PRAIRIE VIEW A&M UNIVERSITY
DATA MINING
Report on

Comparing the performance of K-Means and Hierarchical clustering analysis on Online Retail Dataset

Name: Uboho Victor T. **Instructor: Dr. Lin Li**
Student ID: P21443445

1. Introduction

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, differentiating its properties and results.

These models are distinguished by their organization and type of relationship between them Since this is a very valuable data analysis technique, it has several different applications in the sciences world. Every large data set of information can be processed by this kind of analysis, producing great results with many distinct types of data.

One of the most important application is related to image processing. detecting distinct kinds of pattern in image data. This can be very effective in biology research, distinguishing objects and identifying patterns. Another use is the classification of medical exams.

2. Problem Description

The aim of the project is to carry out model evaluation and performance measurement on two clustering algorithms (K-Means and Hierarchical) on an online retail dataset.

Comparing the performances of different clustering algorithms is critical in machine learning as there always lies the need to find the best model that represents our data and how well the chosen model will work in the future.

This is particularly important because the performance of different cluster algorithms vary with respect to the given dataset and the associated features of each sample or tuple in the dataset and therefore, it is necessary to compare their performances before making any model selection.

3. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Dataset: Online Retail Dataset

Data Set Information: This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Set Characteristics:	Multivariate, Sequential, Time-Series	**Number of Instances:**	541909	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2015-11-06
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	272968

Source: Dr Daqing Chen, Director: Public Analytics group. chend **'@'** lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

Attribute Information:
InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice: Unit price. Numeric, Product price per unit in sterling.
CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

The following steps were carried out in preprocessing the dataset:

1. **Importing the libraries:** Some of the libraries used for the data preprocessing includes NumPy, Pandas, Sklearn, Seaborn, Time, etc.
2. **Splitting the Dataset:** Due to the limited computational power, I used 4% of the dataset (20,000 samples) for the clustering.
3. **Importing the data:** Pandas was used to read the data into the memory as a dataframe. Also, the alphabetic characters were removed from the “InvoiceNo” column as to convert the column to Numeric.

4. **Covert categorical features to numeric:** Pandas was also used to convert the categorical features such as 'Country' and 'InvoiceDate' to numeric features. Furthermore, the 'Description' column was not required for our clustering, hence was removed from the dataframe.
5. **Normalizing the Data:** Two different techniques were used to normalize the data. Firstly, we used the *stats* feature of the *scipy* module to generate the *zscore* of columns with numeric features for the clustering. The second normalization was mainly for the purpose of visualizing the data after clustering, for this, we used the preprocessing feature of *sklearn* module.

4. Clustering Models

K-MEANS Clustering

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.

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

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

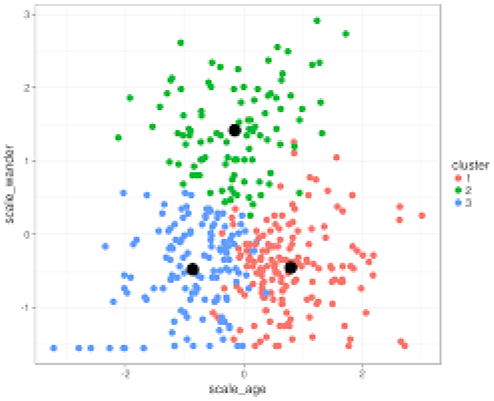

Fig. 1: KMeans clustering with 3 clusters

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Hierarchical Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

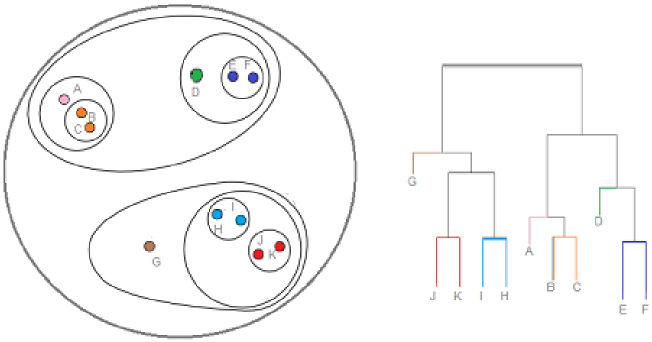

Fig. 1: Hierarchical clustering

How They Work

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering.

5. Experimental Results

For the purpose of the experiment, as earlier stated, only 20,000 samples of our dataset was used for the clustering. I ran the experiment using different parts of the split dataset on the clustering models to help select a more balanced dataset for our experiment.

Furthermore, I experimented using different cluster sizes (2,3,4,5 and 6) to find out which cluster size best suits our dataset and finally selected 3 clusters for both KMeans and Hierarchical clustering.

To visualize our results, we do a scatterplot of ‘UnitPrice’ by ‘InvoiceDate’ and then ‘CustomerID by ‘InvoiceDate’ for each of the clustering and the results are given below.

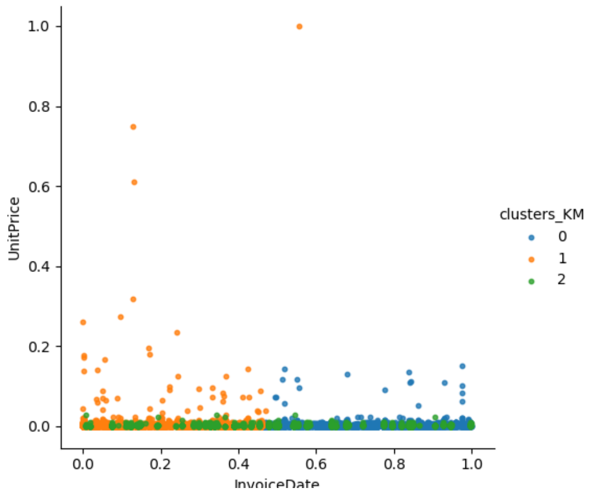

Fig. 3: KMeans Clustering with 3 clusters (UnitPrice x InvoiceDate)

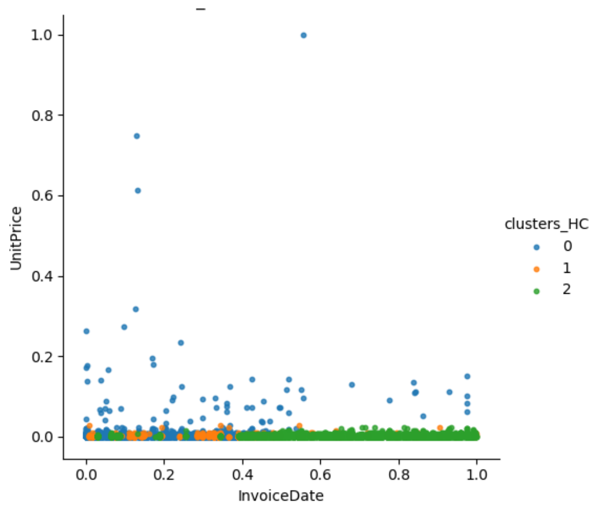

Fig. 4: hierarchical Clustering with 3 clusters (UnitPrice x InvoiceDate)

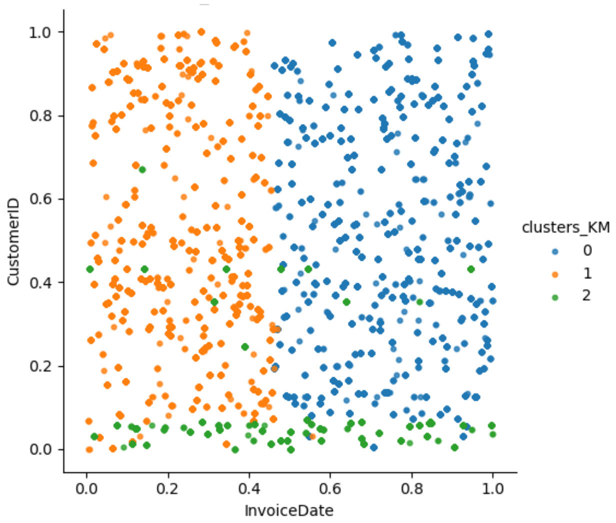

Fig. 5: KMeans Clustering with 3 clusters (CustomerID x InvoiceDate)

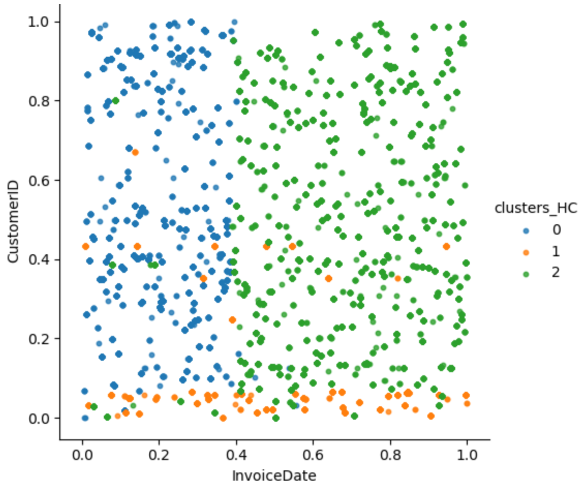

Fig. 6: Hierarchical Clustering with 3 clusters (CustomerID x InvoiceDate)

6. Conclusion

The results show that for the given dataset, the two clustering techniques produces similar work. It was also discovered that for visualization purposes, it is important that when plotting the scatter plot graph, the features used in both axes is important in order to better visualize and evaluate the performance of the clustering models.

For future work, I will research and implement different metrics for evaluating clusters such as cluster cohesiveness, completeness and cardinality

References

Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: *A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management*, Vol. 19, No. 3, pp. 197â€“208, 2012.

www.bigdatamadesimple.com *What is Clustering in Data Mining*, 2015

J. B. MacQueen (1967) *"Some Methods for classification and Analysis of Multivariate Observations*, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

S. C. Johnson, 1967. *"Hierarchical Clustering Schemes"* Psychometrika, 2:241-254