# Finding Best Possible Number of Clusters using K-Means Algorithm

**K. Maheswari**

***Abstract:*** ***Customers are assets for business. The companies are investing more for customer relationship management. Retaining customer for long time is a difficult process in today's trend. On line shopping is also increasing day by day. People are more interested to visit popular web sites and they are spending very less time to choose their products. On line shops are paying more interest to analyze customer preferences, their needs, shopping behaviors through data mining technique. Proper classification is necessary for organizing such data. In this work, Customer with the same buying behavior is grouped based on the features age and salary. K-Means algorithm is applied to form clusters with different K values for original data and normalized data. The within sum of square (wss) is calculated for both the data for different cluster size. The minimum wss is considered to be better which is achieved in normalized data. The validity of cluster is evaluated by elbow, silhouette and gap statistic method to choose the optimal number of clusters. This work is implemented in R software.***

***Keywords : Cluster, Customer Purchase, K-Means and WSS.***

## I. INTRODUCTION

Clustering is an unsupervised learning Algorithm. It deals unlabeled data. It is used to find the group similarity. The important part of clustering is measuring the distance between data points. Clustering permits to find the hidden relationship between data points. Intra cluster minimization and inter cluster maximization will be used for creating good clusters.

Clustering analysis is a process of Identifying groups. The clusters used to find the relationship between variables. The process involved is

1. Preparing a dataset.
2. Preprocessing of data.
3. Generating clusters of the data.
4. Interpreting results –validating clusters

The data set used in this work is downloaded from the internet. The source of data set is (https://github.com/atse0612/Machine-Learning:A=Z/blob/master/Social_Network_Ads.csv)

If the data is good and clean then the models can be constructed easily, so that the performance will increase. To make the data sensible and extract meaningful value from unstructured data, classification [13] and clustering are used. Clustering is one of the techniques [9] [11] to sort and organize the data in to logical grouping before analysis process starts. Clustering is also performed based on the distances. The shape of the cluster is influenced by distance measure only. The distance measures are Euclidean measure, Minkowski and Manhatten measure.

Euclidean distance is calculated by

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \qquad (1)$$

Minkowski distance is calculated by

$$D(x,y) = \sqrt[p]{(\sum_{i=1}^{n} | xi - yi |^2)} \qquad (2)$$

Manhatten distance measure is calculated by

$$D(x,y) = |xi - yi| \qquad (3)$$

Many techniques are available for scaling the data. They are mean re center, median and rescaling between 0 and 1. The standardization is the process of adjusting the range of values between 0 and 1 or between -1 and +1.

Re center is a method which performs standard z score transformation. The mean is subtracted from each data point. Median is the middle value. It separates the higher half value and lower half value. Scaling is a process of multiplying each data point by a constant value to alter the range of values. The units of data point may be different numerical form. The result may be affected. The algorithm takes the value numerically. It can't differentiate whichever is higher. Age value 60 is more important than the salary 5000 per month.

This paper is organized as follows. Chapter I focused introduction of this paper. Chapter II deals with Review of Literature. Chapter III describes Methodology used in this work. Chapter IV presents the experimental results and chapter V concludes the work.

**Revised Manuscript Received on December 16, 2019.**
***** Correspondence Author
**Dr.K.Maheswari***,Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankovil-Virudhunagar , India. Email: maheswarisnr@gmail.com

## II. REVIEW OF LITERATURE

Clustering allows to find similarity and differences exist between datasets. The clustering performed with minimum wss between data and its cluster centroid is more preferable. This is achieved by K-Means algorithm. The following is the study of various methodologies, findings, drawbacks of various authors.

Vladim´ir Holy et al. [1] proposed a method for clustering products. The author framed optimization problem by applying genetic algorithm on simulated data and real data. Number of clusters is used as pa parameter. The author demonstrated and suggested that if the number of cluster is more than the number of categories then the method gives reverse results.

Elham Photoohi Bafghi [2] presented a method for classification. The customers are classified based on their shopping, financial status. Genetic algorithm is applied and accuracy is measured. Customers with the same buying behaviors were put it in same cluster.

Jayant Tikmani et al [3] described customer segmentation. K-Means clustering algorithm is applied on television company customers. The conclusion drawn from this paper is , among 80 customers, 13 customers strongly agreed about the necessity of television. 16 customers only agreed and 18 customers strongly disagreed. 4 clusters are formed and implemented in spss.

Yen-Chung Liu et al [4] implemented K-Means algorithm on sequence data. The customers are grouped based on their buying behavior. Therefore, the same customer may be in different clusters or multiple clusters or in no clusters. It was implemented in MATLAB.

Krzysztof Małecki [5] proposed an approach for decision support system. The author developed topology for on-line customers based on which web site they are visiting. The customers are classified in to information hunters, customer visit on-line stores and opportunity hunter. There is 95.9% variance among these four types of customers.

Chad West et al, [6] measured customer loyalty of a super market. The loyalty of a cutomer is measured by number of times the customer visits the super market and how much they are buying. The algorithms Kohonen neural network and K-Means were applied on supermarket transactional database. The author proposed scoring system, points assigned to each customer based on the purchase. This system does not show major difference between loyal and disloyal clusters. The modified scoring system was developed. The increased score was obtained under this scheme.

Rui Xu et al [7] surveyed clustering algorithms for various data sets. The data set is taken from computer science, statistics, and machine learning, The author discussed proximity measure and cluster validation. The author suggested that the choice of good features will reduce the burden in further design processing.

Vitor Campos et al [8] conveyed that the large volume of data handling is a difficult process in decision making by managers in the organization. A model is developed for classification and clusters for decision making process to make the work easy. The sale of vehicle parts and accessories data set is used for analysis. REPTree, JRip, PART, and J48 algorithms were applied for measuring accuracy. PART algorithm shows higher accuracy. Online shopping using SVM is discuused in [13].

Clustering is a good choice used to detect similarities of buying behavior of a customer.

## III. METHODOLOGY

K Means Clustering is an unsupervised machine learning algorithm. The output is not predicted in unsupervised learning algorithm. It finds similar patterns. The clustering [10] [12] of data is based on the similarity that occurs in the dataset. The k means algorithm finds number of clusters the dataset may be grouped into. For each row, the cluster number will be assigned randomly by the algorithm. The centroid of each cluster is determined. The following two steps are performed repeatedly until the within cluster sum of squares is minimized.

- Reassign data points to the cluster whose centroid point is very nearer.
- Determine new centroid for each cluster obtrained

The within cluster variation is calculated by between_SS / total_SS. X1,X2,…..Xj is a set of observations. The cluster variance is defined as the sum of the squared variations of mean of the cluster of all the rows in the dataset. The goal of clustering algorithm is minimizing the within cluster variance. Minimum value of cluster variance is more preferable. The data set is downloaded from the github (https://github.com/atse0612/Machine-Learning:A=Z/blob/master/Social_Network_Ads.csv).

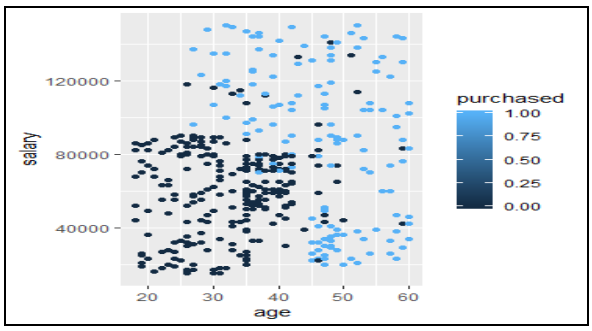


**Fig 1 purchase rate based on age and salary**

The plot for the attributes age and salary based on the product purchase is shown in figure 1. The people whose Age between 20 to 40 and the salary below 1,00,000 was not buy any product. Young people with lower purchase rate are shown in fig. Middle age and upper middle age people are purchasing the product even though they are getting low salary.

**Algorithm 1 : K-Means**

Input : Number of Clusters K, Distance D
Output : K clusters of the dataset

1. Select K clusters for the dataset
2. Assign the centroids randomly
3. For each data point, calculate the closest centroid.
4. Assign the point to the cluster
5. Set each cluster position to the mean of all data points assigned to the cluster
6. Step 4 and 5 is repeated until no more changes.

The output of K-Means algorithm is listed as Cluster, centers, totss, withinss, tot.withinss, betweenss and size. Cluster [14] [15] is a Grouping of similar characters. Cluster centers are represented in matrix form. Total sum of squares is totss. within cluster sum of squares is known as withinss. sum of withinss is tot.withinss. Sum of squares of Between cluster is betweenss. Number of data points in the cluster is defined to be its size.

## IV. EXPERIMENTAL RESULTS

After cleaning process is over, the clustering algorithm k-means is applied on the dataset. The algorithm is applied for various K values. For 4 clusters, the within cluster sum of square is 92.8% with cluster sizes 71,134,87 and 108. For cluster 3, the wss is 87.2% and 64.4% for cluster 2. The Table 1 shows number of cluters ,cluster size and wss before scaling the data. The minimum the wss will be the better for further processing. Table 2 shows the within cluster sum of square value after scaling.

**TABLE 1 WITHIN CLUSTER SUMOF SQUARE – BEFORE NORMALIZING**

| Number of Clusters | Cluster size | Within cluster sum of squares by cluster |
|---|---|---|
| 4 | 71, 134, 87, 108 | (between_SS / total_SS = 92.8 %) |
| 3 | 74, 185, 141 | (between_SS / total_SS = 87.2 %) |
| 2 | 227, 173 | (between_SS / total_SS = 64.4 %) |

**TABLE 2 WITHIN CLUSTER SUMOF SQUARE – AFTER NORMALIZING**

| Number of Clusters | Cluster size | Within cluster sum of squares by cluster |
|---|---|---|
| 4 | 158, 81, 66, 95 | (between_SS / total_SS = 69.2 %) |
| 3 | 139, 81, 180 | (between_SS / total_SS = 59.5 %) |
| 2 | 108, 292 | (between_SS / total_SS = 36.9 %) |

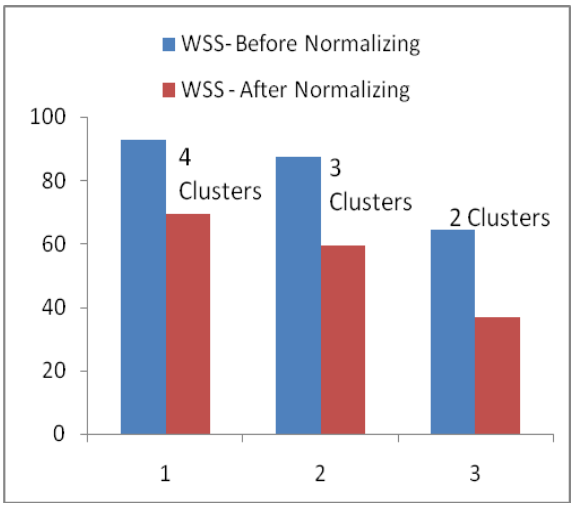


**Fig 2 WSS before normalizing and after normalizing**

The need for normalizing data set is necessary for some problems. Some feature value may range from 10 to 100 other feature may range from 10000 to 100000. The range will vary for features used in the dataset. These different ranges are used for classification or clustering, The greater numeric range impacts more than the lesser numeric range. To improve accuracy, the dataset is normalized or scaled to bring the common range. Table 3 shows cluster means for age and salary before performing scale() function in R. Figure 2 shows the comparison of wss value before normalizing and after normalizing.

**TABLE 3 CLUSTER MEANS BEFORE SCALING**

| Cluster | Age | Salary |
|---|---|---|
| 1 | 42.47887 | 126338.03 |
| 2 | 35.97015 | 80485.07 |
| 3 | 37.26437 | 26873.56 |
| 4 | 36.88889 | 53740.74 |

**TABLE 4 CLUSTER MEANS AFTER SCALING**

| Cluster | Age | Salary |
|---|---|---|
| 1 | -0.1814230 | 0.0411494 |
| 2 | 0.6158721 | 1.5224765 |
| 3 | 1.2730284 | -0.8149687 |
| 4 | -1.1077966 | -0.8003607 |

Table 4 shows cluster means for age and salary after performing scale() function using the following formula

$$Z - Score = \frac{Data\ value - mean}{Standard\ Deviation} \quad (4)$$

The data set can be normalized using Min-Max. This is a normalization technique. The formula is

$$\text{Normalize} = \frac{(X - \min(X))}{\max(X) - \min(X)} \quad (5)$$

The disadvantage of min-max normalization is to bring data towards the mean. It does not handle outliers present in the dataset. So, z-score standardization is a better technique. The scale() function in R is used to achieve z-score standardization.

The elbow method scans wittiness cluster value for each K value. The procedure for the selection of optimal number of clusters is given in algorithm 2.The optimal value chosen by this algorithm is K=4 which is shown in figure 3.

Algorithm 2 : Elbow Method

Input : Number of Clusters K, Wss value
Output : Optimal K value

1. Generate clusters using k-Means algorithm for different K values.
2. For each K value, determine within cluster sum of square wss.
3. Plot the graph of wss for different K values.
4. The location in which there is a bend in the plot indicates the optimal number of clusres.

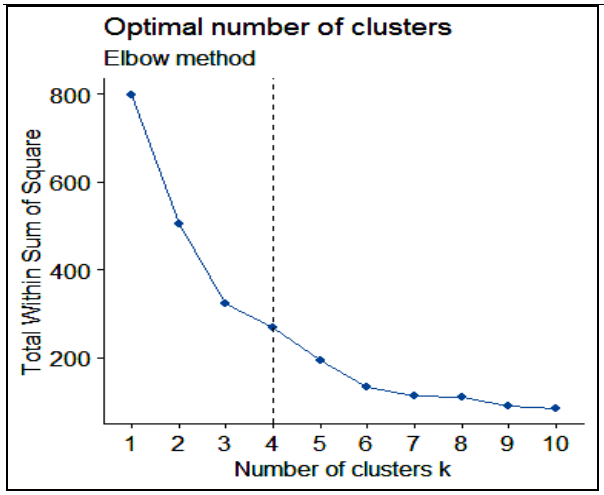


Fig 3 Finding number of clusters using Elbow Method

The drawback of elbow method is ambiguity. The alternative is Average silhouette method which measures the quality of clustering. It shows how the object lies within the cluster in figure 4.

Algorithm 3 : Average silhouette Method

Input : Number of Clusters K, average silhouette value
Output : Optimal K value

1. Generate clusters using k-Means algorithm for different K values.
2. For each K value, determine average silhouette.
3. Plot the graph of wss for different K values.
4. The location in which there is a maximum in the plot indicates the optimal number of clusters.

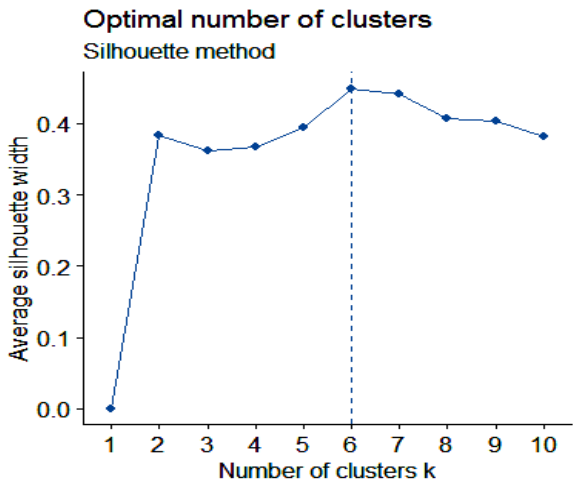


Fig 4 Finding number of clusters using silhouettle Method

The Gap static method compares the wss value for different K values and is shown in figure 5.

Algorithm 4 : Gap Static Method

Input : Number of Clusters K, Wss value
Output : Optimal K value

1. Generate clusters using k-Means algorithm for different K values.
2. For each K value, determine intra cluster variation .
3. Generate a data set with random uniform distribution .
4. Compute wss for this new data set value.
5. Calculate the gap between the wss for original data set with the new data set value.
6. Choose the minimum K value such that the gap statistics is within one standard deviation.

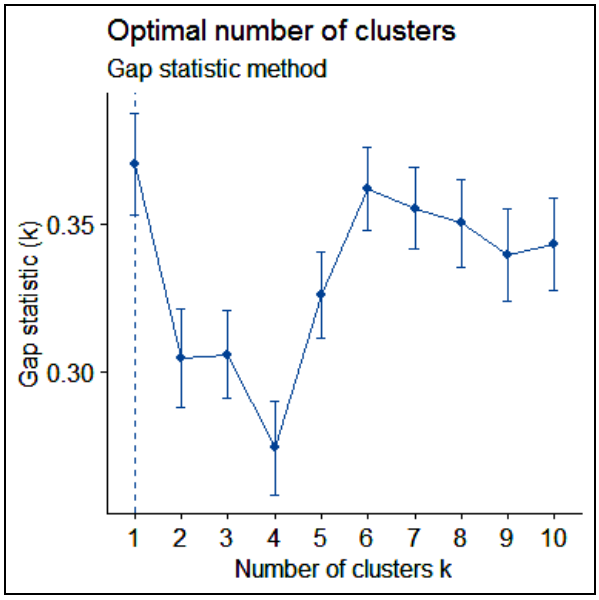


**Fig 5 Finding number of clusters using Gap Statistic Method**

The output of gap statistic method is
**********
Among all indices:
* 5 proposed 2 as the best number of clusters
* 10 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
***** Conclusion *****
* According to the majority rule, the best number of clusters is 3
**********
The best partition is shown in figure 6.

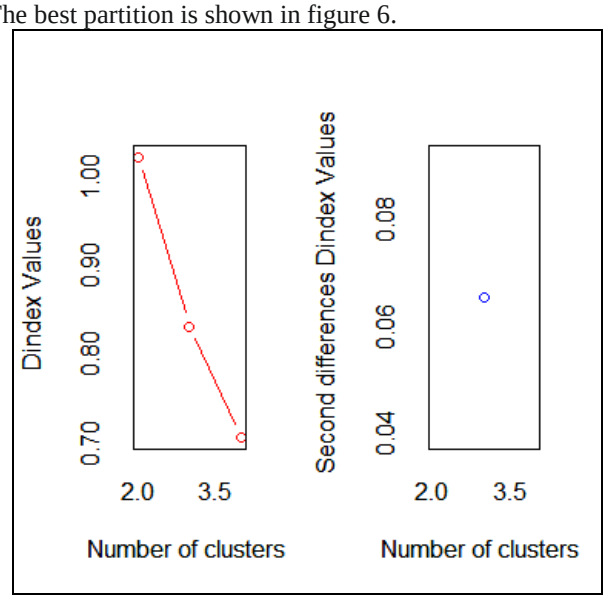


**Fig 6 Finding number of clusters using Dindex Value**

Different performance metrics were used for classification in machine learning algorithms. For clustering, there are two important things. Before Clustering – Number of Cluster is a required parameter for the clustering algorithm. After Clustering – Cluster validation – How good the created clusters are?

Various methods (Elbow, silhouette and gap statistic) are used to choose the optimal number of clusters. But each method gives different solutions. Elbow method suggests 4 clusters silhouette recommends 6 clusters, and gap statistic suggests 3 clusters. So, there is a need for selecting the best method for choosing optimal number of clusters. From the result in figure 6 using Dindex value, both the plots suggest to form 3 clusters.

## V. CONCLUSION

On line shopping does not allow the customers to touch the product. They have to attract the customers by showing images, photos and other displays. Data Mining is a most powerful tool to discover knowledge from the database. In this work, the data set is normalized to produce improved results. Minimum wss was calculated using scale () function in R software. Clustering was applied on normalized data set. In this paper, choosing of right number of clusters was performed. Future work concentrates more advanced algorithms to perform this work.

## REFERENCES

1. Vladimír Holý, Ondřej Sokol, Michal Černý,"Clustering Retail Products Based on Customer Behaviour", Applied Soft Computing, Elsevier Vol 60, PP: 752-762, 2017.
2. Elham Photoohi Bafghi," Clustering of Customers Based on Shopping Behavior and Employing Genetic Algorithms", Engineering, Technology & Applied Science Research Vol. 7, No. 1, 2017, 1420-1424.
3. Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar ,"An Approach to Customer Classification using k-means", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015.
4. 4. Yen-Chung Liu, Yen-Liang Chen," Customer Clustering Based on customer Purchasing Sequence Data", Int. Journal of Engineering Research and Application, ISSN : 2248-9622, Vol. 7, Issue 1, ( Part -1) January 2017, pp.49-58.
5. Krzysztof Małecki, Jarosław Wątróbsk ,"The Classification of Internet Shop Customers based on the Cluster Analysis and Graph Cellular Automata", International Conference on Knowledge Based and Intelligent Information and Engineering Systems,KES2017,6-8,Sep2017, Marseille, France, ScienceDirect , Procedia Computer Science .
6. Chad West, Stephanie MacDonald, Pawan Lingras , and Greg Adams," Relationship between Product Based Loyalty and Clustering based on Supermarket Visit and SpendingPatterns", International Journal of Computer Science & Applications, Vol. II, No. II, pp. 85 – 100, 2005.
7. Rui Xu, Donald Wunsch," Survey of Clustering Algorithms", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.
8. Vitor Campos , Carlos Bueno, Jacques Brancher, Fabio Matsunaga,Rafael Negrao," Knowledge Discovery Using an Integration of Clustering and Classification to Support Decision- making in E-commerce", Advances in Economics and Business 3(8): 329-336, 2015.
9. Ali Rezaeian, Sajjad Shokouhyar, Fariba Dehghan, "Measuring Customers Satisfaction of ECommerce Sites Using Clustering Techniques: Case Study of Nyazco Website", International Journal of Management, Accounting and Economics Vol. 3, No. 1, January, 2016.
10. Takanobu Nakahara, Takeaki Uno, Yukinobu Hamuro," Prediction model using micro-clustering", 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014, Elsevier, ScienceDirect, 35 ( 2014 ) 1488 – 1494.
11. Daqing Chen, Sai Laing Sain, Kun Guo," Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining", Springer, Volume 19, Issue 3, pp 197–208, Sep 2012.

12. Rachid Ait daoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb,"
Customer Segmentation Model in E- commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:9, No:8, 2015.
13..Dr.K.Maheswari,Ms.P.Packia Amutha Priya "Predicting Customer Behavior in Online Shopping Using SVM Classifier", presented paper in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization & Signal Processing, INCOS'17, published in IEEE
Xplore 01 March 2018. SCOPUS INDEXED.
14. Sriramakrishnan Chandrasekaran, Abhishek Kumar," A Clustering Approach for Customer Billing Prediction in Mall: A Machine Learning Mechanism", Journal of Computer and Communications, Scientific Research Publishing, 2019, vol: 7,PP:55-66.
15. Chandrasekaran, S. (2019) Deep Learning Approach for Customer Records Prediction for Future Products Issues. Journal of Computer and
Communications, 7, 44-54 https://doi.org/10.4236/jcc.2019.73005.

**AUTHORS PROFILE**

Dr. K. Maheswari received her B.sc (Computer Science) from Madurai Kamaraj University and MCA. M.Phil. from Bharathidasan University. She has completed her Ph.d at Bharathiar University. She is currently working as an Associate Professor in the Department of Computer Applications, Kalasalingam Academy of Research and Education. She has 23 years of teaching experience. She has presented research papers in several national and international conferences. She has published many research papers in various international journals. Her research interest is VoIP , network security and Data Mining.