model makes is recorded, after comparison to the true targets. Then the percentage of misclassification is calculated. In our dataset accuracy determine how often the model predicts defaulters and non-defaulters correctly.*

*2) **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. If the precision is the high then there will be low false positive rate. Here precision tells us that whenever our models predicts it is defaulter how often it is correct.*

*3) **Recall:** Recall is the ratio of correctly predicted positive observations to the all observations in actual class. In other words, out of all positive class how much we have predicted correctly. When we apply this in our dataset it shows the actual defaulters that the model will predict.*

*4) **Loss function:** Loss functions let the optimization function know how well it is doing. Loss functions are used in the output layer, Layers that support unsupervised layer-wise pretraining.*

V. SUMMARY & FUTURE WORK

We studied the data, checking for data unbalancing, preparation, visualizing features and understanding relationships between various features. We are building the train model with K-means algorithm from python libraries and custom model. Then selected the best cluster value k, compared and checked with elbow graph. Finally, conclude the results and ended with best one.

VI. REFERENCES

[1]. Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). "A comparative study of efficient initialization methods for the k-means clustering algorithm". Expert Systems with Applications.

[2]. Renjith, Shini, and C. Anjali. "A personalized mobile travel recommender system using hybrid algorithm." In Computational Systems and Communications (ICCSC), 2014 First International Conference on, pp. 12-17. IEEE, 2014

[3]. Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.

[4]. Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence 24, no. 12 (2002): 1650-1654.

[5]. Kovács, Ferenc, Csaba Legány, and Attila Babos. "Cluster validity measurement techniques." In a 6th International symposium of hungarian researchers on computational intelligence. 2005.

[6]. David J. Ketchen, Jr; Christopher L. Shook (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique". Strategic Management Journal.

[7]. Trupti M. Kodinariya, Dr. Prashant R. Makwana. “Review on determining number of Cluster in K-Means Clustering”. In International Journal of Advance Research in Computer Science and Management Studies. Volume 1, Issue 6, November 2013

CONTRIBUTIONS

Rajaraman Ganesan: He worked on data preparation and understanding of clustering algorithms. He applies pre-built libraries in python like Sklearn and train model. His role is to find best value of cluster, compare and check with elbow graph. At the end, evaluate the best results after applying pre-built libraries.

Vatsal Shah: He worked on dataset understanding, data pre-processing. His role is to understand and apply k-means algorithms with custom models. It includes create equations, find distance and select best value of cluster K number and evaluate results. Compare the results with of prebuilt libraries accuracy with custom models and conclude the best model.