TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 7844 - 7853

Combined K-Hierarchy Clustering to know the Buying Pattern of Customer and Provide Them with Freebies by Online Site for Future Shopping

Priyanka Desai
e-mail: desaipriyanka2002@yahoo.co.in

Bhavna Arora
e-mail: a12.bhavna@gmail.com

Article Info
Volume 82
Page Number: 7844 - 7853
Publication Issue:
January-February 2020

Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *04 February 2020*

Abstract:
The focus is on buying pattern of the customer based on the discount and its related quantity. Data is available in unsupervised form as the online data being received is not linear. The data has to be put in different chunks, it is not possible as the data has to be analysed, hence this is the gap. The solution is to form clusters first so as to segregate the data that is being received for which the following questions need to be answered What is the kind of data being received? Is the buying pattern related to the discount being offered for a particular quantity or viz. The cluster are formed using K-means cluster and Hierarchical cluster, this is then compared with proposed algorithm explained in the paper..

Keywords: *Unsupervised form, K-means Clustering, Hierarchical Clustering, Combined K hierarchy*

I. Introduction

Tracking and convincing online customers to buy a product being launched or for already available product is a task, hence need to study and come up with a good solution to cluster similar customers in one group. Take a look at online shopping data from 2016 to 2021 in figure 1. It is predicted that there is a rise in customer shopping on the internet in United States of America.

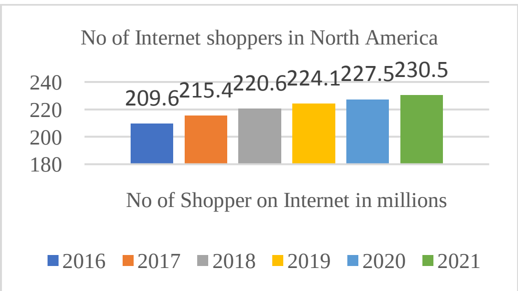

Figure 1:Online shopping data in North America

With the rise in customers there is also a rise in uncertainty. It is difficult to find a good lead. As is seen in the figure2, traffic and lead form 63% of the challenges faced by marketing team. The primary reason for shopping online is that they offer better rates than the retail market. It is more comfortable to shop sitting at home or while travelling, one avoids the crowd and queues in a shop, wide range of product availability that is not available in a retail, easy to find the product one is looking out to buy and more important is the availability to shop 24/7.

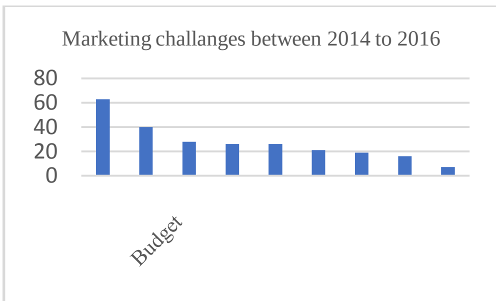

Figure 2: Finding good lead is a marketing challenge

In such a scenario there is a need to understand the customer shopping pattern and to cluster them into like groups so that the shopping web sites can provide them the deals that they are interested. The paper does not focus on how the data is collected? but instead focuses on the data available with the site. Here Kaggle Customer data is used for analysis. The solution is to group relevant data to form different groups combining K and hierarchy. Here use hierarchy to combine nearestgroups until a certain point and then use K to perform allocating proper groups.

The existing solution does not provide proper groups due to availability of similar groups leading to errors while forming groups. If a 3-year-old kid needs to be enrolled in a school then the child will be enrolled in nursery, a 4-year-old child being enrolled in Junior KG this is the norm and hence grouping here is simple. The online shopping scenario is nowhere similar to the kids being enrolled in school. As per banking norms any person between the age of 15 to 18 years can own a debit or credit card. In such a scenario it is difficult to actually segregate the shoppers by age as a 15-year-old can shop for the items a 35-year-old is shopping. Therefore, the focus of the paper is the discount provided on the number of items on a particular offer.

Groups or clusters is an unsupervised method used to form proper clustering first and then this data can be used for further supervised learning analysis. As there is a lot of data available for discounts on number of items, initially it's difficult to segregate them. Hence clustering them into proper groups is the need of the hour. This grouped data can then be used to find the proper clusters to avoid errors. This errors can be reduced by using the proposed method as shown in the results.

Using proposed combined K-hierarchy is helpful to achieve proper results. The errors are reduced as the verification and validation is done in the analysis phase itself rather than after the testing phase. The Software Development Life Cycle model has the following phases for development; requirement, design, analysis, coding, testing. Here train; gather customer data: requirement, design: find the drawbacks in the methods being considered for execution, detect k or hierarchy or why proposed combined k-hierarchy: analysis, proposed combined k-hierarchy: coding and finally test if the trained data give the required outcome.

The basic idea of the paper is to avoid error at any level of grouping or clustering this can be done by the proposed combined k-hierarchy explained in the paper.

II. Related Theory

In K [1,2,3,4] the initial clusters and the compared with the next iteration of clusters until the final iteration of clusters is equal to the iteration-1.

K-means clustering

I/P: $X = (x_1\, x_2\ \dots x_n)$

$Y = (y_1, y_2 \dots y_n)$

in $G=(g_1, g_2\ \dots g_k)$ // initial clusters

O/P: fi $G = (g_1, g_2\ \dots g_k)$ // final clusters

$L=l(x,y)$ where $x=1,2,\dots n$ and $y=1,2,\dots n$ with $g=1,2,\dots.k$ for clusters X and Y

Algorithm

For $G=(x_i, y_i \in G)$

in $g_i = (x_i, y_i) \in\ \subseteq G$

End

For $x_i \in g_i$ and $y_i \in g_i$

$$x_{g_i} = \frac{1}{n} x_i$$

$$y_{g_i} = \frac{1}{n} y_i$$

End

1:For$x_i \in X$ and $y_i \in Y$

$$l(x_i, y_i) = (x_i - x_{g_i})^2 + (y_i - y_{g_i})^2 \quad \forall g_i$$

= calculate $\min D \forall x_i, y_i \in g_i$

End

If in $G = G$ //if change cluster equals initial cluster

January - February 2020
ISSN: 0193 - 4120 Page No. 01 - 11

Converge =true
exit
else
Ite=0
For $x_i \in$ and $y_i \in$ Y and Converge=false
minD= calculate mind $\forall x_i , y_i \in g_i$
if mind $\neq l(x_i , y_i)$
itr++
goto 1
End
End
End

Hierarchical clustering

I/O: x_{g_i}, y_{g_i} in x_i , y_i
O/P: maxDx_{g_i}, y_{g_i}
Algorithm
For $x_{g_i}, y_{g_i} \in G$
Do
Matrix=Calculate minxD$\forall$G
End
For $g_i \in G$ in a matrix
If min$D_{g_i} \in g_i$
minG=compare and replace min (D_{g_i}, D_{g_i}') // find the min distance in the tuples, compare
// the tuples and consider the minimum value between two tuples
End
Until x_{g_i}, y_{g_i} has single cluster remaining

Papers usinghierarchical-k as solution methodology

This paper[5]selects initial groups randomly to generate final cluster allocation.

The paper[6] uses the combined approach for microarray datasets, but the methodology followed is not explained in detail.

III. Research Questions

1.What are the draw backs of K and hierarchy?

2. What is the need to use combined k-hierarchy to form clusters?

3. Will the combined K- hierarchy solve the errors that arise during grouping or clustering?

IV. Proposed Solution

Due to discrepancy in the outcome of K there is a need to compare it with Hierarchy. With better initial clusters K performs better, but hierarchy is a method that generates proper clusters; hence the need to combine the methods to generate better outcome without discrepancy.

Combined K Hierarchy

Due to availability of a large amount of dynamic data grouping data becomes an important aspect of data handling. Though there are a lot of off-the shelf tools available it is mandatory to understand the working of the clusters such that it helps the analyst to find the best method of analysis of the data. The focus of this paper is providing the best cluster method rather than guiding the best tool that can be used to generate the groups.

The method shows the input and the output needed for the formation of clusters so as to avoid major deviations in the result being generated. If there is a major deviation in the out come K generates errors leading to method stopping abruptly in any given scenario. This process of finding the best cluster should be calculated for every data set whether it is used for customer analysis, product analysis or clustering for a search engine. The bigger the data there is a possibility of major deviations.

This lead to finding the best initial cluster so that the following final cluster is generated without any hinderance; this is the reason Hierarchy is used to calculate the best cluster given an initial cluster followed by K to give a final cluster for a given dataset. Let's take a look at the flow chartin figure 3, followed by the algorithm. First select the data set to be used, here it's customer data set. Next prune the data for deals being offered and remove duplications in data; this will lead to generating best outcome. Find the clusters using the Euclidean distance to be used in hierarchy to find proper centroids using Hierarchy; if centroids are not separated properly then perform this method until a proper centroid is found. The centroids should be spaced at a proper distance such that the information should not overlap each other that may

Published by: The Mattingley Publishing Co., Inc.

lead to deviations in the final outcome. This data is then fed into K to check for the proper cluster formation. Initialize the first iteration as zero, now compare the outcome of the first iteration with the clusters that are allocated to the data sets in K. The initial cluster and the first iteration should be the same; this shows that the use of hierarchy was helpful in reducing the iterations thus saving time and nullifying the deviations and thus saving cost incurred to check the process all over again for the best outcome of clusters.

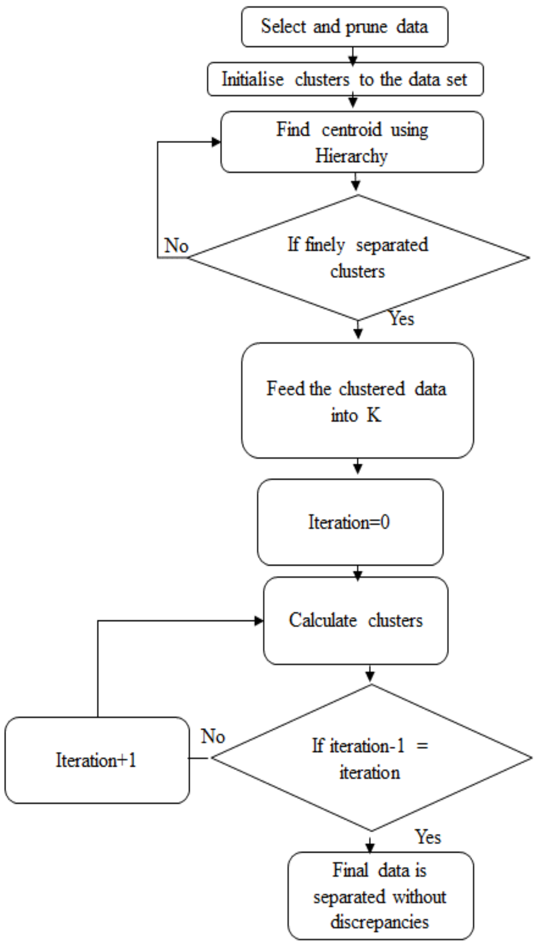

Figure 3: Flowchart for combined K Hierarchy

The flow chart explains the steps involved in processing the proposed method. It is self-explanatory that shows the need to use the pruned data set's where the cluster are allocated and not random clusters. These clusters are then calculated using the Euclidian distance which is used in the Hierarchy to form the centroids, the first two minimum distances are compared and merged together, next two minimum distances are compared and merged together. This step is continued until a proper maximum distance is achieved between the clusters and not until the formation of only one cluster. The reduced clusters are realigned to match the input data sets and are fed into K. The clusters are calculated where the initial frequency is '0', this is compared to the outcome of clusters. The initial clusters and the outcome clusters should match at the first or second instance as this will prove that the said method saves bot time and cost. Finally, the sum of squares for the cluster minimum needs to be found that shows the final sum of squares in smaller than the initial sum of squares.

4.2 Proposed algorithm:

I/P: X = $(x_1\ x_2\ \dots x_n\ )$

Y = $(y_1\ , y_2\ \dots y_n\ )$

in G=$(g_1\ , g_2\ \dots g_k\ )$ // initial clusters

O/P: fi G =$(g_1\ , g_2\ \dots g_k\ )$ // final clusters

L=l(x,y) where x=1,2,…n and y=1,2,…n with g=1,2,….k for clusters X and Y

Algorithm

For G=$(x_i, y_i\ Є G)$ // find the hierarch cluster this will avoid to errors in the K-mean calculation

in g_i = $(x_i\ , y_i\ )Є\ \subseteq G$

End

For $x_{g_i}, y_{g_i} Є\ G$

Do

Matrix=Calculate minxD∀G

End

For $g_i\ Є\ G\ in\ a\ matrix$

If minD_{g_i}Є g_i

minG=compare and replace min $(D_{g_i}, D_{g_i}{}')$

// find the min distance in the tuples, compare // the tuples and consider the minimum value between two tuples

End

January - February 2020
ISSN: 0193 - 4120 Page No. 01 - 11

End Until x_{g_i}, y_{g_i} has clusters that are well separated maxD and not until single cluster

For $x_i \in g_i$ and $y_i \in g_i$

$x_{g_i} = \frac{1}{n} x_i$

$y_{g_i} = \frac{1}{n} y_i$

End

1:For $x_i \in X$ and $y_i \in Y$

$l(x_i, y_i) = (x_i - x_{g_i})^2 + (y_i - y_{g_i})^2 \forall g_i$

= calculate minD $\forall x_i, y_i \in g_i$

SSq=min $\sum_{i=1}^{n} l(x_i, y_i)$ //initial sum of suqaures

inSSq=SSq

End

If in G = G //if change cluster equals initial cluster

Print Converge =true

Print SSq //final sum of squares

Compare(inSSq. SSq)

exit

else

Ite=1

For $x_i \in$ and $y_i \in$ Y and Converge=false

minD= calculate mind $\forall x_i, y_i \in g_i$

if mind $\neq l(x_i, y_i)$

itr++

goto 1

End

End

End

V. Experimental Results

The data used here is Kaggle customer data of 32 offers with a total of 325 transactions. The data is clustered using k-means clustering, hierarchical clustering and combined k hierarchy. The initial data uses hierarchical clusters to find the best cluster. These initial clusters are then fed into the k-means to given an outcome. Hence the name combined K-hierarchy. Better to find the optimal outcome at the start to avoid late detection of errors. As the saying goes detection is better than cure, that same is applicable to the algorithms that are merged to give better results for any given data. A look at section 4.2 that deals with the proposed algorithm with comments gives a better understanding of the experimental results that are taken care of in this section.

K-means clustering

There are four versions created for comparing. In Version1the data is pruned and the initial 5 clusters lead to errors which is shown in the final outcome leading to only three clusters. Version2 uses clusters as per allocation given by Kaggle; the initial clusters again give errors. Similar is the case with version3 and Version4.

Table 1:K-mearns Initial Clustering

Version 1 Initial	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	SSE
Centroid x	144	6.857143	30	60	104	18727.38
Centroid y	75.33333	53.57143	32.66667	58.5	63	
Version 2						
Centroid x	144	6	30	72	81	16280.08
Centroid y	75.33333	53.16667	32.66667	58.625	61.58333	
Version 3						
Centroid x	84	19.2	58.5	88.71429	10	34552.81
Centroid y	63.16667	62.4	58.5	59.14286	50.33333	
Version 4						
Centroid x	144	8	30	72	100.8	17635.198
Centroid y	75.333333	46.555556	56.666667	55.285714	64.9	

January - February 2020
ISSN: 0193 - 4120 Page No. 01 - 11

Table 2:K-mearnsfinal Clustering

Version 1 final	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Number of Iterations	SSE
Centroid x	144	8.181818	Err	72	Err	2	11950.56
Centroid y	51.42857	52	Err	65.64286	Err		
Version 2							
Centroid x	144	8.181818	#DIV/0!	72	72	2	7670.184
Centroid y	51.42857	52	#DIV/0!	50.5	85.83333		
Version 3							
Centroid x	72	Err	72	144	8.181818	6	7670.184
Centroid y	85.83333	Err	50.5	51.42857	52		
Version 4							
Centroid x	144	8.4	6	72	Err	2	10597.829
Centroid y	51.428571	48.5	87	65.642857	Err		

The number of cases in each cluster for each version is mentioned in the table. The table shows initial cluster allocation and final cluster allocation.

Table 3:K-mearns initial data allocation

Initial cases in each cluster	Version 1	Version 2	Version 3	Version4
0	3	3	6	3
1	7	6	5	9
2	3	3	4	3
3	10	8	14	7
4	9	12	3	10

Table 4 :K-mean final data allocation

Final cases in each cluster	Version 1	Version 2	Version 3	Version 4
0	7	7	6	7
1	11	11	-	10
2	-	-	8	1
3	14	8	7	14
4	-	6	11	-

The F tests should be used only for descriptive purposes shown in table 6, because the clusters have been chosen to maximize the differencesamong cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted astests of the hypothesis that the cluster means are equal.

Table 5:K-means Clustering using SPSS

ANOVA

	Cluster Mean Square	df	Error Mean Square	df	F	Sig.
MinimumQtykg	19965.810	4	3.394	27	5882.783	.000
Discount	1781.460	4	231.421	27	7.698	.000

The figures 4,5,6,7 below depict the cluster allocation.

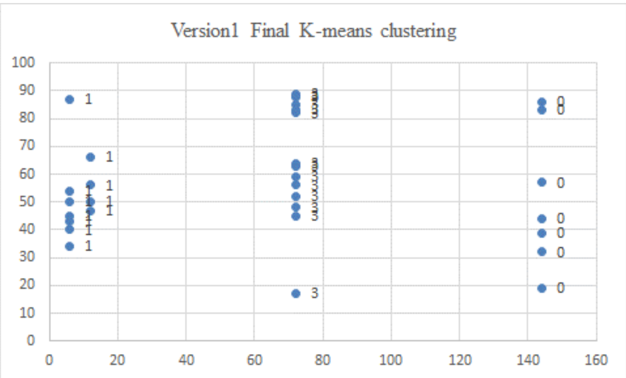

Figure 4: K-means cluster of table-2 version 1

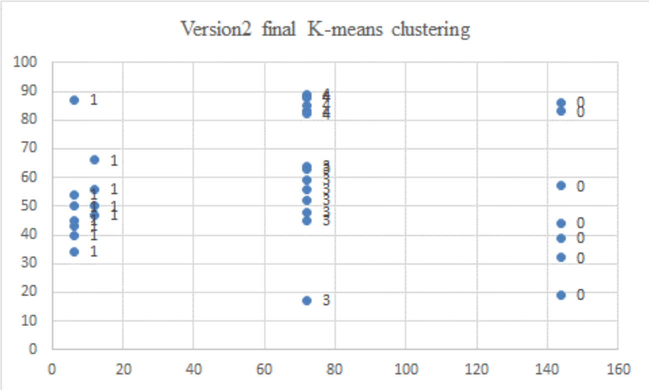

Figure 5: K-means cluster of table-2 version 2

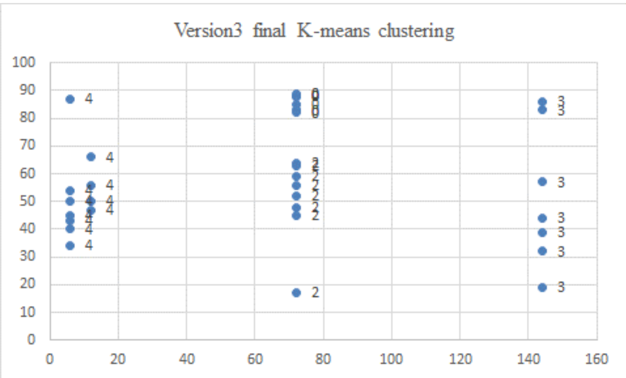

Figure 6: K-means cluster of table-2 version 3

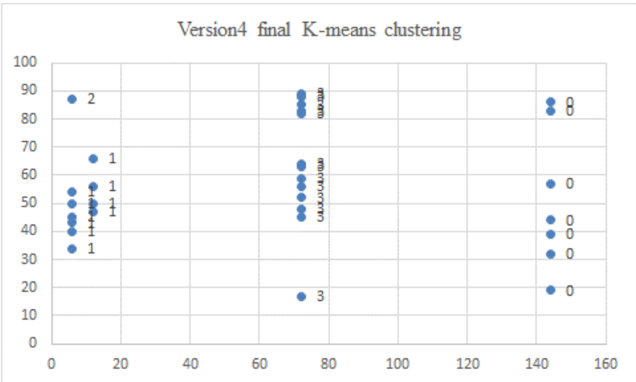

Figure7: K-means cluster of table-2 version 4

The training of 32 data sets of version 4 gives proper clusters; the testing is done on 32 data sets wherein the data are assigned proper clusters based on the trained data

Hierarchical clustering

The hierarchical clustering uses the approach of combining clusters that are having minimum distance. Thus leading to proper seperated clusters forming a maximum distance with every cluster that is combined.The centroids of version4 are shown in the figures 8,9,10,11.

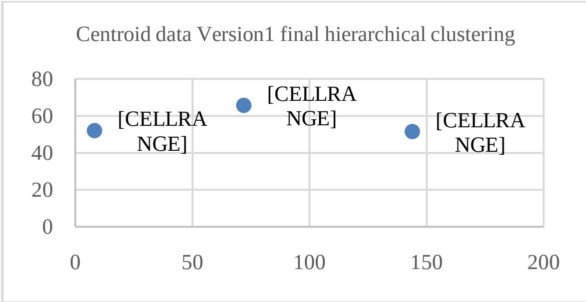

Figure 8: Centroid of Hierarchical cluster of table-2 version 1

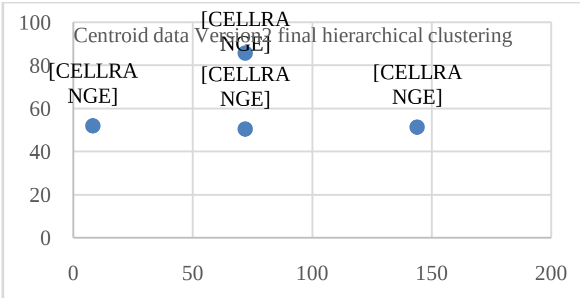

Figure 9: Centroid of Hierarchical cluster of table-2 version 2

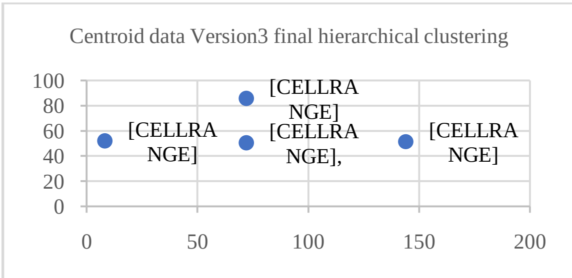

Figure 10: Centroid of Hierarchical cluster of table-2 version 3

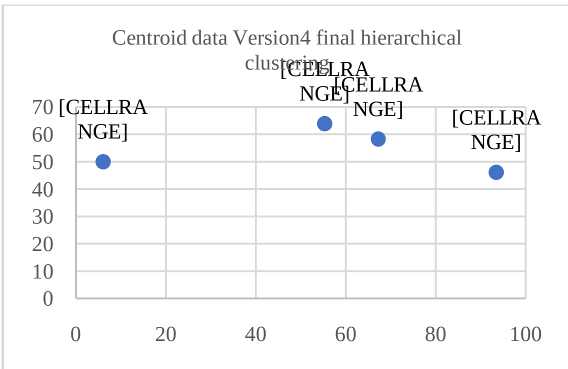

Figure 11: Centroid of Hierarchical cluster of table-2 version 4

January - February 2020
ISSN: 0193 - 4120 Page No. 01 - 11

There is a need to first know the number of clusters per case as shown in figure 12.

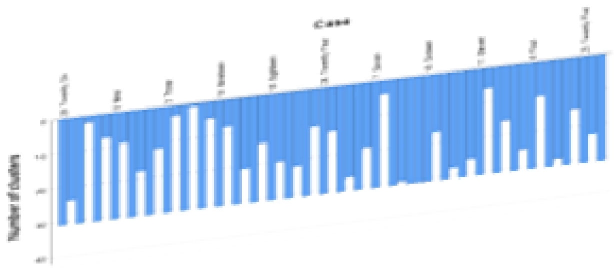

Figure 12: Number of clusters bs cases

Dendrogram until only one cluster remains after combining the clusters that is depicted in the figure13.

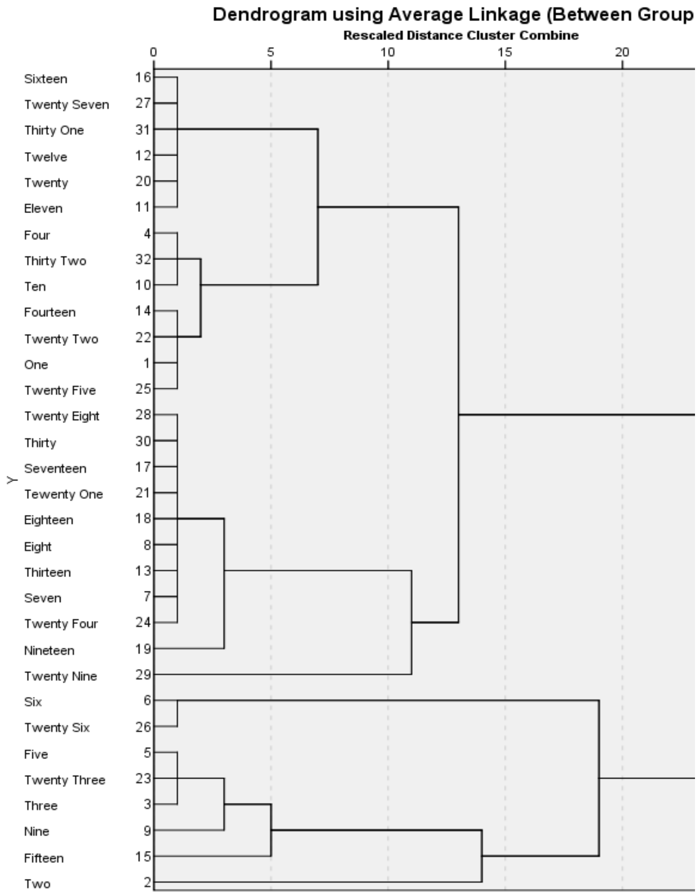

Figure 13: Dendrogram to show merging of clusters

Proposed Combined K hierarchy

The Proposed algorithm is already explained in the previous section 4. The minimum distance is calculated and the nearest distance is calculated for two clusters and combined to form a cluster. This method is repeated until only one cluster is left,the paper uses combining clusters until proper separation and not until the generation of only one cluster as in table 8. This combined cluster data is then fed into k cluster analysis so that it does not lead to any error in the outcome shown in table 5,6. Initially the combined cluster names are not arranged sequentially but on completion the clusters can be reassigned in sequential order shown in table 7,8.

Table 5:Combined K hierarchy initial formayion

Version 1 Initial combined K hierarchy	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	SSE
Centroid x	144	6.857143	30	60	104	18727.38
Centroid y	75.333333	53.571433	32.666667	58.5	63	

Table 6:Combined K hierarchy final formation

Version 1 Final combined K hierarchy	Cluster 0	Cluster 1	Cluster 2	Number of Iterations	SSE
Centroid x	144	8.181818	72	2	11950.56
Centroid y	51.42857	52	65.64286		

Table 7:Combined K hierarchyinitial clusters

Initial cases in	Version1

each cluster	
0	3
1	7
2	3
3	10
4	9

Table 8:Combined K hierarchyfinal clusters

Final cases in each cluster	Version1
0	7
1	11
2	14

VI. Inference of experimental results

K-means and Hierarchical are used in industry and research respectively with initial clusters allocated randomly. In this paper the outcome of each is shown separately, but the K-mean leads to errors as the centroids are not allocated properly. The Hierarchical shows us the formation of maximum distance between clusters.

Hence the combined k hierarchy uses the hierarchical strategy initially to separate out each cluster first and then feed the data into K to get a better outcome without errors

VII. Conclusion

K-means calculates the final proper clusters provided there is no error in the subsequent cluster formation, therefore use hierarchical with some modification to find proper initial clusters. This is done in the proposed combined K hierarchy; the combination of clusters is done until a proper maximum distance cluster separation is achieved using hierarchy this is then fed into the K to get the proper clusters without leading to mixing of data into any of the clusters. The flaws in each method is discussed. Hence the best outcome is achieved by combining K and hierarchy as explained in the proposed algorithm. This is helpful for properly segregation the customer based on historic data and providing them with discounts and benefits for the online products in the future. Thereby generating a win-win situation for the customer and online site.

VIII. References

[1] Tian-Shi Xu., et.al,"Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data ",IEEE Transactions on Power Delivery.,vol. 32,no. 2 ,pp. 609 - 616, April 2017

[2] Tung-Shou Chen., et.al,"A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray", ISPACS 2005. Proceedings of 2005 International Symposium, pp.405-408

[3] Mohammad Shabbir Hasan and Zhong-HuiDuan, (2015) chap4,Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Elsevier In http://www.shabbirhasan.com/files/papers/book_chapter.pdf

[4] Okayama, H.,et al., " Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas". 2012, Canc. Res. 72, pp.100–111

[5] Ihle, N.T., et al.,"Effect of KRAS oncogene substitutions on protein behavior: implications forsignaling and clinical outcome",2012,. J. Natl. Canc. Inst. 104, pp.228–239

[6] Hasan, M.S., Duan, Z.-H.," A hybrid clustering algorithms and functional study of gene expression in lung adenocarcinoma". In: Proceedings of the World Comp: International Conference on Bioinformatics and Computational Biology,2014, pp. 23–29.

Priyanka Desai

Dr.Priyanka Desai Trainer,Analyst,ex-faculty, Mumbai. She has more than 13 years of teaching/ industry experience in various subjects such as Python, Information Retrieval, Software Engineering, Web Technology, Object Oriented programming, Databases, Computer Networks, Web Mining, worked as trainee at i-flex solutions Mumbai. Has worked as M.E. coordinator, member of organizing committee National conference/International Conference, track manager for International conference (ACM, IJCA) at TCET Mumbai and paper setter at University of Mumbai. Is also an ISO certified internal auditor, life

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 01 - 11

member of ISTE and certification of IBM rational rose. Three students have completed their M.E under her guidance at TCET. Published/presented more than 12 papers in International/National Journals and Conferences. Has handled many academic projects at TCET, was a reviewer of International Conference, "ICSCET 18"(IEEE-UoCE), Mumbai. Present inclination towards ML, IoT(Arduino), Python, Salesforce-Admin, Technical content development, Patent law for engineers and scientist and Basic Patent writing.