MP-K-Means: Modified Partition Based Cluster Initialization Method for K-Means Algorithm

Manoj Kumar Gupta, Pravin Chandra

Abstract*: In k-means algorithm, initial cluster centroids are selected arbitrarily which leads to diverse formation of clusters in each run. Consequently, accuracy and performance of k-means is majorly depends on the selection of initial centroids. Thus, the initial cluster centroids shall be chosen carefully to obtain better accuracy and performance of k-means algorithm. In view of this, a new Modified Partition based Cluster Initialization method for k-means called as MP-k-means is proposed in this paper. MP-k-means is an amended version of P-k-means [1] in which the range of values of each dimension is divided into 'k' equi-sized partition based on arithmetic average. This division of range into 'k' equi-sized partition is affected by outliers present in the data. In order to remove the effect of outliers in P-k-means, the partitioning of each dimension is made based on positional average instead of arithmetic average in MP-k-means. Six popular datasets are used for empirical evaluation of the algorithms. The empirical results are compared and validated based on various external and internal clustering validation measures. The comparative results show that MP-k-means is significantly superior to the basic k-means and P-k-means. The proposed method may also be applied to other clustering algorithms which are based on the concept of selection of initial cluster centroids.

Keywords*: K-means Algorithm; Cluster Initialization; Partition based Cluster Initialization; P-k-Means; MP-k-means; Data Mining; Clustering.

I. INTRODUCTION

Both machine learning and data mining are majorly used in several applications. A number of vital functions are used in machine learning and data mining [2, 3]. Clustering is one of them based on unsupervised learning to cluster the data observations based on distance / (dis)similarity among their various characteristics / attributes. Nearby / like observations are grouped together in the same cluster whereas the far off / alike observations are grouped in another cluster(s) [2, 4-6].

Numerous clustering methods are proposed in the literature [7, 8]. Due to simplicity, k-means is broadly used clustering algorithm to identify convex-shaped clusters. Basic k-means algorithm is presented as *Algorithm 1* [1, 2, 7]:

The arbitrary selection of initial cluster centroids, as depicted in Step 2 of *Algorithm 1*, leads to a formation of diverse set of clusters in each run of the basic k-means.

P-k-means algorithm [1], which is modified and presented in this paper, is described as *Algorithm 2*:

Algorithm 1: Basic k-means Algorithm

Step 1: *Decide k (# of clusters)*
Step 2: *Randomly initialize cluster centroids* $C = \{c_1, c_2, ..., c_k\}$
Step 3: *Repeat*
- *a. For each data point* (x_i) *in data set (D)*
  - *i. Compute distance* $dis(x_i, C)$ *between* x_i *and all cluster centroids*
  - *ii. Assign* x_i *to the nearest cluster*
- *b. Re-compute cluster centroids as the mean of all cluster members.*

Step 4: *Until cluster membership stabilizes.*

Algorithm 2: P-k-means: k-means using Partition Based Cluster Initialization Method

Step 1: *Decide k (# of clusters)*
// initialize k cluster centroids as per Steps 2.1 through Step 2.2
Step 2: *Initialize cluster centroids* $C = \{c_1, c_2, ... , c_k\}$ *as:*
// range of values of each dimension (i.e. attribute) is logically divided into 'k' equi-sized partitions based on arithmetic average of the respective attributes
Step 2.1: *Divide the range of data of each dimension,* dim_i *into k equi-ranged partitions.*
// logically model the partitions of each dimension as separate lists of partitions
// randomly select 'k' unique sets (containing one partition from each dimension) and then choose a random value from each chosen partitions as' k' initial centroids
Step 2.2: *Repeat*
- *i. Arbitrarily choose one partition from each dimension* (dim_i)*, which was not selected earlier.*
- *ii. Find the randomized value of each partition selected for centroid.*

Step 2.3: *If all centroids are chosen then go to Step 3 else go to Step 2.2*
// find out the cluster membership of each data point iteratively until cluster membership stabilizes
Step 3: *Repeat*
- *a. For each data point* (x_i) *in data set (D)*
  - *i. Compute distance* $dis(x_i, C)$ *between* x_i *and all cluster centroids*
  - *ii. Assign* x_i *to the nearest cluster*

Revised Manuscript Received on November 15, 2019.
***** Correspondence Author
Manoj Kumar Gupta, Research Scholar, USIC&T, Guru Gobind Singh Indraprastha University, Delhi. E-mail ID: manojkgupta5@gmail.com.
Pravin Chandra, Professor, USIC&T, Guru Gobind Singh Indraprastha University, Delhi, India. E-mail ID: chandra.pravin@gmal.com.

b. Re-compute cluster centroids as the mean of all cluster members.

Step 4: *If cluster membership stabilizes then end else go to Step 3.*

In P-k-means (*Algorithm 2*) [1], the choice of initial cluster centroids is affected by outliers present in one or more characteristics or attributes of the data set, as depicted in *Step 2(a)*, because arithmetic average is used to partition each dimension (i.e. attribute) of the data set. So as to avoid this situation, the *Step 2(a)* of *Algorithm 2* i.e. P-k-means Algorithm is modified and presented in this paper.

Related work in the field of cluster centroid initialization available in the literature is described in Section 2 of this paper. The modified P-k-means called as MP-k-means is proposed and described in Section 3. Experiment design and results of proposed method i.e. MP-k-means along with basic k-means and P-k-means on six popular data sets are described in Section 4. The results are compared using various internal and external clustering validation measures such as Accuracy, Performance, Intra-cluster Compactness, Inter-cluster Separation, Purity/Precision, Recall and F-Measure. Finally in Section 5, the conclusion is drawn. The comparative results exhibit that MP-k-means is significantly better than that of basic k-means and P-k-means in terms of the aforesaid clustering validation measures.

II. RELATED WORK

Accuracy and performance of k-means can be enhanced by careful selection of initial cluster centroids which are nearer to the actual cluster centroids. In view of this, a numerous attempts have been made by the researchers so far and proposed a variety of methods to initialize the cluster centroids [7, 8, 46]. Related work of some of the leading researchers available in the literature in the area of selection of initial cluster centroids is presented in *Table I*.

Table-I: Related Work

Reference	Method Proposed
Forgy [9]	Earliest cluster initialization method based on random basis
McQueen [10]	Similar to Forgy (1965) but differs in assigning the left over objects to one of the closest seed location
Kaufman and Rousseeuw [11]	Method in which most centrally located instance is chosen as first centroid
Katsavounidis *et al.* [12]	Cluster initialization begins by choosing an edge point X and then finds the furthest point.
Bradley and Fayyad [13]	Data is randomly broken into the J random small sub-subsets and then the initial points are selected as cluster centroid
Pei *et al.* [14]	Proposed two new clustering initialization techniques based on potential method. The proposed method is appropriate for data set in which distribution of data is agglomerative in feature space
Khan and Ahmad [15]	Proposed Cluster Center Initialization Algorithm (CCIA) based on the use of Density-based Multi Scale Data Condensation (DBMSDC) using the estimation of density of the data at a point
Su and Dy [16]	PCA-Part (Principal Components Analysis Partitioning) for cluster initialization based on the use of deterministic divisive hierarchical method
Hathaway *et al.* [17]	maximin initialization based on progressive sampling scheme for cluster initialization suitable for data which contain compact, separated clusters
Arai and Barakbah [18]	Cluster initialization method based on Hierarchical algorithm in order to determine the initial centroids
Arthur and Vassilvitskii [19]	Proposed k-means++ in which initial centroids are chosen consecutively with probability proportional to the distance to the nearest centroid.
Wu *et al.* [20]	New initialization method based on density of data points suitable for categorical data sets
Kang and Cho [21]	Cluster initialization method based on centrality, sparsity and isotropy
Maitra [22]	Selection of initial centroids by finding a representative local modes from the most separated ones
Xu *et al.* [23]	Initialization method based on reverse nearest neighbor suitable for continuous data
Dang *et al.* [24]	Initialization Method for Semi-supervised Clustering suitable for data which contain a number of potentially helpful information
Naldi *et al.* [25]	Cluster initialization methods based on evolutionary techniques
Reddy *et al.* [26]	Cluster initialization method based on Minimum Spanning Tree (MST) suitable for computational biology, pattern recognition and image processing
Bai *et al.* [27]	Cluster initialization method based on k-modes for categorical data
Chen [28]	Cluster initialization based of Hierarchical two-division method
Aldahdooh and Ashour [29]	Initialization of centroids based on selection method instead of the random selection

Reference	Method Proposed
Goyal and Kumar [30]	Mean-Based algorithm suitable for data sets in which attributes of data points having positive values
Duwairi and Abu-Rahmeh [31]	Spherical k-means based on distributed seeds across the input space
Poomagal *et al.* [32]	k-means initialization method based on iterative selection
Dhanabal and Chandramathi [33]	Cluster initialization method based on extreme end distance
Golasowski *et al.* [34]	Cluster initialization method based on Brute-force approach using heuristics
Kumar and Reddy [35]	Density based initialization method which is also scalable to large datasets
Ismkhan [36]	Proposed iterative k-means to reduce SSE
Nguyen *et al.* [37]	Propose k-means** to achieve global optimum solution
Sandhya and Sekar [38]	Three variant approaches for centroid initialization suitable for document clustering
Yu *et al.* [39]	Proposed bi-layer k-means and tri-level k-means algorithms
Kurada and Kanadam [40]	Automatic Clustering Using TLBO
Gupta and Chandra [1]	Proposed Partition based Cluster Initialization Method for k-means called as P-k-means

III. THE PROPOSED METHOD

To avoid the effect of outliers present in the features of the dataset and to get better accuracy and performance of P-k-means, an amended version of P-k-means to initialize the cluster centroids called as MP-k-means is devised and proposed in this paper. In the proposed method, the range of each dimension (or attribute), dim_i, of the dataset is logically divided in '*k*' equi-sized partitions based on positional average instead of arithmetic average, where '*k*' refers to the # of clusters. The *Step 2(a)* of *Algorithm 2* i.e. P-k-means is modified as stated above in the proposed algorithm MP-k-means (*Algorithm 3*). Rest of the steps of P-k-means (*Algorithm 2*) remains same. The proposed method i.e. MP-k-means, is presented as *Algorithm 3*:

IV. EXPERIMENT DESIGN AND RESULTS

Basic k-means, P-k-means and MP-k-Means algorithms are implemented in MATLAB and executed on six popular datasets taken from Hartigan and UCI. The results are computed and compared based on the average of 200 runs of each of the algorithms on each of the data sets used. The implementation is the standard one with no special optimizations.

A. Datasets Used

For empirical evaluation, all three algorithms are evaluated on six different datasets: *Animal Milk, Image Segmentation, IRIS, Pen Digit, Spambase* and *Wine*. First data set *Animal Milk* is taken from Hartigan *(https://people.sc.fsu.edu/ ~jburkardt/datasets/hartigan/file02.txt)*. Rest five datasets are taken from UCI Machine Learning Repository *(https://archive.ics.uci.edu/ml/datasets.html)*. Details of these datasets are presented in *Table II*.

Algorithm 3: MP-k-means: Modified k-means using Partition Based Cluster Initialization Method

Step 1: *Decide k (# of clusters)*

// initialize k cluster centroids as per Steps 2.1 through Step 2.2

Step 2: *Initialize cluster centroids* $C = \{c_1, c_2, \ldots, c_k\}$ *as:*

// range of values of each dimension (i.e. attribute) is logically divided in 'k' equi-sized partitions based on positional average of the respective attributes

Step 2.1: *Divide the data values of each dimension,* dim_i *into 'k' equi-sized partitions using positional averages.*

// logically model the partitions of each dimension as separate lists of partitions

// randomly select 'k' unique sets (containing one partition from each dimension) and then choose a random value from each chosen partitions as 'k' initial centroids

Step 2.2: *Repeat*

i. *Arbitrarily choose one partition from each dimension* (dim_i)*, which was not selected earlier.*
ii. *Find the randomized value of each partition selected for centroid.*

Step 2.3: *If all centroids are chosen then go to Step 3 else go to Step 2.2*

// find out the cluster membership of each data point iteratively until cluster membership stabilizes

Step 3: *Repeat*

a. *For each data point* (x_i) *in data set (D)*
   i. *Compute distance* $dis(x_i, C)$ *between* x_i *and all cluster centroids*
   ii. *Assign* x_i *to the nearest cluster*
b. *Re-compute cluster centroids as the mean of all cluster members.*

Step 4: *If cluster membership stabilizes then end else go to Step 3.*

Table-II: Datasets Used

Dataset	# of Clusters	# of Attributes	# of Instances
Animal Milk	5	4	16
Image Segmentation	7	19	2100
IRIS	3	4	150

Pen Digit	10	16	7494
Spambase	2	57	4601
Wine	3	13	178

B. Clustering Evaluation and Validation Measures

Clustering evaluation and validation measures are used to assess the validity of goodness of the clustering [41]. These are also used for the comparison of experiments and results of the clustering algorithms. These measures are broadly classified into two categories external measures and internal measures [42]. Various notations used in clustering evaluation measures are described in *Table III*.

Table-III: Notations used in Clustering Evaluation Measures

Notation	Meaning
k	# of Clusters
C_i	i^{th} Cluster
T_j	j^{th} Partition or Ground Truth
n	# of data objects
n_i	# of data objects assigned to C_i
m_j	# of data objects belongs to T_j
n_{ij}	# of data objects of C_i belong to T_j
T_{j_i}	partition which contains the maximum # of data objects from C_i
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
TP_i	True Positives of C_i
TN_i	True Negatives of C_i
FP_i	False Positives of C_i
FN_i	False Negatives of C_i
S_k	set of data objects in C_i
r	# of attributes or characteristics of the data set
x_{lj}	j^{th} attribute of the l^{th} data object belong to C_i
x_{ij}	j^{th} attribute of the i^{th} data object belong to C_i
$\overline{x}_{ij}$	j^{th} attribute of the cluster centroid of C_i
$\overline{x}_i$	the centroid of C_i

(i) *External Measures*

External measures are based on supervised learning in which clustering results are evaluated against the ground truth without employing criteria intrinsic to the dataset [41, 43]. External measures used in this paper are described as follows:

- **Purity** – Purity quantifies the degree that cluster C_i contains data objects only from one partition or ground truth. It is suitable for balanced data. Purity of cluster C_i and total purity of clustering C are defined by the *eq. 1* and *eq. 2* respectively [43].

$$purity(C_i) = \frac{1}{n_i} \max_{i=1}^{k} \{n_{ij}\} \quad (1)$$

$$purity(C) = \sum_{i=1}^{k} \frac{n_i}{n} purity(C_i) = \frac{1}{n} \sum_{i=1}^{k} \max_{j=1}^{k} \{n_{ij}\} \quad (2)$$

Clustering is called as perfect clustering if total purity of clustering *purity(C)* = 1. Total purity of clustering C is also defined by the *eq. 3*.

$$purity(C) = \frac{TP}{TP + FP} \quad (3)$$

- **Precision** – Precision is the fraction of data objects in cluster C_i from the majority partition or ground truth T_{j_i} (i.e., the same as purity). Like purity, it is also suitable for balanced data. Precision of cluster C_i and total precision of clustering C are expressed by the *eq. 4* and *eq. 5* respectively [43].

$$precision(C_i) = \frac{1}{n_i} \max_{i=1}^{k} \{n_{ij}\} = \frac{n_{ij}}{n_i} \quad (4)$$

$$precision(C) = \frac{1}{k} \sum_{i=1}^{k} precision(C_i) \quad (5)$$

Precision of the cluster C_i and total precision of the clustering C are also defined by the *eq. 6* and *eq. 7* respectively [44].

$$precision(C_i) = \frac{TP_i}{TP_i + FP_i} \quad (6)$$

$$precision(C) = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_i}{TP_i + FP_i} \quad (7)$$

- **Recall** – Recall is the fraction of data object in partition or ground truth T_{j_i} shared in common with cluster C_i, where $m_{j_i} = |T_{j_i}|$. Recall of cluster C_i is expressed by the *eq. 8* [43].

$$recall(C_i) = \frac{n_{ij}}{|T_{j_i}|} = \frac{n_{ij}}{m_j} \quad (8)$$

Recall of the i^{th} cluster C_i and total recall of the clustering C are also defined by the *eq. 9* and *eq. 10* respectively [44].

$$recall(C_i) = \frac{TP}{TP + FN} \quad (9)$$

$$recall(C) = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i} \quad (10)$$

- **F-measure** – Both purity and precision are not suitable for balanced data. In case of imbalanced data, false negatives play a major role in the cluster evaluation. In view of this, both precision and recall are used for cluster evaluation. F-measure is the harmonic mean of precision and recall. The F-measure of cluster C_i and F-measure of clustering C are defined by the *eq. 11* and *eq. 12* respectively [43, 44].

$$F(C_i) = \frac{precision(Ci)recall(Ci)}{precision(Ci)+recall(Ci)} = \frac{2n_{ij}}{n_i + m_j} \quad (11)$$

$$F(C) = \frac{1}{k}\sum_{i=1}^{k} F(C_i) \quad (12)$$

Consider the two hypothetical datasets presented in *Table IV* and *Table VI*. The values of Purity, Precision, Recall and F-measure computed using the datasets given in *Table IV* and *Table VI* are presented in *Table V* and *Table VII* respectively. In *Tables IV* to *Table VII*, T_1, T_2 and T_3 are the partitions based on ground truth; C_1, C_2 and C_3 are the clusters identified by the algorithm.

Table-IV: Hypothetical Dataset-1

Cluster (C)	Ground Truth (T)			Sum (n_i)
	T_1	T_2	T_3	
C_1	0	20	25	**45**
C_2	0	20	5	**25**
C_3	30	0	0	**30**
Sum (m_j)	**30**	**40**	**30**	**100**

Table-V:: Purity, Precision, Recall and F-measure based on Hypothetical Dataset-1

Metric	Cluster			Total
	C_1	C_2	C_3	
Purity	25/45 = 0.56	20/25 = 0.80	30/30 = 1.00	(0.56+0.80+1.00)/3 = 0.79
Precision	25/45 = 0.56	20/25 = 0.80	30/30 = 1.00	(0.56+0.80+1.00)/3 = 0.79
Recall	25/30 = 0.83	20/40 = 0.50	30/30 = 1.00	(0.83+.0.50+1.00)/3 = 0.78
F-measure	50/75 = 0.67	40/65 = 0.62	60/60 = 1.00	(0.67+0.62+1.00)/3 = 0.76

Table-VI: Hypothetical Dataset-2

Cluster (C)	Ground Truth (T)			Sum (n_i)
	T_1	T_2	T_3	
C_1	0	30	25	**55**
C_2	0	20	5	**25**
C_3	20	0	0	**20**
Sum (m_j)	**20**	**50**	**30**	**100**

Table-VII: Purity, Precision, Recall and F-measure based on Hypothetical Dataset-2

Metric	Cluster			Total
	C_1	C_2	C_3	
Purity	30/55 = 0.55	20/25 = 0.80	20/20 = 1.00	(0.55+0.80+1.00)/3 = 0.78
Precision	30/55 = 0.55	20/25 = 0.80	20/20 = 1.00	(0.55+0.80+1.00)/3 = 0.78
Recall	30/50 = 0.60	20/50 = 0.40	20/20 = 1.00	(0.60+0.40+1.00)/3 = 0.67
F-measure	60/105= 0.57	40/75 = 0.53	40/40 = 1.00	(0.57+0.53+1.00)/3 = 0.70

*(ii) **Internal Measures***

Internal measures are based on unsupervised learning to assess the goodness of clustering by employing criteria derived from the dataset itself [41, 45]. These are mostly based on two major criteria intra-cluster compactness (or cohesion) and inter-cluster separation. There is a trade-off to maximize inter-cluster separation and intra-cluster compactness. Both the criteria are described below:

- **Cluster Compactness or Cohesion**– Cluster cohesion refers to how data observations are closely related in a cluster. Variance is the common measure of it [41]. It is measured using *Sum of Squares of distances within Cluster* (SS_W) which should be minimized. SS_W of cluster C_i and average SS_W for all

clustering *C* are defined by the eq. *13* and *eq. 14* respectively.

$$SS_W(C_i) = \sum_{l \in S_k} \sum_{j=1}^{k} (x_{lj} - \bar{x}_{ij})^2 \quad (13)$$

$$SS_W(C) = \frac{1}{k} \sum_{i}^{k} \sum_{l \in S_k} \sum_{j=1}^{k} (x_{lj} - \bar{x}_{ij})^2 = \frac{1}{k} \sum_{i=1}^{k} SS_W(C_i) \quad (14)$$

- **Cluster Separation** – Cluster separation refers to how clusters are well-separated or distinct from other clusters [41]. It is measured using *Sum of Squares of distances between* Cluster Centroids (SS_B). Let the i^{th} and j^{th} Clusters be x_i and x_j respectively; then distance between clusters x_i and x_j is defined by the *eq. 15*.

$$D_{ij} = \sqrt{(x_i - x_j)^2} \quad (15)$$

The average distance among all clusters may be defined by the *eq. 16*.

$$D = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{k} \sum_{j=i+1}^{k} D_{ij} \quad (16)$$

In *eq. 16*, D is the SS_B of the clustering. $SS_B(C)$ may also be defined by the *eq. 17*.

$$SS_B(C) = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \quad (17)$$

C. Results and Discussions

The results of comparative empirical evaluation of basic k-Means, P-k-means and MP-k-means algorithms are presented in *Table VIII* to *Table XIV*. The results of all these three methods are evaluated and compared based on (i) Performance of Clustering i.e. # of Iterations taken to converge, (ii) Accuracy of Clustering, (iii) Intra-cluster Compactness (i.e. SS_W), (iv) Inter-cluster Separation (i.e. SS_B), (v) Purity / Precision, (vi) Recall and (vii) F-Measure. As the purity and precision gives the same result hence the results of both purity and precision is presented in the same table.

Comparative performance of all three methods is presented in *Tables VIII*. In *Table IX*, the accuracy based on cluster assignments compared with ground truth is presented. *Table X* and *Table XI* present the SS_W and SS_B respectively. In *Table XII* to *Table XIV*, the Purity / Precision, Recall and F-Measure of the clustering compared based on the ground truth are presented respectively.

Table-VIII: Performance i.e. # of Iterations taken to Converge

Dataset	Basic K-means	P-k-means	MP-k-means
Animal Milk	38.42	38.90	46.01
Image Segmentation	13.89	13.79	13.87
IRIS	9.22	8.62	9.11
Pen Digit	28.65	27.69	27.08
Spambase	7.13	6.87	7.63
Wine	11.87	10.68	10.86

Table-IX: Accuracy

Dataset	Basic K-means	P-k-means	MP-k-means
Animal Milk	96.96%	97.49%	97.80%
Image Segmentation	97.09%	97.19%	97.25%
IRIS	88.81%	88.83%	88.85%
Pen Digit	75.89%	76.09%	76.00%
Spambase	98.92%	98.92%	98.83%
Wine	71.70%	72.57%	71.09%

Table-X: Intra-cluster Compactness (i.e. SSW)

Dataset	Original Values			Normalized Values		
	Basic K-means	P-k-means	MP-k-means	Basic K-means	P-k-means	MP-k-means
Animal Milk	7.67	7.28	6.91	100.00	48.52	0.00
Image Segmentation	4274503.45	4380699.66	4364428.36	0.00	100.00	84.68
IRIS	27.14	26.83	26.40	100.00	57.65	0.00
Pen Digit	3510918.33	3502339.52	3496404.27	100.00	40.89	0.00
Spambase	620211350.51	620211350.51	617591265.95	100.00	100.00	0.00
Wine	841436.80	867154.49	829168.91	32.30	100.00	0.00

Table-XI: Inter-cluster Separation (i.e. SS_B)

Dataset	Original Values			Normalized Values		
	Basic K-means	P-k-means	MP-k-means	Basic K-means	P-k-means	MP-k-means
Animal Milk	682.73	692.07	698.35	0.00	59.76	100.00
Image Segmentation	2273924.87	2317511.20	2240836.25	43.16	100.00	0.00
IRIS	13.05	13.08	13.13	0.00	37.54	100.00
Pen Digit	109851.99	109844.49	109119.04	100.00	98.98	0.00
Spambase	7088425076.79	7088425076.79	6970417099.94	100.00	100.00	0.00
Wine	296495.06	305144.18	291245.56	37.77	100.00	0.00

Table-XII: Precision

Dataset	Basic K-means	P-k-means	MP-k-means
Animal Milk	0.9582	0.9643	0.9660
Image Segmentation	0.8042	0.8028	0.8049
IRIS	0.8972	0.9020	0.9004
Pen Digit	0.7718	0.7651	0.7712
Spambase	0.6671	0.6875	0.6678
Wine	0.7406	0.7433	0.7351

Table-XIII: Recall

Dataset	Basic K-means	P-k-means	MP-k-means
Animal Milk	0.9160	0.9280	0.9320
Image Segmentation	0.2162	0.2175	0.2106
IRIS	0.8807	0.8869	0.8877
Pen Digit	0.7167	0.7202	0.7231
Spambase	0.5073	0.5265	0.5080
Wine	0.6561	0.6474	0.6650

Table-XIV: F-Measure

Dataset	Basic K-means	P-k-means	MP-k-means
Animal Milk	0.9125	0.9246	0.9280
Image Segmentation	0.1151	0.1163	0.1100
IRIS	0.8789	0.8852	0.8862
Pen Digit	0.7079	0.7102	0.7139
Spambase	0.3945	0.4379	0.3959
Wine	0.6703	0.6632	0.6777

Table VIII shows that performance of MP-k-means is better than basic k-means for Image Segmentation, IRIS and Wine datasets whereas it is better than other two methods for Pen Digit dataset. Accuracy of MP-k-means is better as compared to other two methods for all datasets except Spambase and Wine datasets as shown in *Table IX*. *Table X* and *Table XI*, shows that SS_W and SS_B of MP-k-means are also better than that of other two methods. The F-Measure of MP-k-means is also better than that of other two methods for all datasets except Image Segmentation dataset as shown in *Table XIV*.

V. SUMMARY AND CONCLUSION

Basic k-Means algorithm is commonly used due its simplicity. The accuracy and performance of basic k-means is majorly affected due to the selection of initial cluster centroids. Hence, careful selection of initial cluster centroids is desired. A new method of initialization of the cluster centroids is proposed in this paper called as Modified Partition Based Cluster Initialization Method for k-means (MP-k-means). In MP-k-means, the dimensions of the data are partitioned in such a manner that if 'd' is the dimensionality of data, then 'd' lists consisting of 'k' equi-sized partitions based on positional average are created. Out of these 'd' lists, the centroids for initialization of the k-means algorithm are chosen in a random manner by choosing 'k' unique sets where each set is a collection of one arbitrary partition from each dimension. This ensures that no two cluster centroids are same. The proposed algorithm, MP-k-means is also easy to implement as it is also based on random selection of initial centroids. In MP-k-means, '*k*' centroids are also arbitrarily

chosen with the high probability for the closeness to the actual cluster centroids.

The empirical results presented in *Table VIII* through *Table XIV* show that MP-k-means is significantly better than basic k-means and P-k-means in terms of Performance, Accuracy, SS_W, SS_B, Purity / Precision, Recall and F-Measure. In view of the above, the MP-k-means outperformed the P-k-means algorithm. Though, the proposed method is suggested and implemented with k-means algorithm for careful selection of initial centroid. However, it may also be applied to other clustering algorithms which are based on selection of initial cluster centroids.

REFERENCES

1. Gupta, M.K., Chandra, P: P-k-means: k-means using partition based cluster initialization method, In Proc. ICACM 2019, Elsevier SSRN, pp 567-573 (2019)
2. Han J., Kamber M., Pei J.: Data Mining Concepts and Techniques, Elsevier, 3rd Edition (2012)
3. Arora R.K., Gupta M.K.: e-Governance using data warehousing and data mining, I.J. of Computer Applications 169(8), 28-31 (2017)
4. Jain, A. K., Dubes, R. C.: Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ 1988)
5. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review, ACM Comput. Surv. 31(3) (1999)
6. Gupta, M.K., Chandra, P.: An empirical evaluation of k-means clustering algorithm using different distance/similarity metrics, In Proc. of ICETIT 2019 Emerging Trends in Information Technology, LNEE 605 pp 884-892 DOI: https://doi.org/10.1007/978-3-030-30577-2_79 (2019)
7. Jain, A.K.: Data clustering: 50 years beyond k-means, Pattern Recognition Letters, Elsevier 31, pp. 651-666 (2010)
8. Gupta, M.K., Chandra, P.: A comparative study of clustering algorithms, In Proc. of the 13th INDIACom-2019, 6th International Conference on "Computing for Sustainable Global Development", IEEE Xplore (2019)
9. Forgy E.: Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, J. Biometrics, 21(3), 768 (1965)
10. McQueen, J.B.: Some methods for classification and analysis of multi-variate observation, Symposium on Mathematical Statistics and Probability, University of California Press (1967)
11. Kaufman, L., Rousseeuw, P.J.: Finding groups in data - an introduction to cluster analysis, Wiley, Canada (1990)
12. Katsavounidis, I, Kuo, C. Zhang, Z.: A new initialization technique for generalized Lloyd iteration, IEEE, 1(10), 144-146 (1994)
13. Bradley, P.S., Fayyad: Refining initial points for k-means clustering, Proc. 15th Intl. Conf. on Machine Learning, San Francisco, CA, pp 91-99 (1998)
14. Pei, J., Fan, J., Xie, W.: A new initialization method of cluster centers. J. of Electron. 16(4), pp 320–326, https://doi.org/10.1007/s11767-999 -0033-3 (1999)
15. Khan, S.S., Ahmad, A.: Cluster Centre initialization algorithm for k-means clustering, Pattern Recognition Letters 25(11), pp 1293-1302 (2004)
16. Su, T., Dy, J.: A deterministic method for initializing k-means clustering, Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference, pp. 784 – 786 (2004)
17. Hathaway R.J., Bezdek J.C., Huband J.M.: Maximin Initialization for Cluster Analysis. In: Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2006. Lecture Notes in Computer Science, vol 4225. Springer (2006)
18. Arai, K., Barakbah, A.R.: Hierarchical k-means: an algorithm for centroids initialization for k-means, Rep. Fac. Sci. Engrg, Saga Univ. , vol. 36 (2007)
19. Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding, ACM-SIAM Symposium on Discrete Algorithms (SODA 2007) pp. 1–11 (2007)
20. Wu S., Jiang Q., Huang J.Z.: A new initialization method for clustering categorical data, In: Advances in Knowledge Discovery and Data Mining. PAKDD 2007. Lecture Notes in Computer Science, vol 4426. Springer (2007)
21. Kang P., Cho S.: K-means clustering seeds initialization based on centrality, sparsity, and isotropy. In: Intelligent Data Engineering and Automated Learning - IDEAL 2009. IDEAL 2009. Lecture Notes in Computer Science, vol 5788. Springer (2009)
22. Maitra, R.: Partition-optimization algorithms, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, pp. 144–157 (2009)
23. Xu, J., Xu, B., Zhang, W.: Stable initialization scheme for k-means clustering, Wuhan University Journal of Natural Sciences 14(1), pp 24–28 (2009) https://doi.org/10.1007/s11859-009-0106-z
24. Dang Y., Xuan Z., Rong L., Liu M.: A Novel Initialization Method for Semi-supervised Clustering. In: Knowledge Science, Engineering and Management. KSEM 2010. Lecture Notes in Computer Science, vol 6291. Springer (2010)
25. Naldi, M.C., Campello, R.J.G.B., Hruschka, E.R., Carvalho, A.C.P.L.F.: Efficiency issues of evolutionary k-means, Applied Soft Computing 11, pp. 1938–1952 (2011)
26. Reddy D., Mishra D., Jana P.K.: MST-based cluster initialization for k-means. In: Advances in Computer Science and Information Technology. CCSIT 2011. Communications in Computer and Information Science, vol 131. Springer (2011)
27. Bai, L., Liang, J., Dang, C., Cao, F.: A cluster centers initialization method for clustering categorical data, Expert Systems with Applications 39(9), pp 8022-8029 https://doi.org/10.1016/ j.eswa.2012.01.131 (2012).
28. Chen G.H.: Cluster center initialization using hierarchical two-division of a data set along each dimension. In: Advances in Computer Science and Information Engineering. Advances in Intelligent and Soft Computing, vol 168. Springer (2012)
29. Aldahdooh, R.T., Ashour, W.: DIMK-means: distance-based initialization methods for k-means clustering algorithms, I.J. Intelligent Systems and Applications 2, pp 41-51 (2013)
30. Goyal, M., Kumar, S.: Improving the initial centroids of k-means clustering algorithm to generalize its applicability, Journal of The Institution of Engineers (India): Series B 95(4), pp 345–350 https://doi.org/10.1007/s40031-014-0106-z (2014)
31. Duwairi, R., Abu-Rahmeh, M.: A novel approach for initializing the spherical k-means clustering algorithm, Simulation Modelling Practice and Theory 54, pp 49-63 https://doi.org/10.1016/j.simpat.2015.03.007 (2015)
32. Poomagal, S., Saranya, P., Karthik, S.: A novel method for selecting initial centroids in K-means clustering algorithm, International Journal of Intelligent Systems Technologies and Applications 15(3) https://doi.org/10.1504/IJISTA.2016.078347 (2016)
33. Dhanabal, S., Chandramathi, S.: Enhancing clustering accuracy by finding initial centroid using k-minimum-average-maximum method, International Journal of Information and Communication Technology 11(2) (2017) https://doi.org/10.1504/IJICT.2017.086252
34. Golasowski M., Martinovič J., Slaninová K.: Comparison of k-means clustering initialization approaches with brute-force initialization. In: Advanced Computing and Systems for Security. Advances in Intelligent Systems and Computing, vol 567. Springer (2017)
35. Kumar KM, Reddy ARM: An efficient k-means clustering filtering algorithm using density based initial cluster centers, Information Sciences, Vol 418–419, pp 286-301 https://doi.org/10.1016/j.ins. 2017.07.036 (2017)
36. Ismkhan, H.: I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means, Pattern Recognition 79, pp 402-413 (2018) https://doi.org/10.1016/j.patcog.2018.02.015
37. Nguyen, C.D., Duc, T., Duong, T.H.: K-means** – a fast and efficient K-means algorithms, International Journal of Intelligent Information and Database Systems 11(1) (2018) DOI: 10.1504/IJIIDS.2018.091595
38. Sandhya N., Sekar M.R.: Analysis of variant approaches for initial centroid selection in k-means clustering algorithm. In: Smart Computing and Informatics. Smart Innovation, Systems and Technologies, vol 78. Springer (2018)
39. Yu, S., Chu, S., Wang, C., Chan, Y., Chang, T.: Two improved k-means algorithms, Applied Soft Computing 68, pp 747-755, (2018) https://doi.org/10.1016/j.asoc.2017.08.032.
40. Kurada R.R., Kanadam K.P.: A novel evolutionary automatic clustering technique by unifying initial seed selection algorithms into teaching–learning-based optimization. In: Soft Computing and Medical Bioinformatics. Briefs in Applied Sciences and Technology. Springer (2019)
41. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques, Journal of Intelligent Information Systems, 17:2/3, pp 107-145 (2001)
42. Theodoridis, S., Koutroubas, K.: Pattern Recognition, 2nd Ed., Elsevier Academic Press (2003)
43. Rendón, E., Abundez, I., Arizmendi, A, Quiroz. E.M.: Internal versus external cluster validation indexes, International Journal of Computers And Communications 5(1), pp 27-34 (2011)
44. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms, Ann. Data. Sci., Springer (2015) https://doi.org/10.1007/s40745-015-0040-1
45. Cluster Analysis in Data Mining "https://www.coursera.org"

46. Gupta, M.K., Chandra, P: HYBCIM: Hypercube based cluster initialization method for k-means, International Journal of Innovative Technology and Exploring Engineering 9(10), pp 3584-3587 DOI: 10.35940/ijitee.J9774.0881019 (2019).

AUTHORS PROFILE

MANOJ KUMAR GUPTA is research scholar of University School of Information, Communication & Technology, GGSIPU, Delhi, India. He worked as Professor at Rukmini Devi Institute of Advanced Studies (Aff. to Guru Gobind Singh Indraprastha University), Delhi, India. He was also Dean Examination, Admission and Administration in the Institute as additional charge. He has more than 20 years of experience in teaching and administration. His interest areas are Database Systems, Data Warehousing and Data Mining. He has 4 books and 20+ international / national research papers to his credit.

PRAVIN CHANDRA received the M.Sc. degree in Physics from the University of Delhi, India, in 1993, the M.Tech. degree from the Indian School of Mines, Dhanbad, India, in 1998, and the Ph.D. degree from Guru Gobind Singh Indraprastha University in 2004. He is currently a Professor at the University School of Information, Communication and Technology, Guru Gobind Singh Indraprastha University, Delhi, India where he was the Controller of Examinations also as the additional charge. His research interests are in the areas of artificial neural networks, soft computing, finger print analysis, ad-hoc networks, and software engineering.