# K-Means Clustering of Cloud Data using Weka and R Language

Banshidhar Choudhary
Department of Computer Science
Babasaheb Bhimrao Ambedkar University
Lucknow, 226025, India

Vipin Saxena
Department of Computer Science
Babasaheb Bhimrao Ambedkar University
Lucknow, 226025, India

## ABSTRACT
From the literature, it is observed that the there is tremendous growth of cloud data which is increasing day by day in an exponential manner. The cloud data contains large files in the form of text, audio and video formats. Therefore, for optimizing the search timings for said files, there is a need of clustering of data. In the present work, K-means clustering is applied for the large data of banking sector and for this purpose, Weka and R language are used which give optimize results to search the desired information. Computed results are depicted through figures and tables.

## Keywords
Cloud data, File Formats, Clustering, K-Means, Search, Credit/Debit Cards.

## 1. INTRODUCTION
Data mining is a grey area of the research using the important techniques to find out the pattern from the raw data set stored over the cloud servers in which servers are well connected using the concept of distributed computing network. There are two types of learning techniques for analysis of data available in the form of the hidden pattern, which is important for data analysis one is called as supervised while another is known as unsupervised learning. Further, a cluster is defined as a collection of objects which are similar anddissimilar to the objects belonging to other clusters. It is a common technique which is used for statistical data analysis in many fields like bioinformatics, machine learning, image analysis, pattern recognition and many more. In the current scenario, data is stored by many organizations over the cloud servers which are well connected through network topologies and one of the organizations is the bank which stores huge amount of data over thecloudservers and the data may be homogeneous or heterogeneous in nature.Banks access the data from clouds and provide the services to the customersaround the 24*7*365 and from that service customersget benefit for accessing the accounts and other information. Due to exponential growth of internet services and innovation of the 5G technology, the telecommunication organizations are providing the 5G services to the users which is much faster than the 3G and 4G services and the database can be accessed through the hand-held devices like mobile phones, smart watches, laptops, desktop and PDA, etc., which are well connected through the distributed network system. On these devices, the data can be viewed within fraction of seconds. For this purpose, various methods are used for searching of the desired information from the big data set and one of the important methods is clustering of the data and further the desired information can be found using the K-means technique which is used in the present work on the data set of the banking sector. This method is a widely used partition-based clustering methods, it can be easily implemented and most efficient one in terms of the computation of the execution time for K-mean clustering group items into k groups. This group is selected because of minimizing the sum of squared of distance between objects and the corresponding centroid. Let us describe some of the important references related to the present work in the next section.

## 2. RELATED WORK
Many researchers have studied the clustering algorithms as advances in the data management techniques. In this regard, Ahmad and Dey [1] have studied clustering algorithm based on K-mean for the data with mixed numerical and categorical feature on real world data set. Rujasiri and Chomtee [2] focused on the effectiveness of the five-cluster methods through comparison between them with multivariate data and finally founded that methods considered were most suitable for cluster analysis. Bharti, et al. [3] have proposed a hybrid model of data mining technique for intrusion detection and overcome the problem of class dominance, force assignment and selection of class problem. The performance of the proposed model is evaluated over KDD Cup 1999 data set. Srivastava et al. [4] have compared two clustering techniques K-mean and C-mean clustering used for content-based image retrieval system. Ahmad [5] also presented the performance of k-mean clustering algorithm for various values of k, and k is the number of clusters using this technique for finding a network intrusion detection. Aggarwal and Aggarwal [6] have presented a midpoint-based K-mean clustering algorithm that gives a solution to the limitation of the original K-means algorithm. Chaturvedi and Rajavat [7] proposed a new K-mean clustering algorithm in which authors computed the initial centroids systematically instead of randomly assigned due to which accuracy and time improved. Sharma [8] has analyzed the four clustering algorithms like K-mean algorithm, hierarchical algorithm, density-based algorithm and EM algorithm with respond to time to build a model, instance of clusters square error and log likelihood by using wekatool.Yuan et al. [9] have defined the concept of transaction database clustering which is based on K-mean. Institute of Electrical and Electronics Engineers [10] focused on performance of different data set using fuzzy K-mean clustering in map reduced on hadoop, which shows the execution time. Asnani [11] served about the cloud computing storage, with the analysis of clustering analysis for various business-integrated application. Shaaban et al. [12] presented evaluation of K-mean and fuzzy C-mean image, segmentation using cluster classifier

and discovered using the results which confirm the effectiveness of the proposed fuzzy C-mean, image segmentation-based cluster classifier. Rathore [13] proposed a new modified clustering algorithm and implemented by improving the data quality by removing the outlier point, and finally compared with the traditional k-mean clustering algorithm. Abualigah et al. [14] proposed multi-objective based text clustering technique using K-mean which showed the better performance in term of accuracy and F-measure comparatively with other techniques. Verma et al.[15] improved the clustering technique by defining the clusters automatically, and to assign required cluster to un-clustered points and used backtracking method to find exact number of clusters. Chahal and Kaur et al. [16] proposed a hybrid system, which is divided into two parts first, cluster the data points using K-mean algorithm and second using Adaptive–SVM algorithm and classified the training data. Patel et al. [17] compared own cluster algorithm from algorithms like K-mean algorithm, K-medoids, distributed K-mean, hierarchical cluster and density-based clustering and describes which clustering algorithm should use in different conditions for query for production of best results. Kumar et al. [18] also proposed K-mean clustering algorithm for automatic detection of the acute Leukemia. Bansal et al. [19] proposed improved K-mean clustering algorithm which is to be used for defining the number of clusters automatically and assigned the require cluster to un-cluster points and improved in accuracy and reducing the time by the member assigned to the cluster to predict cancer. Khan et al. [20] studied perception of student using of mobile data by K-mean clustering algorithm and explored the gap using K-mean clustering algorithm. Kalra et al. [21] proposed a framework for datamining of heterogeneous data for a multiple heterogeneous data using k-mean clustering on real life dataset and analysed the result in the form of clusters. Kuraria et al. [22] proposed a centroid selection process using WCSS and Elbow method for K-mean clustering algorithm in data mining. Rashid et al. [23] outlined a method for storage data for health care system and preserved privacy using constant key length encryption technique to secure data on cloud storage irrespective of the type of users. Vats and Sagar [24] evaluated the performance of the K-mean algorithm in different ways like k-mean (using machine learning library), K-mean simple (using java code on map reduce) with Initial Equidistance Centers (IEC) and K-mean on spark (Machine Learning Library ) on static IP address data sets over the map reduce framework. Wye et al. [25] discussed a novel approach to localize human location based on the current zone via K-mean clustering. The K-mean clustering was proposed to validate antenna placement to divide the testbed into zones using Received Signal Strength Indicator (RSSI). ICICCS 2020 [26] reviewed various outlier and clustering technique and after that, problems are identified using the K-mean clustering algorithm for outlier detection. Shang et al. [27] proposed analysis of simple K-mean and parallel K-mean clustering for software product and organizational performance using k-mean education sector dataset. Zada et al. [28] proposed performance evaluation of simple K-mean and parallel K-mean clustering algorithm for big data business process and showed that parallel algorithm reduces the number of execution and the time it takes to complete a task. Omer et al. [29] presented a widely used Markovchain based on probabilistic modelling approach and k-mean clustering together form a hybrid method for prediction of mobile network accessibility and readability.

# 3. PROPOSED METHOD

For applying the K-mean clustering, a data set has been collected for the credit/debit cards provided by the different banks to customers. Further, the algorithm is implemented using Weka and R language. The steps involved in the K-mean algorithm are given below:

*Step 1: Select the Number of Clusters as K;*

*Step 2: Randomly Select the Centre Point as k's for each Cluster*

*Step 3: Assign all the Points to Closet Centre Centroid;*

*Step 4: Re-compute the Centroids from newly formed Clusters;*

*Step 5: Repeat Steps 3 and 4 until reaching Stable Cluster.*

The above steps are pictorial represented in the following figure 1.

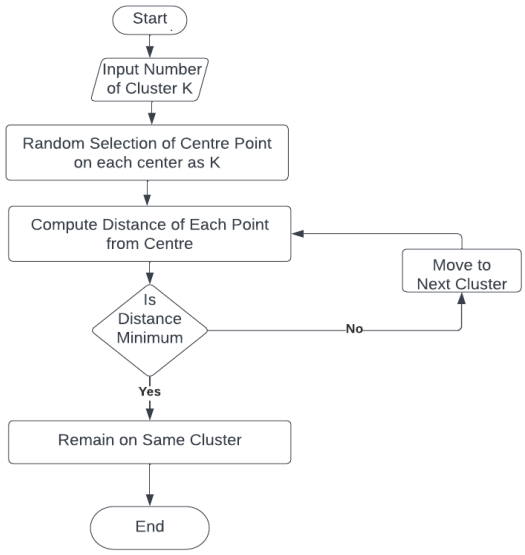


**Figure 1 Pictorial Representation of K-Mean Clustering**

The above algorithm is also known as the flat clustering algorithm, and it is suitable for the large data sets. In the algorithm, sum of the squared distance of the data sets would be minimum from the centroid and further it consists of less variation inside the clusters, therefore, one can get similar data sets inside the cluster.

# 4. RESULTS AND DISCUSSION

Let us implement the above algorithm for optimizing the search timings for the long data sets of banking sector, which has a large data set of debit/credit cards provided to the customer for availing the services given by bank. The details of card are represented below in the table 1 which contains attributes like ID, Customer, Age, Card1, Card2 and Card3. Multiple cards are considered as one customer may have more than one card. For clustering purpose, card number is converting into data point by summing up all the digits that is provided in the table 2.

Using table 2, it is given in bdcard.csv and is used by WEKA's implementation for clustering reasons. Weka is a tool for clustering; K-mean clustering is applied in this tool, whereas for calculating the distance between centroid to

data points. By executing it over Weka, it is observed that the time required for developing a model with Euclidian distance is computed as 0.02 seconds, and the value of k is 3. Further, it is noticed that twenty-five instances and five attributes are considered for the first clustering, represented by cluster0 which has 10 elements, another cluster is represented by cluster1 which has 8 elements, and finally the third cluster is represented by cluster2 which has 7 elements. As per

**Table 1 Details of Cards of the Customer issued by Bank**

| ID | Customer | Age | Card1 | Card2 | Card3 |
|---|---|---|---|---|---|
| 101 | S1 | 52 | 4200778927351900 | 4492907363241340 | 4423916322437560 |
| 102 | S2 | 32 | 4351245711358470 | 4835263951256210 | 4139636597873050 |
| 103 | S3 | 44 | 4200101521325820 | 4682817921471080 | 4384717548838970 |
| 104 | S4 | 38 | 4050453813353890 | 4503946857111580 | 4934193188723190 |
| 105 | S5 | 35 | 4361664384555660 | 4426437886798500 | 4331593599718930 |
| 106 | S6 | 28 | 4183026391245030 | 4009911666236670 | 4053467692969280 |
| 107 | S7 | 25 | 4874657217235480 | 4690062165422420 | 4129564158434530 |
| 108 | S8 | 43 | 4053057452828780 | 4086269975915990 | 4153727943447130 |
| 109 | S9 | 36 | 4153727943447130 | 4367063511982550 | 4071956557833210 |
| 110 | S10 | 27 | 4275688577132560 | 4800024388761150 | 4392132741528100 |
| 111 | S11 | 45 | 4809729378151410 | 4247913936211590 | 4490293574881720 |
| 112 | S12 | 42 | 4009162922114070 | 4750324891411770 | 4194291258997530 |
| 113 | S13 | 30 | 4445466134688480 | 4828609892739930 | 4154868788586040 |
| 114 | S14 | 47 | 4782495158244830 | 4853907545314310 | 4629317268614880 |
| 115 | S15 | 53 | 4926851896119880 | 4414018673565270 | 4338473626563550 |
| 116 | S16 | 55 | 4321402897333950 | 4366718211745060 | 4124422612745050 |
| 117 | S17 | 50 | 4819825422548530 | 4148244593838090 | 4079566211668550 |
| 118 | S18 | 48 | 4476131559731100 | 4165021232865680 | 4811964665957910 |
| 119 | S19 | 25 | 3733803757525570 | 4190874265258930 | 4510461921512960 |
| 120 | S20 | 33 | 4449521734654070 | 4148815184118600 | 4148931417744020 |
| 121 | S21 | 39 | 4843288947855130 | 3728929234194840 | 4472372596232140 |
| 122 | S22 | 44 | 4906992218896500 | 4252335329317210 | 4471306376741890 |
| 123 | S23 | 34 | 4758977475625990 | 4070622149611660 | 4085046159573380 |
| 124 | S24 | 29 | 4912716221948540 | 4121866593879090 | 4471306376741890 |
| 125 | S25 | 40 | 4605826661673470 | 4314679136824100 | 4038581676739020 |

**Table 2 Conversion of Details of Cards into Data Set**

| ID | Age | Card1 | Card2 | Card3 |
|---|---|---|---|---|
| 101 | 52 | 73 | 64 | 66 |
| 102 | 32 | 61 | 70 | 89 |
| 103 | 44 | 42 | 74 | 91 |
| 104 | 38 | 67 | 73 | 80 |
| 105 | 35 | 80 | 86 | 84 |
| 106 | 28 | 59 | 71 | 89 |
| 107 | 25 | 65 | 59 | 67 |
| 108 | 43 | 84 | 69 | 91 |
| 109 | 36 | 65 | 68 | 72 |
| 110 | 27 | 84 | 58 | 54 |
| 111 | 45 | 77 | 73 | 81 |
| 112 | 42 | 48 | 70 | 80 |
| 113 | 30 | 82 | 96 | 86 |
| 114 | 47 | 82 | 70 | 76 |
| 115 | 53 | 94 | 68 | 79 |
| 116 | 55 | 66 | 70 | 57 |
| 117 | 50 | 70 | 72 | 72 |
| 118 | 48 | 61 | 61 | 85 |
| 119 | 25 | 76 | 82 | 75 |
| 120 | 33 | 73 | 68 | 66 |
| 121 | 39 | 79 | 83 | 65 |
| 122 | 44 | 87 | 57 | 100 |
| 123 | 34 | 101 | 60 | 76 |
| 124 | 29 | 65 | 79 | 71 |
| 125 | 40 | 75 | 62 | 78 |

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: bdcard, Instances: 25 Attributes: 5
ID, Age, Card1, Card2, Card3
Test mode: evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 7
Within cluster sum of squared errors: 5.023034233086975
Initial starting points (random):
Cluster 0: 114, 47,82,70,76
Cluster 1: 113, 30,82,96,86
Cluster 2: 121, 39,79,83,65
Missing values globally replaced with mean/mode
Final cluster centroids:

| Attribute | Full Data (25.0) | Cluster# 0 (10.0) | 1 (8.0) | 2 (7.0) |
|---|---|---|---|---|
| ID | 113 | 114.5 | 106.75 | 118 |
| Age | 39.68 | 48.8 | 35.625 | 31.2857 |
| Card1 | 72.64 | 79.5 | 63 | 73.8571 |
| Card2 | 70.52 | 66.4 | 76 | 70.1429 |
| Card3 | 77.2 | 78.3 | 83.875 | 68 |

Time taken to build model (full training data): 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 10 (40%)
1 8 (32%)
2 7 (28%)

**Figure 2 Cluster Formation by Weka's Tool**

the Weka's implementation, the above figure 2 gives information about the computation time of 0.02 seconds cluster formation of similar kinds of data points which may lead to desired information in the optimized time. Further, the above data represented in the table 2 is clustered using R language. The details str(df), head(df) and df which are shown below in the figure 3. It is observed that R language provides a variety of libraries related to the Statistical methods alongwith excellent Weka's implementation as shown in the figure 2 for computing environment especially related to the Statistical methods. It is used a programming tool for importing and cleaning of the data so that one can search the desired information within fractions of seconds. With the help of the R programming language, one can cluster data points using the K-mean with Euclidian distance to produce three clusters: Cluster1 has 11 elements, Cluster2 has 8. elements, and Cluster3 has 6 elements. In addition, one can obtain cluster mean, clustering vector, and cluster identification for each observation, as represented in the figure 4. After that, one can get the confusion matrix of Age with its cluster and for plotting a cluster, Age and Card1 attributes are considered, and confusion matrix are shown below in the figure 5 and re-centering of the data points is represented in the figure 6.

```
> df=read.csv("bdcard.csv")
> str(df)
'data.frame':   25 obs. of  5 variables:
 $ ID   : int  101 102 103 104 105 106 107 108 109 110 ...
 $ Age  : int  52 32 44 38 35 28 25 43 36 27 ...
 $ Card1: int  73 61 42 67 80 59 65 84 65 84 ...
 $ Card2: int  64 70 74 73 86 71 59 69 68 58 ...
 $ Card3: int  66 89 91 80 84 89 67 91 72 54 ...
> head(df)
   ID Age Card1 Card2 Card3
1 101  52    73    64    66
2 102  32    61    70    89
3 103  44    42    74    91
4 104  38    67    73    80
5 105  35    80    86    84
6 106  28    59    71    89
> df
    ID Age Card1 Card2 Card3
1  101  52    73    64    66
2  102  32    61    70    89
3  103  44    42    74    91
4  104  38    67    73    80
5  105  35    80    86    84
6  106  28    59    71    89
7  107  25    65    59    67
8  108  43    84    69    91
9  109  36    65    68    72
10 110  27    84    58    54
11 111  45    77    73    81
12 112  42    48    70    80
13 113  30    82    96    86
14 114  47    82    70    76
15 115  53    94    68    79
16 116  55    66    70    57
17 117  50    70    72    72
18 118  48    61    61    85
19 119  26    76    82    75
20 120  33    73    68    66
21 121  39    79    83    65
22 122  41    87    57   100
23 123  54   101    60    76
24 124  29    65    79    71
25 125  40    75    62    78
```

**Figure 3 Creation of Data Frames by R Language**

```
> df1 <- df[, -6]
> head(df1)
  ID Age Card1 Card2 Card3
1 101  52    73    64    66
2 102  32    61    70    89
3 103  44    42    74    91
4 104  38    67    73    80
5 105  35    80    86    84
6 106  28    59    71    89
> # Fitting K-Means clustering Model
> # to training dataset
> set.seed(240) # Setting seed
> kmeans.re <- kmeans(df1, centers = 3, nstart = 20)
> kmeans.re
K-means clustering with 3 clusters of sizes 11, 8, 6

Cluster means:
       ID     Age   Card1   Card2   Card3
1 115.3636 37.45455 71.90909 69.54545 67.54545
2 113.8750 43.50000 85.87500 72.37500 84.12500
3 107.5000 38.66667 56.33333 69.83333 85.66667

Clustering vector:
 [1] 1 3 3 3 2 3 1 2 1 1 2 3 2 2 2 1 1 3 1 1 1 2 2 1 1

Within cluster sum of squares by cluster:
[1] 3455.636 2832.500 1142.333
 (between_SS / total_SS =  41.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
> # Cluster identification for
> # each observation
> kmeans.re$cluster
 [1] 1 3 3 3 2 3 1 2 1 1 2 3 2 2 2 1 1 3 1 1 1 2 2 1 1
>
```

**Figure 4 Identification of Clusters through R Language**

Finally, a cluster card is created by the R language which is represented in the figure 7. In the above, it is represented that data points having similarities are grouped together for finding the desired information by the user in the optimized time frame and these groups are three and above concept is very useful for the search of information available in the large files in the form of text, audio or video.

```
> # Confusion Matrix
> cm <- table(df$Age, kmeans.re$cluster)
> cm

    1 2 3
  25 1 0 0
  26 1 0 0
  27 1 0 0
  28 0 0 1
  29 1 0 0
  30 0 1 0
  32 0 0 1
  33 1 0 0
  35 0 1 0
  36 1 0 0
  38 0 0 1
  39 1 0 0
  40 1 0 0
  41 0 1 0
  42 0 0 1
  43 0 1 0
  44 0 0 1
  45 0 1 0
  47 0 1 0
  48 0 0 1
  50 1 0 0
  52 1 0 0
  53 0 1 0
  54 0 1 0
  55 1 0 0
>
> # Model Evaluation and visualization
> plot(df1[c("Age", "Card1")])
> plot(df1[c("Age", "Card1")],
+      col = kmeans.re$cluster)
> plot(df1[c("Age", "Card1")],
+      col = kmeans.re$cluster,
+      main = "K-means with 3 clusters")
>
> ## Plotiing cluster centers
> kmeans.re$centers
       ID      Age    Card1    Card2    Card3
1 115.3636 37.45455 71.90909 69.54545 67.54545
2 113.8750 43.50000 85.87500 72.37500 84.12500
3 107.5000 38.66667 56.33333 69.83333 85.66667
```

**Figure 5 Confusion Matrix by R Language**

```
> kmeans.re$centers[, c("Age", "Card1")]
       Age    Card1
1 37.45455 71.90909
2 43.50000 85.87500
3 38.66667 56.33333
>
> # cex is font size, pch is symbol
> points(kmeans.re$centers[, c("Age", "Card1")],
+        col = 1:3, pch = 8, cex = 3)
>
> ## Visualizing clusters
> y_kmeans <- kmeans.re$cluster
> clusplot(df1[, c("Age", "Card1")],
+          y_kmeans,
+          lines = 0,
+          shade = TRUE,
+          color = TRUE,
+          labels = 2,
+          plotchar = FALSE,
+          span = TRUE,
+          main = paste("Cluster Card"),
+          xlab = 'Age',
+          ylab = 'Card1')
> cm <- table(df$Card1, kmeans.re$cluster)
> cm

     1 2 3
  42  0 0 1
  48  0 0 1
  59  0 0 1
  61  0 0 2
  65  3 0 0
  66  1 0 0
  67  0 0 1
  70  1 0 0
  73  2 0 0
  75  1 0 0
  76  1 0 0
  77  0 1 0
  79  1 0 0
  80  0 1 0
  82  0 2 0
  84  1 1 0
  87  0 1 0
  94  0 1 0
```

**Figure 6 Re-centering of Data Points by R Language**

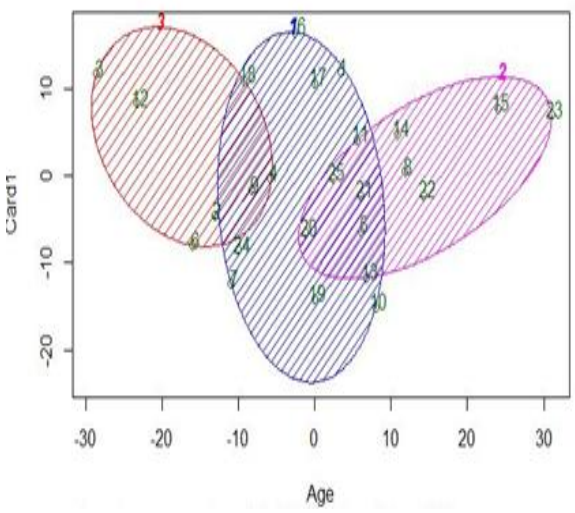


**Figure 7 Creation of Cluster Card for Data Points by R Language**

## 5. CONCLUSIONS

It is fact the K-mean clustering is very useful for handling the large data sets available over the clouds and day by day, it is having exponential growth. The large files may be in the form of text, audio and video formats. From the above work, it is concluded that the Weka and R language are very efficient tool and programming language, respectively for optimizing the search timings for these files which is the need of the customer in the present scenario. The concept of clustering has been implemented on the case study of cards issued by the banks to the customers and computed the time for formation of clusters as 0.02 seconds. The present work may be extended for the large finite data sets of any organization.

## 6. REFERENCES

[1] Ahmad, A., and Dey, L. (2007), "A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data". *Data and Knowledge Engineering*, *63*(2), 503–527. https://doi.org/10.1016/j.datak.2007.03.016.

[2] Rujasiri, P., and Chomtee, B. (2009), "Comparison of Clustering Techniques for Cluster Analysis". In *Nat. Sci.* 43, 378-388.

[3] Bharti, K. K., Shukla, S., and Jain, S. (2010), "Intrusion Detection using Clustering", *International Journal of Computer and Communication Technology*, 248–255. https://doi.org/10.47893/ijcct.2010.1052.

[4] Shrivastava, R., Upadhyay, K., Bhati, R., and Mishra, D. K. (2010), "Comparison between K-Mean and C-Mean Clustering for CBIR", *Proceedings of 2nd International Conference on Computational Intelligence, Modelling and Simulation, CIMSim 2010*, 117–118. https://doi.org/10.1109/CIMSiM.2010.66.

[5] Ahmad, A. (2010), "Data Clustering Using K-Mean Algorithm for Network Intrusion Detection*"*,M.Tech. Report in Lovely Professional University, Chandigarh, India.

[6] Aggarwal, N. and Aggarwal, K. (2012), "A Mid-Point based K-mean Clustering Algorithm for Data mining". *International Journal on Computer Science and Engineering*, 4(6), 1174-1180, Springer Singapore, ISSN: 0975-3397.

[7] Chathurvedi, N. and Rajavat, A. (2013), "An Improvement in K-mean Clustering Algorithm using Better Time and Accuracy", *International Journal of Programming Language and Application*, 3(4), 13-19. https://doi.org/10.5121/ijpla.2013.3407.

[8] Sharma, I. (2014), "Comparison of Different Clustering Algorithms using WEKA Tool", 1(2), 20-22. *ISSN*: 2349-7173. www.ijartes.org.

[9] Yuan, D., Cuan, Y. and Liu, Y. (2014), "An Effective Clustering Algorithm for Transaction Databases Based on K-Mean". *Journal of Computers*, *9*(4). https://doi.org/10.4304/jcp.9.4.812-816.

[10] Garg, D. and Trivedi, K (2014), "Fuzzy K-Mean Clustering in Map Reduce on Cloud Based Hadoop", *ICACCCT :Proceedings of 2014 IEEE International Conference on Advanced Communication Control & Computing Technologies : May 08-10, 2014*. ISBN: 978-1-4799-3914-5/14.

[11] Asnani, R. (2015), "A Distributed K-mean Clustering Algorithm for Cloud Data Mining". *International Journal of Engineering Trends and Technology*, *30*(7), ISSN: 2231-5381. http://www.ijettjournal.org.

[12] Shaaban, H. R., AbdulkaremHabib, A., and Abbas Obaid, F. (2015), "Performance Evaluation of K-Mean and Fuzzy C-Mean Image Segmentation Based Clustering Classifier", *International Journal of Advanced Computer Science and Applications*, 6(12), 176-183. www.ijacsa.thesai.org.

[13] Rathore, P. (2016), "Analysis and Performance Improvement of K-means clustering in Big Data Environment". *International Conference of Communication Network,* http://DOI10.1109/ICCN.2015.9.

[14] Abualigah, L. M., Khader, A. T., and Al-Betar, M. A. (2016), "Multi Objectives Based Text Clustering Technique using K-mean Algorithm", *Proceedings - CSIT 2016: 7th International Conference on Computer Science and Information Technology*. https://doi.org/10.1109/CSIT.2016.7549464.

[15] Verma, V., Bhardwaj, S., and Singh, H. (2016), "A Hybrid K-Mean Clustering Algorithm for Prediction Analysis". *Indian Journal of Science and Technology*, *9*(28). https://doi.org/10.17485/ijst/2016/v9i28/98392.

[16] K. Chahal, J., and Kaur, A. (2016). "A Hybrid Approach based on Classification and Clustering for Intrusion Detection System". *International Journal of Mathematical Sciences and Computing*, *2*(4), 34–40, https://doi.org/10.5815/ijmsc.2016.04.04.

[17] Patel, Archana K. M. and Thakral, Pratel (2016), "The Best Clustering Algorithm in Data Mining", *Adhiparasakthi Engineering College. Department of Electronics and Communication Engineering, Institute of Electrical and Electronics Engineers. Madras Section, & Institute of Electrical and Electronics Engineers (ICCSP), ISBN: 9781509003969.* https://doi.org/978-1-5090-0396-9/16/$31.00@2016IEEE.

[18] Kumar, S., Mishra, S., and Asthana, P. (2017), "Automated Detection of Acute Leukemia using K-mean Clustering Algorithm*". Advance in Computer and Computation Science, 655-670.*

[19] Bansal, A., Sharma, M., and Goel, S. (2017), "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining". *International Journal of Computer Applications*, *157*(6), 35–40, https://doi.org/10.5120/ijca2017912719.

[20] Khan, A., Baseer, S., and Javed, S. (2017), "Perception of Students on Usage of Mobile Data by K-mean Clustering Algorithm", *International Journal of Advanced and Applied Seinces*, *4*(2), 17–21. https://doi.org/10.21833/ijaas.2017.02.003.

[21] Kalra, M., Lal, N., and Qamar, S. (2018), "K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data". In *Lecture Notes in Networks and Systems* (Vol. 10, pp. 61–70), Springer. https://doi.org/10.1007/978-981-10-3920-1_7.

[22] Kuraria, A., Jharbade, N., and Soni, M. (2018), "Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining". *International Journal of Scientific Research in Science, Engineering and Technology*, 190–195, https://doi.org/10.32628/ijsrset21841122.

[23] Rashid, M., Singh, H., and Goyal, V. (2019), "Cloud storage privacy in health care systems based on IP and geo-location validation using k-mean clustering technique". *International Journal of E-Health and Medical Communications*, *10*(4), 54–65, https://doi.org/10.4018/IJEHMC.2019100105.

[24] Vats, S., and Sagar, B. B. (2019), "Performance Evaluation of K-means Clustering on Hadoop Infrastructure", *Journal of Discrete Mathematical Sciences and Cryptography*, *22*(8), 1349–1363, https://doi.org/10.1080/09720529.2019.1692444.

[25] Wye, K. F. P., Kanagaraj, E., Zakaria, S. M. M. S., Kamarudin, L. M., Zakaria, A., Kamarudin, K., and Ahmad, N. (2019), "RSSI-based Localization Zoning using K-Mean Clustering", *IOP Conference Series: Materials Science and Engineering*, *705*(1), https://doi.org/10.1088/1757-899X/705/1/012038.

[26] Vaigai College of Engineering, and Institute of Electrical and Electronics Engineers, *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020):13-15 May 2020*.

[27] Shang, R., Ara, B., Zada, I., Nazir, S., Ullah, Z., and Khan, S. U. (2021), "Analysis of Simple K- Mean and Parallel K- Mean Clustering for Software Products and Organizational Performance Using Education Sector Dataset". *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/9988318.

[28] Zada, I., Ali, S., Khan, I., Hadjouni, M., Elmannai, H., Zeeshan, M., Serat, A. M., and Jameel, A. (2022), "Performance Evaluation of Simple K -Mean and Parallel K -Mean Clustering Algorithms: Big Data Business Process Management Concept", *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/1277765

[29] Omer, A. S., Yemer, T. A., and Woldegebreal, D. H. (2022), "Hybrid K-Mean Clustering and Markov Chain for Mobile Network*", Accessibility and Retain ability Prediction.* 9. https://doi.org/10.3390/engproc 2022018009.

# 7. AUTHORS' PROFILES

**Banshidhar Choudhary** received Post Graduate Degree in Computer Applications (M.C.A.) from Dr. Indira Gandhi National Open University, New Delhi in 2002 and M.Phil. Degree from Madurai Kamraj University in 2007 and currently a research scholar in the Department of Computer Science, Babasaheb Bhimrao Ambedkar University. He has 15 years of teaching experience in Computer Science field in the various Indian Universities and 03 years in the Al-Jabal Al-Garbi University, Libya. Currently, he is solving the research problems related to security of cloud data and data mining in the Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India under fellowship program of University Grant Commission (UGC), New Delhi.

**Prof. Vipin Saxena** received his Ph.D. degree from Indian Institute of Technology, Roorkee, Uttarakhand, India. Presently, he is working as Professor in Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He has published more than 190 research articles in the International and National Journals and Conferences, authored 05 books in the field of Computer Science and Scientific Computing, attended 55 International and National Conferences and received three National Awards for meritorious research work in the field of Computer Science and other details are available on www.profvipinsaxena.com. His research interests are Scientific Computing, Computer Networking and Software Engineering.