K-means Clustering Algorithm Based on E-Commerce Big Data

Indivar Shaik, Swapna Suhasini Nittela, Trayabak Hiwarkar, Srinivas Nalla

Abstract: ***As the technology improving, huge volumes of different types of data is being generated rapidly. Mining such data is a challenging task. One of the important tasks of mining is to group similar objects or similar data into cluster which is very much useful for analysis and prediction. K-means clustering method is a popular partition based approach for clustering data as it leads to good quality of results. This paper focuses on K-means clustering algorithm by analyzing the E-commerce big data. In this research, geographical location and unique identification number of the customer are considered as constraints for clustering.***

Keywords: Big data, Clustering, E-commerce, K-means.

I. INTRODUCTION

E-commerce is an application area where clustering algorithms are widely used. E-commerce websites continuously generates different types of data. Analyzing such data is really a challenging task. Clustering algorithms helps to find out the hidden similarities between the data, such as similarity between the price of the product or services provided to the purchasing history of the customers. It helps to group the customers with the similar patterns. [1]

Clustering of data is a method of grouping data into particular patterns or a method for classifying the information mountain into meaningful stacks. The goal of the clustering method is to divide a dataset into multiple groups so that the resemblance within a group is greater than between the groups. The objective of the clustering is to find out from the given data the underlying intrinsic structure of each cluster. [2] The problem clustering can be found in many application areas, such as pattern classification, pattern recognition, data mining, and data compression. The concept of what organizes a vigorous cluster depends on the execution and many techniques of discovering clusters that are substance to different requirements, that are systematic and ad hoc. [3]

There are different approaches and algorithms are available to deal with clustering as stated in the earlier sections, such as Partition based, Density based, Hierarchical based and Model based approaches. There are different algorithms proposed within the above mentioned approaches such as k-means clustering, c-means clustering, DBSCAN, CLARA, BIRCH and Gaussian Mixture Model. [4]

These algorithms are widely used and are showed very promising performance in clustering and are broadly used in many applications. [4] K-means a partition based clustering approach which is broadly used in different application areas, but it is time taking and also final results of clusters is depends on the selection of the initial cluster size. [2]

The clustering comes under unsupervised learning process as the clusters of similar objects form automatically. We can cluster anything, and the better our clusters are, the more similar items are in the cluster. This concept of similarity depends on measurement of the similarity. Because there is no target value like in Classification. This is the reason why it is known as unsupervised learning. For example, from a customer dataset by using unsupervised learning we can retrieve the top geographic locations based on their zip code or we can retrieve the top selling products that are bought by the customers in the particular group. [4]

The similarity and dissimilarity between the objects is calculated by measuring the distance between them. There are different methods available for measuring distance, such as Euclidean distance, Manhattan and Minkowski distance. [5]

Different clustering algorithms are available and they can be categorized as follows:

Partitioning based approach: This method breaks down a dataset into a collection of clusters. It builds several different partitions of this dataset, where each partition represents a cluster. K-means clustering, K-medoids clustering or Partition Around Medoids(PAM) and Clustering Large Applications (CLARA) algorithms are the widely used partition based clustering methods. [6] [7]

Density based approach: In this approach a cluster continuous to grow as long as it's neighborhood density meets or exceeds some threshold.

Hierarchical based approach:

i) Agglomerative approach (bottom up): In this approach each object is treated as a single unique cluster. Once all single clusters are formed, clusters are paired and combined successively until all clusters are merged into single cluster that contains all dataset items. [6] [7]

ii) Divisive approach (top down): It begins with all objects in the dataset being included in a single big cluster. The most heterogeneous cluster is split into two at each stage of iteration. The process is repeated until all the objects of dataset are in a cluster of their own. [6] [7]

Grid based approach: It quantizes the objects into grid cells

Distribution or Model based approach: It uses statistical distribution models such as expectation maximization (EM) analysis to fit objects.

Revised Manuscript Received on August 10, 2019.

Indivar Shaik*, Research Scholar, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.

Swapna Suhasini Nittela, Research Scholar, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.

Dr. Tryambak Hiwarkar, Professor, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.

Dr. Srinivas Nalla, Principal, Sahaja Institute of Technology & Science for Women, KarimNagar, Telangana, India.

II. K-MEANS CLUSTERING

K-means clustering is considered to be a very famous and widely used partitioning based approach. It is a numerical, non-deterministic, an unsupervised and an iterative approach. In K-means clustering, the average value (mean value) of objects in the group represents individual cluster. The main objective of the K-means clustering algorithm is to acquire the stable number of clusters that reduces the Euclidean distances between data objects and center of clusters. [8]

Let $D=\{d_i\}$, i=1,2,……,n is a dataset of n-objects, and let K-is the numbers of clusters to be formed.
C_j is the center of the cluster where j=1,2,…..,k.

The distance between centroid and data object is measured by using Euclidean distance formula. The Euclidean distance is calculated by using sum of squares of distances [1]. Let A and B are two objects in a dataset. The Euclidean distance between A and B is calculated by using the below formula.

Euclidean Distance (A, B) = $(|A_1-B_1|^2+ |A_2-B_2|^2 + \ldots\ldots\ldots\ldots +|A_{n-1}-B_{n-1}|^2+|A_n-B_n|^2)^{1/2}$

It can be summarized as; Euclidean distance $=\sqrt{\Sigma|A_i-B_i|^2}$, i=1,2,…….n-1,n.

Each cluster is filled with the objects which are near to the center of the cluster in the data space of cluster nodes. Later, center of the cluster updated by itself as the number of objects increases. This process is repeated until it reaches all the objects of cluster. [9]

For various real-world applications, the k-means clustering algorithm has been shown to be efficient in generating excellent cluster outcomes. Generally, k-means clustering algorithm's required time is proportional to the number of clusters per every iteration and to the number of patterns generated. This method is particularly very expensive for big datasets. [9]

Algorithm: [9][10]
Input: Given dataset D with n-items to cluster D= $\{d_1, d_2, d_3,\ldots\ldots,d_n\}$

K- Number of clusters (random selection)
K is set of subset of D as temporary cluster
K= $\{k_1, k_2, k_3,\ldots\ldots\ldots,k_k\}$
Where $k_1=\{d_1\}$, $k_2=\{d_2\}$, $k_3=\{d_3\}\ldots\ldots$ $k_k=\{d_k\}$

C is set of centroids of those clusters.
C= $\{ c_1, c_2, c_3,\ldots\ldots\ldots,c_k\}$
And $c_1 = d_1$, $c_2 = d_2$, $c_3 = d_3,\ldots\ldots\ldots$ $c_k = d_k$,
Where k<=n

Output: K is set of subset of D as final cluster and C is set of centroids of these clusters.

K= $\{k_1, k_2, k_3,\ldots\ldots\ldots\ldots\ldots, k_k\}$
C= $\{c_1, c_2, c_3,\ldots\ldots\ldots\ldots\ldots, c_k\}$

Steps:
Step 1: Initialize the cluster size. Divide the given dataset into k number of clusters where k-is predefined.
Step 2: Select k-points as cluster centers on a random basis.
Step 3: Calculate mean or centroid of all data objects in each cluster. Use Euclidean distance to measure the distance between cluster centers and each data object.
Step 4: Assign each data object to its nearest cluster center with a minimum distance from the cluster center.
Step 5: Recalculate center of the cluster and repeat Step 3
Step 6: If there is no data object reassigned then stop, else repeat from step 3

III. EMPIRICAL ANALYSIS AND RESULTS

The experiment was conducted by R programming language and R-Studio. The dataset was collected from UCI Machine Learning Repository which consists of 541909 records of the transactions. Centroids are selected randomly in K-means clustering approach. K-means clustering algorithm takes numerical values, so all the objects which we would like to cluster should be numeric variables. As the dataset is from online retail market, this paper focuses on the sales of the products from different countries (i.e.; geographical). It clusters the available countries into different clusters with respect to the sales of the product.

The purpose of this research is to form the clusters of the E-Commerce big data using k-means clustering algorithm. The main constraint of k-means clustering algorithm is that it accepts numerical variables.

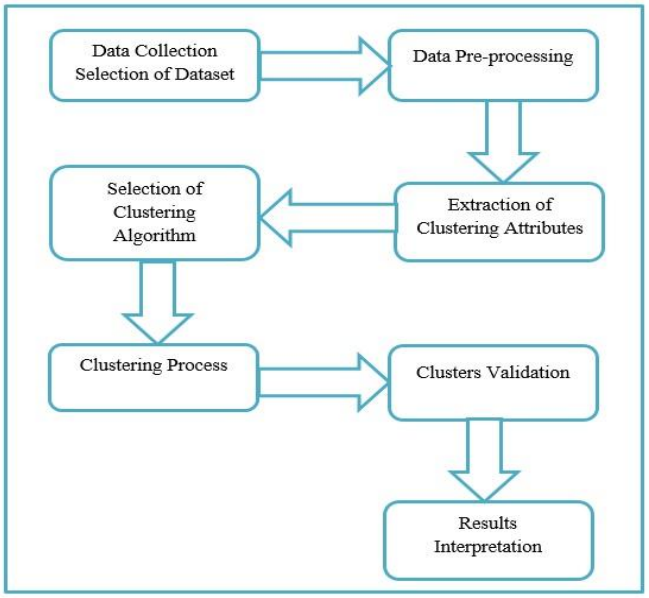

Figure 1: Clustering Process

In this research clustering has performed based on Country and based on CustomerID. We have calculated total number of invoices made by each country and each customer, total quantity of sales and finally calculated total revenue generated by each country and each customer. First performed k-means clustering based on Country and results were analyzed followed by k-means clustering based on CustomerID.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.10	13047	United Kingdom
12	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.10	13047	United Kingdom
13	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3.75	13047	United Kingdom
14	536367	22310	IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom
15	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/1/2010 8:34	4.25	13047	United Kingdom
16	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/1/2010 8:34	4.95	13047	United Kingdom
17	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/1/2010 8:34	9.95	13047	United Kingdom
18	536367	21754	HOME BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
19	536367	21755	LOVE BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
20	536367	21777	RECIPE BOX WITH METAL HEART	4	12/1/2010 8:34	7.95	13047	United Kingdom

Figure 2: Contents of Dataset

K-means clustering based on Country: The dataset was filtered according to the k-means clustering algorithm constraints. First we found the different unique countries available in the dataset. Totally 37 countries were present and some transactions were made from the unknown location, we treat them as unspecified location and also we included it.

	Frequency	TotalQuantity	TotalSales
Australia	1259	83653	137077.27
Austria	401	4827	10154.32
Bahrain	19	260	548.40
Belgium	2069	23152	40910.96
Brazil	32	356	1143.60
Canada	151	2763	3666.38
Channel Islands	758	9479	20086.29
Cyprus	622	6317	12946.29
Czech Republic	30	592	707.72
Denmark	389	8188	18768.14

Figure 3: Country wise Dataset after Preprocessing

We considered 3 different variables for clustering the market. They are frequency of the customer (how many times transactions happened from the particular country), total quantity purchased from the particular country and finally total amount of sales made by the particular country. Initially number of clusters are considered as 3 (As we used k-means clustering algorithm for clustering, need to give number of cluster as input).

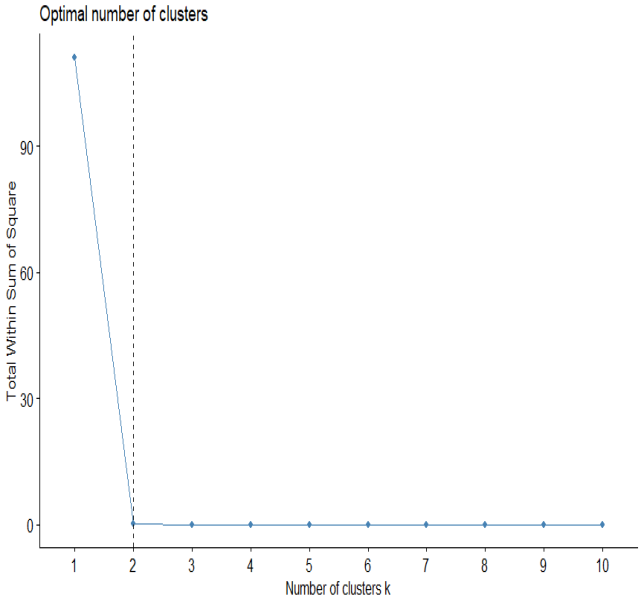

Figure 4: Country wise Optimal number of clusters

When we performed clustering using k-means clustering algorithm using 3 clusters as the input. It gave results as the sizes of 3 clusters are 32, 1 and 5. Among the available 38 countries 32 countries are grouped into cluster 1, only one country in cluster 2 and 5 countries in cluster 3. When considering the percentages of clusters cluster 1 consists of 84.21%, cluster 2 consists of 2.6% and cluster 3 consists of 13.15%. we can conclude that k-means clustering algorithm provided good result as cluster 1 consists of majority of countries in it.

```
> km.result$size
[1] 32  1  5
```

Figure 5: Country wise Cluster size

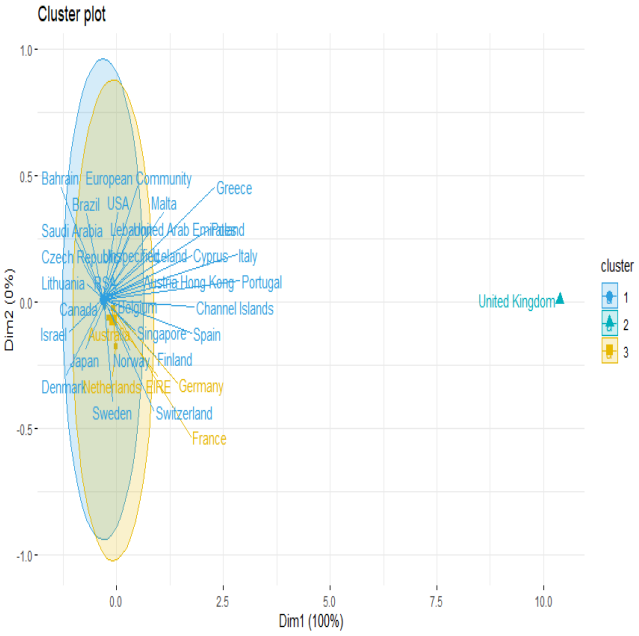

Figure 6: Country wise Visualization of Clustering

K-means clustering based on Country: First we identified the different unique customers available in the dataset. Totally 4372 different customers with unique id were present We considered 3 different variables for clustering the market as we performed in the previous section.

	Frequency	TotalQuantity	TotalSales
12346	2	0	0.00
12347	182	2458	4310.00
12348	31	2341	1797.24
12349	73	631	1757.55
12350	17	197	334.40
12352	95	470	1545.41
12353	4	20	89.00
12354	58	530	1079.40
12355	13	240	459.40
12356	59	1591	2811.43
12357	131	2708	6207.67
12358	19	248	1168.06
12359	254	1612	6245.53
12360	129	1165	2662.06
12361	10	91	189.90
12362	274	2212	5154.58

Figure 7: CustomerID wise Dataset after Preprocessing

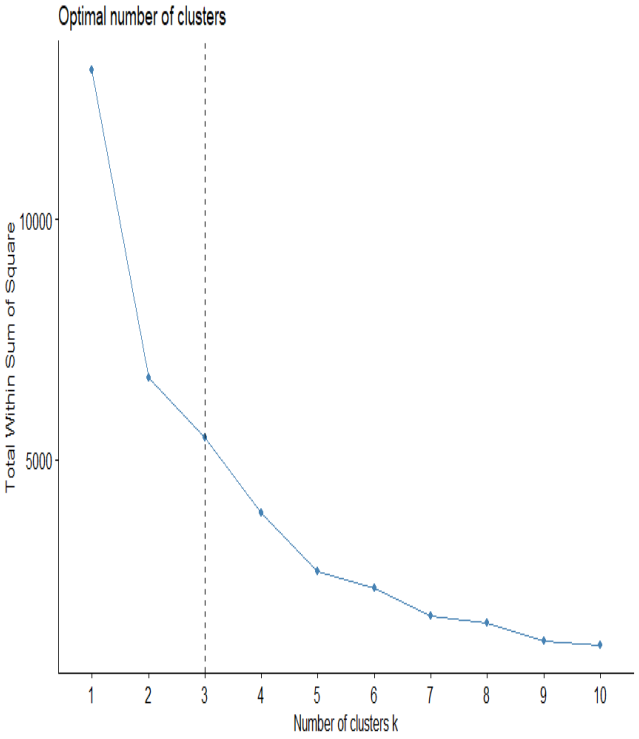

Figure 8: CustomerID wise Optimal number of clusters

When we performed clustering using k-means clustering algorithm using 3 clusters as the input. It gave results as the sizes of 3 clusters are 317, 4040 and 51. Among the available 4372 customers 317 customers are grouped into cluster 1, 4040 customers are in cluster 2 and 15 customers are in cluster 3. When considering the percentages of clusters cluster 1 consists of 7.25%, cluster 2 consists of 92.4% and cluster 3 consists of only 0.34%. we can conclude that k-means clustering algorithm provided good result as cluster 2 consists of majority of countries in it.

```
> km.result1$size
[1]  317 4040   15
```

Figure 9: CustomerID wise Cluster size

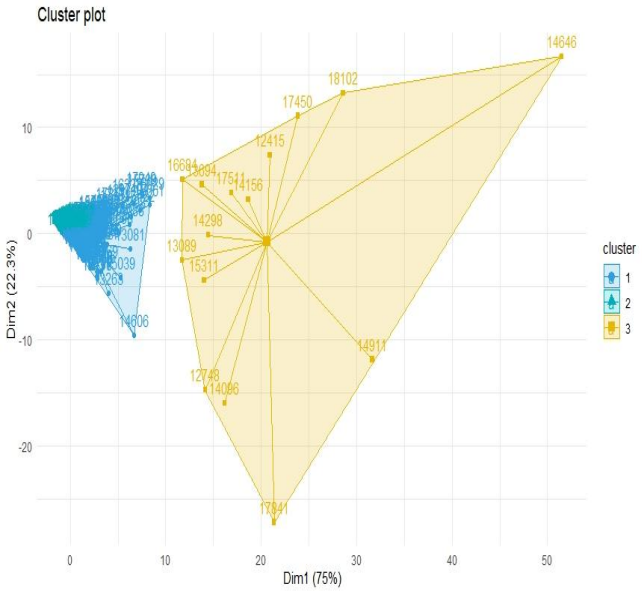

Figure 10: CustomerID wise Visualization of Clustering

IV. CONCLUSION

In this research, we have discussed a simple k-means clustering algorithm with an appropriate example. This paper analyzes the E-commerce dataset by using the k-means clustering algorithm. In this paper we have used a qualitative approach to analyze the clustering algorithm and we have used Euclidean distance for measuring the distance between objects. We demonstrated the k-means clustering algorithm on an E-Commerce dataset which consists of 541909 records of transactions. As per the theoretical analysis and results obtained from the experiment in this research, k-means algorithm delivers better efficiency for clustering huge amounts of data. A drawback of k-means clustering algorithm is that it uses fixed number of clusters. It can be further analyzed by applying different clustering methods.

REFERENCES

1. Shaik, I., Hiwarkar, T., & Nalla, S. (2019). "Customer Segmentation Analysis of E- Commerce Big Data". *International Journal of Engineering and Advanced Technology*. Vol.8 Issue.5, pp. 582–587.
2. Singh, S. P. & Yadav, A. (2013). "Study of K-Means and Enhanced K-Means Clustering Algorithm". *International Journal of Advanced Research in Computer Science*, Vol.*4 Isuue.*10. pp. 103-107
3. Kanungo, T., & Netanyahu, N. S. (2002). "An Efficient k-Means Clustering Algorithm: Analysis and Implementation". *IEEE Transactions on pattern Analysis and Machine Intelligence*, Vol.24 Issue.7, pp. 881-892
4. Wang, S., Li, M., Hu, N., Zhu, E., Hu, J., Liu, X., & Yin, J. (2019). "K-Means Clustering With Incomplete Data". *IEEE Access*, *7*, pp. 69162–69171.

5. Yadav, J., & Sharma, M. "A Review of K-mean Algorithm". (2013). *International Journal of Engineering Trends and Technology*, Vol.*4 Isuue*.7, pp. 2972–2976.
6. Patel K.M.A & Thakral.P, "The best clustering algorithms in data mining," *2016 International Conference on Communication and Signal Processing(ICCSP)*,Melmaruvathur,2016,pp.2042-2046. doi: 10.1109/ICCSP.2016.7754534
7. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). "Clustering algorithms: A comparative approach". In *PLoS ONE* (Vol. 14). https://doi.org/10.1371/journal.pone.0210236
8. Ali, H. H., & Kadhum, L. E. (2017). "K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition". *International Journal of Science and Research (IJSR)*, Vol.*6 Issue*.8, pp.1577–1584. https://doi.org/10.21275/ART20176024
9. Singh,S., & Gill, N.S. (2013). "Analysis and Study of K-Means Clustering Algorithm". *International Journal of Engineering Research & Technology*. Vol.2 Issu.7, pp. 2546-2551
10. Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). "Application of k Means Clustering algorithm for prediction of Students Academic Performance". *International Journal of Computer Science and Information Security*.Vol.*7 No.1*,pp. 292–295.