I J C T A, 8(5), 2015, pp. 2085-2094

A Variant of K-Means Clustering through Heuristic Initial Seed Selection for Improved Clustering of Data

R. Geetha Ramani* and Lakshmi Balasubramanian**

Abstract: Unsupervised clustering algorithms have been used in many applications to group the data based on relevant similarity metrics. K-Means clustering is one of the most widely used clustering techniques owing to its simplicity. Many improvements and extensions have been proposed for this algorithm in view to improve its performance. Out of the various dimensions that have been explored in this regard such as mean computation, centroid representation, initial seed/cluster centre selection and similarity calculation methods, the choice of initial cluster centre is found to have a profound impact in the performance of the algorithm. Existing methods chose the cluster centres either randomly or based on heuristics such as maximum distance property, maximum probability of the squared distance, points with maximum points lying close to it etc. In this paper, a strategy to select relevant initial cluster centres for two-cluster grouping problems is proposed based on the measures indicating the statistical distribution of the data in view to improve the clustering performance in terms of accuracy. These measures include minimum, maximum, median, mean and skew of the data. The algorithm is validated on datasets from UCI repository viz. Balance, BloodDonate, Diabetes, Ionosphere, Parkinsons and Sonar and synthetic datasets. The performance of the proposed algorithm is compared with K-Means and its variants and found to achieve better performance in terms of accuracy. An increase is accuracy of approximately 0.25%-18% is observed across the datasets.

Keywords: Data Mining, Unsupervised Learning; Clustering; K-Means; Initial Cluster Centre Selection

1. INTRODUCTION

The concept of data mining [1] finds its application in all fields in view to discover hidden patterns in voluminous data. The major techniques adopted in data mining are supervised and unsupervised techniques. The supervised techniques [2][3] require labelled training data for its learning while unsupervised techniques do not demand labelled training data. Clustering is one of the most extensively used unsupervised learning techniques. Clustering groups the data into clusters such that the instances within the cluster are of high similarity while those between clusters have less similarity or high dissimilarity. The similarity can be calculated in terms of various distance or similarity measures as suited for the application. Out of the many clustering algorithms, a few take the number of clusters to be grouped as input while the others compute it during the partitioning process. Out of the many clustering algorithms, K-Means algorithm [4] is one of the most commonly used clustering procedure [5][6]. Its overall process involves selecting random initial centres for the clusters and assigning the other instances to the selected centres based on the similarity measure. The selection of these initial centres is greatly related to the performance of the algorithm. The traditional K-Means selects these cluster centres randomly. Randomness gives inconsistent results and can lead to worst or best partitioning. Later, strategies were proposed to choose the initial seeds/centres. These heuristics considered the distance between the instances and chose cluster centres such that they are (i) points at maximum distance (ii) points with probability proportional to squared distance from existing cluster centres etc. in view to elevate the performance. These ideas also included partial randomness as the

* Associate Professor, College of Engineering, Guindy, Anna University, Chennai-600025

** Research Scholar, College of Engineering, Guindy, Anna University, Chennai-600025

first cluster centre was chosen randomly leading to inconsistent results. On analysing all combinations of instances as centres on a few datasets with less number of instances, it was found that the maximum possible accuracy was greater than those achieved by these methods. Hence the necessity for improved initial centre selection was realised.

This paper presents a strategy that involves the concept of statistical measures namely minimum, median, maximum, mean and skew of the data for initial seed selection. To the best of our knowledge, this is the first attempt on identifying initial centres utilising this idea. Since these measures represent the distribution of data, the heuristic is expected to achieve better performance in partitioning. The paper is structured as follows: Section 2 presents the various extensions of K-Means algorithm in different perspectives. Section 3 explains the proposed strategy in identifying the initial cluster centres incorporating the idea of statistical distribution of data. Section 4 reports the results when evaluated on UCI and synthetic datasets and Section 5 concludes the paper. The following section discusses the existing approaches towards unsupervised clustering.

2. LITERATURE SURVEY

Clustering utilises the concept of partitioning the given data based on some similarity measure. K-Means is one of the most popularly used clustering algorithms owing to its simplicity [5][6]. Many variations have been put forth towards improvement of this algorithm. To begin with, the original version of K-Means algorithm is presented. Then, a concise note on the variations is discussed.

K-Means algorithm [4] utilises the distance between two instances as its similarity measure. Lesser is the difference, more is the similarity between them. In a broader view, the algorithm initially selects k cluster centres randomly and assigns the instances to nearby centres. The centre of each group is then represented by the mean value of the instances. The process iterates until there is no change in mean of the cluster across iterations. The algorithm holds a rich variety of extensions in view to improve its performance. The variations in the algorithm pertain to strategies to calculate cluster means, statistical methods to illustrate the centroid, similarity calculations and selection of initial centres.

Two strategies viz. Forgy [7] and McQueen [8] approach for updating cluster centres were put forth. Forgy approach computes the mean of the cluster after assigning all the instances to a cluster while the McQueen approach computes the average after assigning every instance in the data to a cluster. Then, an approach towards initial seed selection in view that two points at maximum distance [9] could achieve better results was proposed. This methodology, initially selected a random centre. The subsequent centres were chosen such that it was at the maximum distance from the selected centre. This was done in view to increase the cluster radius and termed as Farthest First Clustering algorithm.

Subsequently in 1998, to handle with categorical data, mode was utilised to compute the cluster centre and the algorithm was termed as K-Modes clustering algorithm [10]. The algorithm was tested on soybean disease and credit approval data to exhibit the clustering performance. It was also used to cluster large real time datasets. In 2002, the concept of fuzzy logic was incorporated to K-Means clustering [11]. This technique, Fuzzy C-Means, assigned a degree of belonging for every data point to clusters, rather than completely to one cluster. A co-efficient was assigned to every data point describing the degree of being in a cluster. With fuzzy C-Means, the centred of a cluster is the mean of all points, weighted by their degree of belonging to the cluster: The algorithm minimized intra-cluster variance. The algorithm was applied to estimate bias correction in MRI brain datasets.

In 2007, another heuristic to select initial seeds was proposed based on the intuition that spreading the initial cluster centres could improve the performance of clustering and termed it as K-Means++ algorithm [12]. In this approach, the first cluster centre was chosen uniformly at random from the entire set of points, after which each subsequent cluster centre was chosen from the remaining instances with probability

proportional to its squared distance from the point's closest existing cluster centre. The algorithm yielded better error rates and early convergence. The algorithm was tested on four datasets namely NORM-10, NORM-25, Cloud and Intrusion dataset and exhibited reduced error when compared to K-Means algorithm.

Then, in 2009, an alternative approach was postulated to represent the cluster centre [13]. In scenarios, where the updated centre must be one of the cluster instances itself, median was used to compute the updated cluster centre, leading to K-Medoids algorithm. In 2014, Jaskowlak et al [14] have analysed the impact of selection of distance measures during clustering. The study was made on fifteen different distance measures belonging to the category of correlation co-efficient, traditional distance measures and time series specific distance measures with microarray gene expression data and the effect of appropriate distance measure was realised. The distance measures were implemented and test using traditional K-Medoids algorithm. Again in 2015, impact of distance metrics viz., Manhattan and Euclidean distance on K-Means clustering in KDD99 Network Intrusion Dataset was presented [15]. The random seed selecting K-Means was utilised for this purpose.

The study on the previous works highlights the impact of the initial seed selection on the performance of clustering. In this work, a heuristic to choose initial seeds is put forth by incorporating the idea of statistical distribution of the data identified by its minimum, median, mean, maximum and skew. The next section explains the proposed heuristic to identify the initial cluster centres.

3. PROPOSED METHODOLOGY

A variant of K-Means algorithm is proposed by incorporating a heuristic for initial cluster centre selection through statistical distribution of data. K-Means clustering, one of the most widely used clustering methods

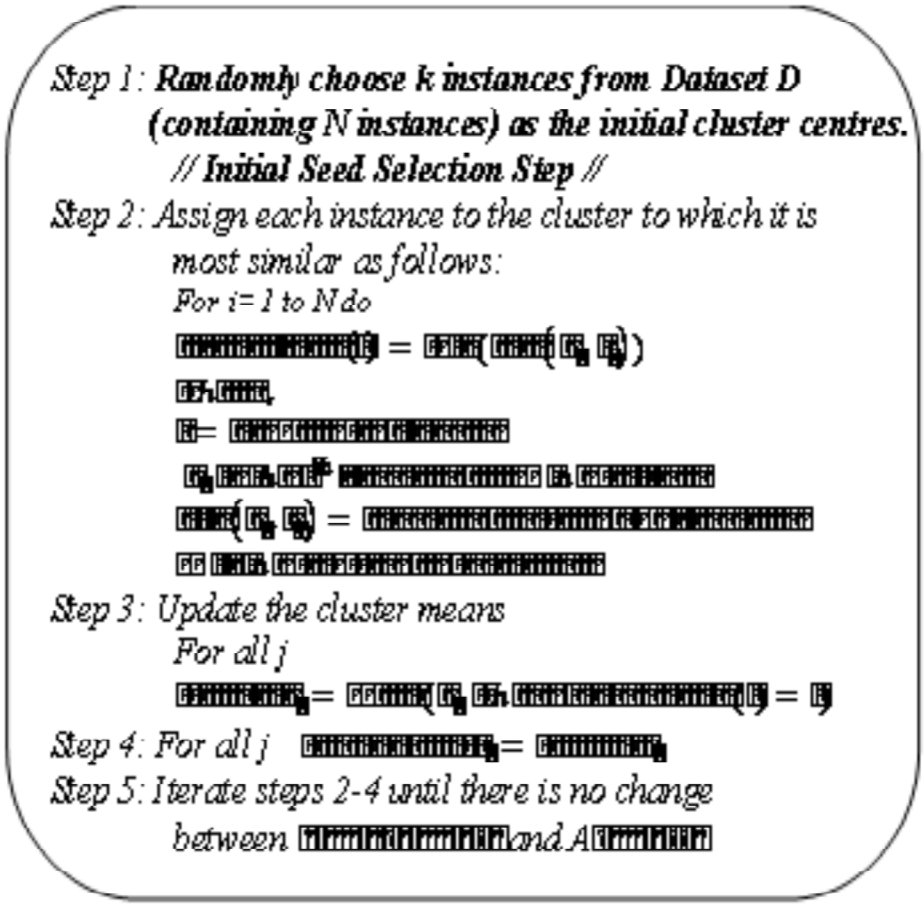

Figure 1: Procedure for K-Means Algorithm

to partition the data into groups, utilises random selection to choose these cluster centres. The procedure of K-Means clustering [4-6] is presented in Figure 1 [5] for immediate reference.

The algorithm first randomly selects *k* instances as the initial cluster centres. Then the remaining instances are assigned to the cluster to which it is the closest. After assigning all instances to any one of the cluster, the cluster centre is updated to the mean of the cluster. Again the instances in the data are allotted to the cluster based on the updated cluster centre. This process continues until the cluster centre remains unchanged across the iterations. Once the process stops, the instances are allocated the clusters to which they are currently assigned to.

It was observed that the efficiency of the algorithm is greatly depended on the selection of initial cluster centres. The proposed methodology utilises statistical measures determining the distribution of data such as minimum, median, mean, maximum and skew for this task. Since skew is meaningful for continuous data, the algorithm suits well for numeric data [16]. The algorithm has been designed to select two initial clusters. Hence the application which demands two group partitions can take advantage of this methodology. Image segmentation applications [5][6], where the region of interest is the foreground and the remaining forms the background, are perfect applications in this regard. The proposed algorithm involving identification of initial cluster centres through statistical distribution of the data is depicted in Figure 2.

The algorithm is explained with sample data points (1, 3) (12, 4) (3, 2) (5, 3) (4, 5) (2000, 1000) (3000, 2000) and (6000, 1000). The first five data points are expected to be in one group while the remaining is expected to lie in another group. With this as an example case, the proposed algorithm is detailed below.

The proposed algorithm initially computes the skew of the entire data by computing the sum of the skew of the individual dimensions. In the presented example, the skew of first dimension and second dimension were computed as 1.2993 and 1.0629 and hence the entire skew is 2.3621. Then, a minimum feature vector is formulated. To obtain minimum feature vector, the covariance of all attributes of the data with the first attribute is calculated. In the presented example, the covariance of second attribute with first attribute resulted in a value of $1.2092*10^6$ indicating it has a positive covariance with the first attribute.

Then, minimum feature vector is formed such that if the attribute has a positive or zero covariance, then the minimum value of the attribute is taken; else the maximum value of the attribute is chosen. Hence the minimum feature vector in the discussed scenario is (1,2). Then the distance of every data instance is computed with respect to the minimum feature vector formulated. The distance calculation on the presented scenario revealed that the instances (in order) are at 0.0005, 0.0021, 0.0003, 0.0008, 0.0016, 0.6004, 1.1100 and 1.1178 units from the minimum feature vector. The data instances are then sorted in ascending order based on the distance obtained from the previous calculation. The instances then take the order as follows: (3,2) (1,3) (5,3) (4,5) (12,4) (2000,1000) (6000,1000) and (3000,2000). The instances that are at minimum, median and maximum distance from the minimum feature vector are shortlisted as candidate for first cluster centre (FCC) and investigated further. In the considered case the candidates for first cluster centre (FCC) are (3,2) (4,5) (12,4) and (3000,2000).

Now, skew of the data is calculated after removing the minimum, median and maximum instance with replacement (remskew). The difference (diff) between the skew of the data after removing the instance and skew of the entire data is computed. For the given scenario, these values were calculated as 0.3981, 0.3971, 0.3977 and-0.2507. The instance, whose removal led to the minimum difference, is chosen as the first cluster centre (FC). Hence the maximum instance (3000.2000) was selected as the FC in this case.

Then, the other centre is chosen based on the following notion. Having chosen the FC, it should be either the minimum, maximum or median instance (In the example, it is the maximum instance). A mean instance is formulated for the subsequent investigations. In the example, the mean instance corresponds to (2000, 1000). The mean would lie between the median and maximum, if the data is positively skewed. It

Step 1: Compute the skew of entire data by summing the skew of every dimension.
Skew of every dimension A_j is calculated as follows:

$$\text{skew}(A_j) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\right)^3}$$

Where $\bar{x}_j$ is the mean of the data and n is the number of samples

$$\text{entireskew} = \sum_{j=1}^{m} \text{skew}(A_j)$$

Where m is the number of dimensions

Step 2: Find the co-variance of all attributes with respect to the first attribute

$$\text{cov}(A_1, A_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{ij} - \bar{x}_j)$$

Step 3: Compute a minimum feature vector such that
For each attribute, find the
min_value if it is has positive or zero covariance
max_value if it has negative covariance.
Step 4: Find euclidean distance between the formed feature vector and every instance of the dataset and arrange the data in ascending order with respect to its distance from the minimum feature vector.
Step 5: The instances having minimum, median and maximum distance are chosen as candidates for the first centre (FCC).
Step 6: For each instance I in FCC, remove the instance and find skew of the remaining instances remskew(I)
Step 7: Find its difference diff(I) =entireskew-remskew (I).
Step 8: The first centre (FC) is the instance I which has maximum diff(i) and the corresponding remskew be skewfc.
Having chosen the centre for the first cluster, it should be either the minimum, median or maximum instance
Step 9: Formulate a mean feature vector and find the instance which lies very close to this feature vector. This is the mean instance.
Step 10: Candidates for second cluster centre (SCC) is chosen as below:

```
if FC is minimum instance
    If skew of entire data is positive or zero
        SCC = median to maximum instance
    Else    SCC=mean to maximum instance
    End
Else if FC is a maximum instance
    If skew of entire data is positive or zero
        SCC = minimum to mean instance
    Else    SCC = minimum to median instance
    End
Else
    If skew of entire data is positive or zero
        SCC = mean to maximum instance
    Else    SCC = minimum to mean instance
    End
End
```

Step 11: For each instance I of SCC,
Remove the instance and find he skew of the remaining instances remsccskew (1) and calculate distance d(I) and covariance cv(I) between FC and SCC(I).
Then calculate the heuristic metric:
*Metric (I)=d(I) *covar(I)/(skewfc-remsccskew (I))*
Step 12: Instance corresponding to highest metric is chosen as SC.
Step 13: With the two centres fixed, assign all the other instances based on its closeness to the cluster centre.

Figure 2: Proposed Clustering Algorithm

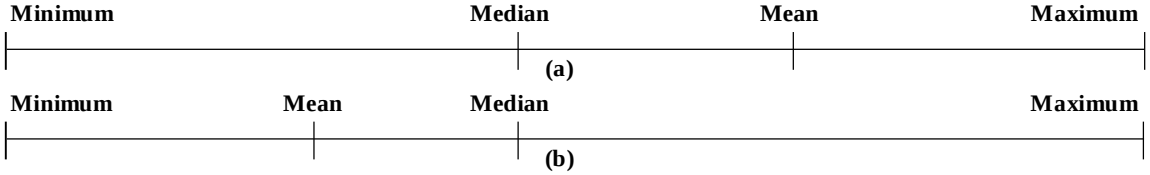

Figure 3: Statistical distribution of (a) Positively and (b) negatively skewed data

would be located between the minimum and median if the data is negatively skewed. Figure 3 illustrates this idea.

So, if the FC is minimum or maximum instance, the other centre (SC) that produces best grouping is expected to be in another subset, where the subsets are separated by median or by mean instance. The superset between the two subsets is always chosen as the candidates for second cluster centre (SCC).

Based on this concept, SCC are chosen as follows: If FC is a minimum instance and entire skew is positive, the SC that yielded best partitioning is expected to lie between the median instance and the maximum instance (superset between subsets formed from (i) median instance and maximum instance and (ii) mean instance and maximum instance). Otherwise, if FC is minimum and skew is negative, and then the SC is expected to lie between the mean and the maximum instance. Similarly, if the FC is a maximum instance, then the candidates for SC would be located between minimum and mean instance if the data is positively skew while it is expected to lie between minimum and median instance if the data is negatively skewed. If the FC is a median instance and the data is positively skewed, then the other centre would lie between mean and the maximum instance. If the FC is a median instance and the data is negatively skewed, then SC is chosen between minimum and mean instance. Considering the given scenario, the data points are positively skewed. It can also be seen that the mean instance lies between the median and the maximum instance. In this case, the SCC includes the instances that lie between minimum and mean instance viz. (3,2) (1,3) (5,3) (4,5) (12,4) and (2000,1000).

Out of the instances in SCC, the distance and covariance between each candidate and FC and skew of the data after removal of each candidate and FC is calculated and a metric is formulated usi8ng these measures. The metric values computed for the considered second cluster candidate corresponds to-234, 469, -469, 234, -1873 and -68897. The SC is the instance that has the highest merit ((1,3) in this case). Having chosen the cluster centres, the remaining instances are assigned to the closest centre. Hence first five data points were grouped to one group while the remaining was put in another group. Table 1 presents the cluster grouping on sample data points by K-Means, K-Means++, Farthest First, FCM and proposed algorithm. C1 and C2 refer to cluster grouping to cluster 1 and cluster 2.

Table 1
Cluster Assignments for the sample data points

Points	*K-Means*	*Farthest First*	*FCM*	*K-Means++*	*Proposed*
(1, 3)	C1	C1	C1	C1	C1
(12, 4)	C1	C1	C1	C1	C1
(3, 2)	C1	C1	C1	C1	C1
(5, 3)	C1	C1	C1	C1	C1
(4, 5)	C1	C1	C1	C1	C1
(2000, 1000)	C1	C1	C1	C1	C2
(3000, 2000)	C2	C1	C2	C2	C2
(6000, 1000)	C2	C2	C2	C2	C2

The initial cluster centres assigned by K-Means are (1,3) and (5,3), Farthest First correspond to (3,2) and (6000,1000); K-Means++ are (3000,2000) and (6000,1000); FCM are (1475.8, 437.4) and (1190, 465.8) and the proposed algorithm are (3000,2000) and (1,3). The algorithm was validated on different datasets from UCI repository and synthetic datasets. The experimental results are elaborated in the next section.

4. RESULTS AND DISCUSSION

The experiments to exhibit the performance of the proposed technique were implemented using Matlabr2008a and Weka, a data mining tool. The performance is assessed in terms of accuracy [17]. The

experiments were conducted on datasets viz. Balance, Blood Donate, Diabetes, Ionosphere, Parkinsons and Sonar datasets obtained from UCI repository [18]. The methodology was also evaluated on synthetic datasets generated from two random data generators in Weka for varying number of instances and attributes. The first random generator (RDG) [19] generates data randomly by producing a decision list, consisting of rules. Instances are generated randomly one by one. If decision list fails to classify the current instance, a new rule according to this current instance is generated and added to the decision list. The second random generator (RandomRBF) [19] initially creates a random set of centres for each class. Each centre is randomly assigned a weight, a central point per attribute, and a standard deviation. To generate new instances, a centre is chosen at random taking the weights of each centre into consideration. The particular centre chosen determines the class of the instance. RandomRBF data contains only numeric attributes as it is non-trivial to include nominal values. As a reference, the datasets generated by RDG and RandomRBF are prefixed with their identities in the dataset names. The details of the datasets are narrated in Table 2.

Table 2
Details of Experimental Data

Dataset	*Number of Instances*	*Number of attributes*
Balance	19	2
Blooddonate	748	4
Diabetes	768	8
Ionosphere	351	34
Parkinsons	196	22
Sonar	208	60
Random RBF1	997	97
Random RBF2	100	12
Random RBF3	1020	11
Random RBF4	100	1012
RDG1	1000	103
RDG2	115	10
RDG3	1000	13
RDG4	100	1009

Initially experiments were performed on a few datasets to analyse the maximum possible accuracy that could be achieved. This was done through exhaustively assigning all pair of instances as cluster centres. Since the search was exhaustive, a few datasets that had only a few instances ranging from ~20 to 50 were considered. The maximum accuracy achieved was higher than that obtained through the popular heuristics. Hence, the need for new heuristics was realised and an attempt was made through the concept of statistical measures that represent the distribution of data. The results show encouraging performance. Farthest First algorithm, FCM and K-Means++ have been widely used in many applications owing to their promising results [20-22]. Hence, Table 3 presents the accuracy obtained through the proposed approach with the accuracy obtained through these algorithms (Farthest First Algorithm, FCM, K-Means++). The comparison is also presented as against original K-Means algorithm as it gives good results many a times due to its randomness.

The reported results in Table 3 exhibit the improved performance achieved through the proposed technique. It is also observed that K-Means outperforms FCM, Farthest First and K-Means++ in some cases and vice versa. The graphical representation of the performance of the proposed clustering is projected in Figure 4.

Table 3
Performance comparison of the proposed clustering with existing methods

Dataset	*K-Means*	*Farthest First*	*FCM*	*K-means++*	*Proposed*
Balance	58.82	58.82	58.82	58.82	**76.47**
Blooddonate	58.82	70.59	70.72	70.59	**72.33**
Diabetes	66.80	65.76	65.89	66.02	**67.84**
Ionosphere	53.28	62.11	70.94	70.94	**71.23**
Parkinsons	62.76	71.94	71.79	72.82	**80.00**
Sonar	51.45	50.48	53.84	54.33	**55.29**
Random RBF1	53.61	52.58	49.48	57.73	**61.86**
Random RBF2	65.00	57.00	63.00	66.00	**76.00**
Random RBF3	51.43	51.23	51.92	49.36	**53.50**
Random RBF4	51.00	61.00	53.00	55.00	**57.00**
RDG1	50.00	56.30	52.20	53.40	**69.30**
RDG2	52.73	47.27	58.18	63.64	**65.46**
RDG3	47.20	54.30	59.70	52.40	**66.20**
RDG4	58.00	58.00	52.00	55.00	**67.00**

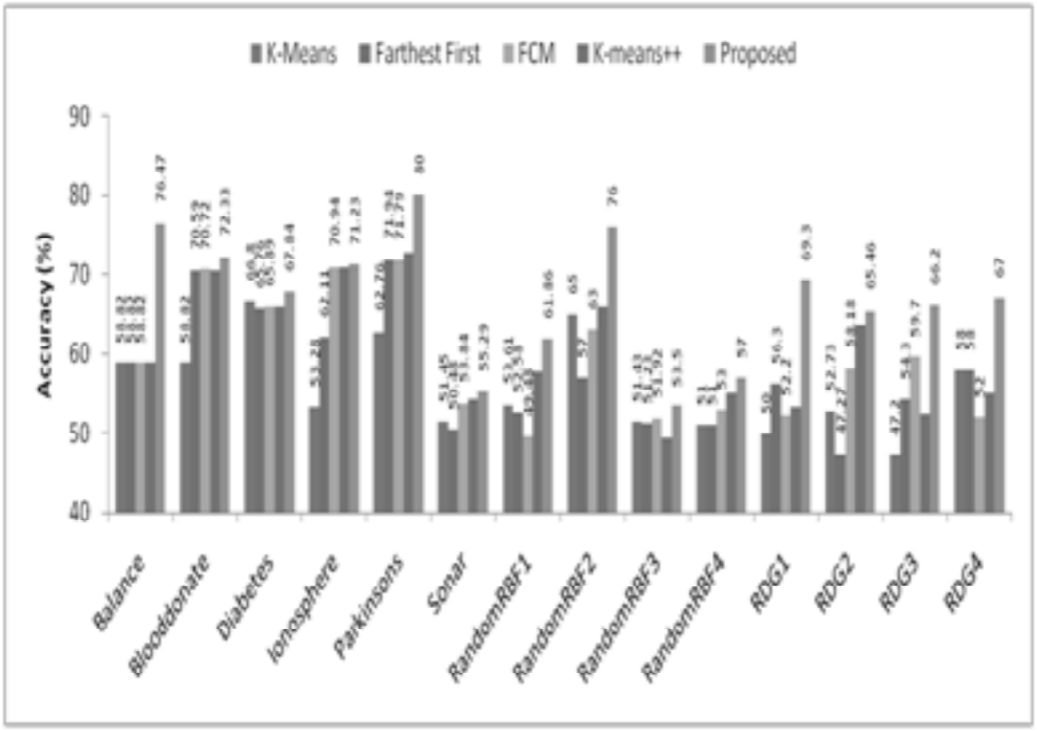

Figure 4: Performance of proposed clustering method

It is seen that the proposed technique exhibits consistent improvement in all datasets. Results encourage the utilisation of the proposed technique in real world scenarios.

5. CONCLUSION

Unsupervised clustering techniques have been widely used in many applications to partition the data without the need of the class label. K-means clustering have been most widely used algorithm in clustering. Many extensions and variations of the algorithm have been proposed in different perspectives in view towards

improvement in performance. The proposed heuristic in this paper targeted on the selection of initial cluster centres. The measures representing the statistical nature of the data (minimum, median, mean, maximum and skew) was utilised to find the initial cluster centres. The methodology was validated on datasets from UCI repository and synthetic datasets generated through Weka tool. Results reported exhibit the improved performance of the proposed heuristic. The proposed method works well suited for applications with continuous data. This work targeted on two group partitioning. Since, image segmentation applications mostly extract features that are continuous and requires two groups partitioning, the algorithm should be useful in this case. Future enhancements could be testing its usefulness in image segmentation applications and/or could be in the direction of either increasing the number of initial centres of partitioning the formulated clusters.

References

[1] J. Han, M. Kamber, & J. Pei, Data Mining: Concepts and Techniques (Third Edition), Morgan Kaufmann Publishers, 2011.

[2] R. GeethaRamani, B. Lakshmi & A. Alaghu Meenal, "Hybrid Decision Classifier Model Employing Naive Bayes and Root Guided Decision Tree for Improved Classification", International Journal of Applied Engineering Research. Vol. 10, no. 17, pp. 13245-13249, 2015.

[3] R. GeethaRamani, B. Lakshmi & A. Alaghu Meenal, "Decision Tree Variants (Absolute Random Decision Tree and Root Guided Decision Tree) for Improved Classification of Data", International Journal of Applied Engineering Research. Vol. 10, no. 17, pp. 13190-13195, 2015.

[4] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, Inc., 1975.

[5] R. GeethaRamani & B. Lakshmi, "Automatic Segmentation of Blood Vessels from Retinal Fundus Images through Image Processing and Data Mining Techniques", Sadhana. Vol. 40, no. 6, pp. 1715-1736, 2015.

[6] R. GeethaRamani & B. Lakshmi, "Retinal Blood Vessel Segmentation employing Image Processing and Data Mining Techniques for Computerized Retinal Image Analysis", Biocybernetics and Biomedical Engineering. Vol. 36, no. 1, pp. 102-118, 2015.

[7] E.W. Forgy, "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classification", Biometrics. Vol. 21, no. 3, 1965.

[8] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, no. 14, pp. 281-297, 1967.

[9] D.S. Hochbaum, "A Best Possible Heuristic for the K-Center Problem", Mathematics of Operations Research. Vol. 10, no. 2, pp. 180-184, 1985.

[10] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery. Vol. 2, no. 3, pp. 283-304, 1998.

[11] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag & T. Moriarty, "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data", IEEE Transactions on Medical Imaging. Vol. 21, no. 3, pp. 193–199, 2002.

[12] D. Arthur & S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding", Eighteenth Annual ACM-SIAM symposium on Discrete algorithms, Philadelphia, USA, pp. 1027-1035, 2007.

[13] H-S. Park, & C-H. Jun, "A Simple and Fast Algorithm for K-Medoids Clustering", Expert Systems with Applications. Vol. 36, pp. 3336-3341, 2009.

[14] P.A. Jaskowlak, R.J. Campello & I.G Costa, "On the Selection of Appropriate Distances for Gene Expression Data Clustering", BMC Bioinformatics. Vol. 15 (Suppl 2), no. 2, 2014.

[15] H. Nabooti, M. Ahmadzadeh, M. Keshtgary & S. Varid Farrahi, "The Impact of Distance Metrics on K-Means Clustering Algorithm using Network Intrusion Detection Data", International Journal of Computer Networks and Communications Security. Vol. 3, no. 5, pp. 225-228, 2015.

[16] David M. Lane, "Online Statistics Education: An Interactive Multimedia Course of Study", Available at http://onlinestatbook.com/2/index.html

[17] R. GeethaRamani, B. Lakshmi & J.G Shomona, "Data Mining Method of Evaluating Classifier Prediction Accuracy in Retinal Data", IEEE International Conference on Computational Intelligence & Computing Research, Coimbatore, India, pp. 426-429, 2012.

[18] M. Lichman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. Available at [http://archive.ics.uci.edu/ml].

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I.H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations. Vol. 11, no. 1, 2009.

[20] Priyanka Sharma, "Comparative Analysis of Various Clustering Algorithms using WEKA", International Research Journal of Engineering and Technology. Vol. 2, no. 4, pp. 107-112, 2015.

[21] R. GeethaRamani, S. SIvashankari & B. Lakshmi, "Automatic Concept Clustering for Ontological Structure through Data Mining Techniques", 2013 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1-9. 2013.

[22] S.S. Lee, D. Won & D. McLeod, "Discovering Relationships among Tags and Geotags", Second International Conference on Weblogs and Social Media, Seattle, Washington, USA, AAAI Press, 2008.