Automatic Clustering Approaches Based On Initial Seed Points

G.V.S.N.R.V.Prasad
Department of Computer Science &Engineering,
Gudlavalleru Engineering College
Gudlavalleru,A.P.India
gutta.prasad1@gmail.com

V.Venkata Krishna
Professor & Principal,
C.I.E.T.,
Rajamundry,
A.P, India.
vakula_krishna@yahoo.co.in

V.Vijaya Kumar
Dept of Computer Science
Engineering & Information Technology,
G.I.E.T,
Rajamundry,A.P,India
vijayvakula@yahoo.com

***Abstract*--Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in the literature. This paper proposes a novel approach to address the major problems in any of the partitional clustering algorithms like choosing appropriate K-value and selection of K-initial seed points. The performance of any partitional clustering algorithms depends on initial seed points which are random in all the existing partitional clustering algorithms. To overcome this problem, a novel algorithm called Weighted Interior Clustering (WIC) algorithm to find approximate initial seed-points, number of clusters and data points in the clusters is proposed in this paper. This paper also proposes another novel approach combining a newly proposed WIC algorithm with K-means named as Weighted Interior K-means Clustering (WIKC). The novelty of this WIKC is that it improves the quality and performance of K-means clustering algorithm with reduced complexity. The experimental results on various datasets, with various instances clearly indicates the efficacy of the proposed methods over the other methods.**

***Keywords-- Clustering, partitioning, data mining, unsupervised learning, hierarchical clustering, k-means*.**

I. INTRODUCTION

Most clustering algorithms are based on two popular techniques known as hierarchical and partitional clustering [3, 7]. Partitional clustering algorithms divide the data set into a specified number of clusters. These algorithms try to minimize certain criteria (e.g. a squared error function) and can therefore be treated as optimization problems. However, these optimization problems are generally NP-hard and combinatorial [7]. The advantages of hierarchical algorithms happen to be the disadvantages of the partitional algorithms and vice versa. Because of their advantages, partitional clustering techniques are more popular than hierarchical techniques in pattern recognition [4]. Hence our study concentrates on partitional techniques.

As the initial clustering is random because of random initial seed points, a different clustering solution may be found each time the algorithm is executed. To find the optimal K-means clustering solution (objective function), the algorithm should be repeated many times with different random initial seed-points.

There is no commonly accepted or standard “best” way to determine either the number of clusters or the initial starting point values. The resulting set of clusters, both their number and their centroids, depends on the specified choice of initial starting-point values. Two simple approaches to cluster initialization are either to select the initial values randomly or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen and the set, which is closest to optimal, is selected. However, testing different initial sets is considered impracticable especially for a large number of clusters. Therefore, different methods have been proposed to find the initial centroids given in the literature [10]. Fahim [1] proposed an efficient method for assigning data points to clusters. The original K-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim’s approach[1] makes use of two distance functions for this purpose, one similar to K-means algorithm and the other one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original K-means algorithm. Hence there is no guarantee for the accuracy of the final clusters. Fahim A. M [2] proposed a method to select a good initial solution by partitioning dataset into blocks and applying K-means to each block. But here the time complexity is slightly more. Though the above algorithms can help find good initial centres for some extent, they are quite complex and some researchers use the K-means algorithm as part of their algorithms, which still needs the use of the random method for cluster-centre initialization.

Nazeer [9] proposed an enhanced K-means to improve the accuracy and efficiency of the K-means clustering algorithm. In this algorithm two methods are used: one method for finding the initial centroids and another method for an efficient way of assigning data- points to appropriate clusters. Madhu Yedla [8] proposed a method to find better initial centroids with reduced time-complexity. For assigning the data points they followed the Fahims and Nazim's approach [1, 9]. This method first checks whether the given dataset contains the negative value attributes or not. Then it transform all the data-points in the dataset to the positive space by subtracting each data-point attribute with the minimum attribute-value in the given dataset. Next, calculate the distance for each data-point from the origin. Then it sorts the data points in accordance with the distances, and partition the sorted data point into k equal sets and the middle point in each set is taken as the initial centroid. This method[8] is used the heuristics approach to assign the data-points to the initial centroid. But all the above methods do not work well for high dimensional data sets. Thus, a direct implementation of the K-means method can be computationally very time consuming. This is especially true for typical data-mining applications with a large number of pattern vectors. To overcome the above disadvantages the present paper proposes two novel approaches called Weighted Interior Clustering (WIC) and Weighted Interior K-means Clustering (WIKC). The proposed novel methods produces comparable clustering results with much better performance by simplifying distance calculations and reducing total execution time.

The present paper is organized as follows.

The section two deals with a novel approach called Weighted Interior Clustering (WIC) algorithm. Proposed Weighted Interior K-means Clustering (WIKC) was explained in chapter three. Results analysis and conclusions are given in sections four and five respectively.

II. Weighted Interior Clustering (WIC) Algorithm based on Initial Seed Points

In any partitional clustering algorithm the number of runs of clustering algorithm to converge to its objective function depends on the selection of initial seed points. Heuristic methods like Mountain Method introduced by Yager and Filev [12] is a simple and effective approach for approximate estimation of initial cluster centres that are required for more complex cluster algorithms. This method is based on gridding on the space, construction of a mountain function and then destruction of a mountain to obtain cluster centres. It also requires 3 parameters (α,β,γ). The inadequate selection of parameters results in undesirable parasitic cluster centres, as the efficiency of clustering depends on the selection of these parameters.

To overcome these disadvantages of mountain method [12] requiring three parameters, called the Advanced Mountain Method [5]was proposed, which results in success with only one parameter 'w' instead of three parameters required in mountain method[12]. In this method the value of the parameter 'w' was selected on the basis of the average distance between data-points. This method determines the number and location of cluster centres but not the number of patterns in each cluster and cannot run efficiently when data dimension is increasing.

To overcome the above disadvantages 'Mountain Means Clustering' (MMC) Algorithm [6] was proposed. In the MMC algorithm a Hill valley function [11] is used to classify the patterns distributed on each mountain after the construction of a mountain function for each possible cluster centre. The patterns which are distributed on the same mountain are considered as points in the respective clusters and the centre as the mean of the data patterns belongs to the corresponding cluster. The MMC method can identify (a) the number of cluster centers, (b) the location of the cluster centres, and (c) patterns belonging to the corresponding clusters.

The main disadvantage of the MMC algorithm is it cannot form qualitative clusters. To overcome this, the present paper proposes a novel approach called Weighted Interior Clustering (WIC) algorithm. This WIC algorithm improves the quality and reduces the complexity by eliminating gridding and construction of the mountain function in the data space. The proposed WIC algorithm completely eliminates the gridding and construction of mountain function. The WIC also reduces the number of iterations over data set. This proposed WIC algorithm can be implemented for any dimension. The WIC method contains eight steps as given below.

Algorithm 1: Pseudo code of Weighted Interior Clustering (WIC) algorithm.

```
Begin
Step 1:   Choose the number of clusters K to 1
Step 2:   Create a new cluster $C_k$.
Step 3:   The pattern $x_i$ corresponds to $C_k$.
Step 4:   Consider pattern $x_{i+1}$.
Step 5:   Verify whether $x_{i+1}$ corresponds to $C_k$ or not by using Similarity function.
              • If Similarity function returns zero, it means the pattern $x_{i+1}$ corresponds to $C_k$.
```

```
        • Delete $x_{i+1}$ from the database .
        • Repeat steps 4 & 5 until the end of the database.
Step 6: Create another cluster $C_{k+1}$.
Step 7: Repeat steps 2-6 until all patterns are assigned to the respective clusters.
Step 8: Compute the cluster-centre which is the mean of patterns belonging to Ck.
End
```

The procedure for evaluating Similarity function for step (5) of WIC algorithm is given below. Similarity function is used to determine the relationship of two points in search-space. The Similarity function returns 0 when patterns x_i and x_j correspond to one cluster. The similarity function contains five steps.

Algorithm 2: Pseudo code of similarity function of WIC algorithm.

```
Begin
Step1:  Calculate the weight of $i_p$ and $i_q$ using weight function.
Step 2: Compare the weights of $i_p$ and $i_q$ and take the minimum weight as $min_{wt}$.
            $min_{wt}=min(weight(i_p),weight(i_q))$
Step 3: Consider the samples in-between $i_p$ & $i_q$ and calculate $i_{in}$ for each sample using equation 1.
            $i_{in}= i_p +( i_q - i_p).samples[j]$        (1)
Step 4: Calculate the weight of each $i_{in}$ .
Step 5: If the $min_{wt}$ is greater than twice the weights of $i_{in}$ for all the samples in between $i_p$ & $i_q$ then return 1
        otherwise return 0.
End
```

The Similarity function returns 0 if the weight of all the interior points is greater than the minimal weight of i_p, and i_q,, otherwise it returns 1. If the Similarity function returns 0 then i_p, and i_q are similar i.e, they belong to the same cluster, otherwise they do not belong to the same cluster.

The other novelty of the proposed WIC scheme lies with evaluating the weight function given in equation 2. The weight function of step(1) of similarity function is evaluated in the following way.

$$\omega_j = \sum_{i=1}^{N} \exp(-\frac{\left\| x_j - x_i \right\|^2}{2\mu^2}) \qquad (2)$$

The positive integer 'µ' can be calculated by using the average distance between data-points given in equations 3 to 6.

$$\mu = 2.34\tilde{d}^2 - 1.26\tilde{d} + 0.23 \qquad (3)$$

$$\tilde{d} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\left\| x_i - x_j \right\|_2}{d_{\max}}}{n} \qquad (4)$$

$$d_{\max} = (\sum_{j=1}^{s} D_j^2)^{1/2} \qquad (5)$$

$$D_j = \max_k (x_{kj}) - \min_k (x_{kj}) \qquad (6)$$

For a collection of n data points {xi, ...,x_n} in the S-dimensional space R^s, x_{kj} denotes the jth dimension of the kth data point, where k=l,2,3 .. . ,n and j=1,2,3,...,S.

III. WEIGHTED INTERIOR K-MEANS CLUSTERING (WIKC)

Further to improve the performance, a novel scheme called Weighted Interior K-means Clustering (WIKC) scheme is developed which is a combination of WIC and K-means algorithm.

IV. RESULT ANALYSIS

The proposed study of WIC and WIKC algorithms are tested with synthetic data-set generated from Gaussian distribution function for both uni-dimensional and multi-dimensional data sets. The major problem in all the clustering algorithms is visualization of clustering results. There is no proper tool to visualize the clustering results. Hence for better visualization of results the DATASET1 which is well- separated one-dimensional data set, DATASET2 which is multi-dimensional synthetic data set and DATASET-3 which is the most popular Iris real dataset (150 records with 3 class lables) is considered. All these data sets are tested and verified for the quality of clustering in terms of cluster density, *IntraCD*, *InterCD* and validity by Compactness Separation Index(CSI).

From Table (1) it is evident that the MMC algorithm has separated the given data set into three clusters and has given three centroids for clusters which are approximate initial seed points for any of the partitional clustering algorithms, but it is observed that some of the datapoints are wrongly clustered. For example, the data-points which are 0.3, 0.31 and 0.323 are assigned to cluster C1 instead o f clustering into C2 which is ranging from 0.3 to 0.39 as shown in Table (1). Similarly the data points 0.80 and 0.81 are assigned to C2 instead of cluster C3.

TABLE 1. COMPARISON OF ALGORITHMS FOR ONE DIMENSIONAL WELL SEPARATED SYNTHETIC DATA SET DATASET-1.

DATASET-1	Existing M-Means			Weighted Interiar			WIKC			K-means		
				Clustering(WIC)								
0.1	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
0.113	0.1	0.323	0.825	0.1	0.34	0.8	0.1	0.3	0.8	0.1	0.3	0.8
0.117	0.113	0.34	0.83	0.113	0.345	0.81	0.113	0.31	0.81	0.113	0.31	0.81
0.12	0.117	0.345	0.838	0.117	0.35	0.825	0.117	0.323	0.825	0.117	0.323	0.825
0.125	0.12	0.35	0.85	0.12	0.36	0.83	0.12	0.34	0.83	0.12	0.34	0.83
0.127	0.125	0.36	0.87	0.125	0.369	0.838	0.125	0.345	0.838	0.125	0.345	0.838
0.13	0.127	0.369	0.88	0.127	0.37	0.85	0.127	0.35	0.85	0.127	0.35	0.85
0.14	0.13	0.37	0.89	0.13	0.39	0.87	0.13	0.36	0.87	0.13	0.36	0.87
0.15	0.14	0.39	0.895	0.14		0.88	0.14	0.369	0.88	0.14	0.369	0.88
0.16	0.15	0.8		0.15		0.89	0.15	0.37	0.89	0.15	0.37	0.89
0.3	0.16	0.81		0.16		0.895	0.16	0.39	0.895	0.16	0.39	0.895
0.31	0.3			0.3								
0.323	0.31			0.31								
0.34				0.323								
0.345	Centre	Centre	Centre	Centre	Centre	Centre	Centre	Centre	Centre	Centre	Centre	Centre
0.35	0.157	0.445	0.859	0.17	0.36	0.848	0.128	0.345	0.848	0.128	0.345	0.848
0.36												
0.369												
0.37												
0.39												
0.8												
0.81												
0.825												
0.83												
0.838												
0.85												
0.87												
0.88												
0.89												
0.895												

Only few data points which belong to C2 are wrongly placed in cluster C1. This factor clearly indicates the improvement of the proposed WIC algorithm over Mountain Means Clustering algorithm. In the next step, centroids which are calculated from the clusters formed by the proposed WIC algorithm are given as input to K-means clustering algorithm known as WIKC. The results in Table(1) show that the clusters formed by the WIKC are more appropriate and accurate than the clusters formed by all the previous methods.

The same data set is given to K-means clustering with K=3 and the same is also given by WIC algorithm with three initial centroids. The results are similar to clusters formed by WIKC. This may not be true in all the data sets or in all the cases. In some of the cases K-means is unable to form as many clusters as WIKC because a randomly selected K initial seed points may not be appropriate to cluster K number of clusters, resulting in some empty clusters. But WIKC is rectifying this problem in most of the cases as it is selecting most appropriate initial seed points from WIC.

TABLE 2. COMPARISON OF ALGORITHMS FOR DATASET-1.

	Appointed Number of clusters	Gained Number of clusters	Number of Iterations Over Data Set	Custer Density	IntraCD	InterCD	CSI
MMC	--------	3	8	2.889	0.0461	0.19	0.243
WIC	--------	3	3	2.347	0.0406	0.19	0.213
WIKC	3	---------	4	1.372	0.021	0.217	0.097
K-Means	3	---------	7	1.372	0.021	0.217	0.097

Tables (2) to (4) show the unique phenomenon that the MMC and WIC algorithms gain a number of clusters. The gained clusters of WIC algorithm are assigned to K-means. This phenomenon of assigning K clusters to K-means is called WIKC. The least values in Compactness Separation Index(CSI) also show a better clustering scheme.

TABLE 3. COMPARISON OF ALGORITHMS FOR DATASET-2.

	Appointed Number of clusters	Gained Number of clusters	Number of Iterations Over Data Set	Custer Density	IntraCD	InterCD	CSI
MMC	--------	12	18	8.89	0.04215	0.15	0.281
WIC	--------	12	12	8.75	0.04102	0.15	0.274
WIKC	12	---------	13	8.19	0.0395	0.148	0.266
K-means	12	---------	17	8.19	0.0395	0.148	0.266

TABLE 4. COMPARISON OF ALGORITHMS FOR DATASET-3.

	Appointed Number of clusters	Gained Number of clusters	Number of Iterations Over Data Set	Custer Density	IntraCD	InterCD	CSI
MMC	------------	3	8	356.01	2.231	0.772	2.89
WIC	------------	3	3	298	1.9866	0.772	2.573
WIKC	3	--------	4	100.08	0.6671	1.9262	0.346
K-means	3	--------	7	100.08	0.6674	1.9262	0.347

Tables (1) to (4) show that the clusters formed by K-means and WIKC are the same. But the number of clusters and their centroids calculated by the proposed algorithm are not known in advance to implement K-means clustering alone. Hence the K-means clustering algorithm has to run a number of times with different random initial seed points at every K value which is time-consuming.

Cluster Misclassification Matrix (CMM) is the important and crucial measurement for any clustering algorithm. The CMM indicates how many data points are wrongly placed in clusters. The present paper tabulates CMM for MMC, WIC, K-means and WIKC algorithms as shown in Tables (5) to (8). It is evident that the proposed algorithms WIC and WIKC show better performance than MMC and K-means algorithms. Based on these tables the clustering accuracy percentages are evaluated and shown in terms of bar graphs in Figure 1.

TABLE 5. CMM OF MMC ALGORITHM.

	CUSTER 1	CLUSTER 2	CLUSTER 3
CLASS 1	-	-	-
CLASS 2	15	-	-
CLASS 3	13	5	-

TABLE 6. CMM OF WIC ALGORITHM.

	CUSTER 1	CLUSTER 2	CLUSTER 3
CLASS 1	-	-	-
CLASS 2	11	-	-
CLASS 3	6	3	-

TABLE 7. CMM OF K-MEANS ALGORITHM.

	CUSTER 1	CLUSTER 2	CLUSTER 3
CLASS 1	-	-	-
CLASS 2	-	1	-
CLASS 3	-	-	1

TABLE 8. CMM OF WIKC ALGORITHM.

	CUSTER 1	CLUSTER 2	CLUSTER3
CLASS 1	-	-	-
CLASS 2	-	-	-
CLASS 3	-	1	-

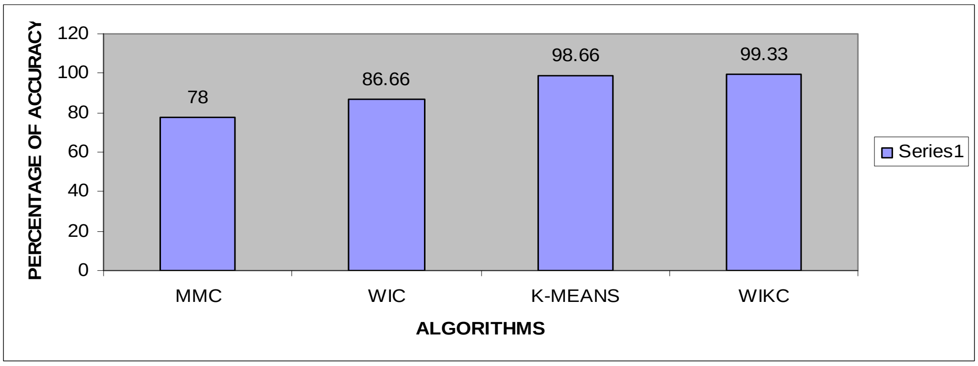

Figure 1. Comparison of Accuracy of Algorithms.

V. CONCLUSION

The convergence of any partitional clustering algorithm depends on number and selection of initial seed points. To over come these problems the present paper proposes two approaches called WIC and WIKC.

Tables 1, 2 and 3 show that the proposed WIC algorithm outperforms the existing Mountain Means Clustering (MMC) algorithm which requires three input parameters. The WIC algorithm improves cluster quality compared to MMC algorithm. Further the proposed WIC reduces the complexity when compared to existing MMC and its variants by eliminating construction and destruction of mountains in the two dimensional grid space. Tables 2 to 4 show the unique phenomenon, that WIKC shows a better compactness and separation when compared with all other algorithms. This is due to the result of minimum values in *IntraCD*, high values in *InterCD* and less value in cluster density. Moreover, the WIKC performs a less number of iterations than the other algorithms due to the initial selection of seed points. After WIKC approach, the WIC algorithm shows better results.These algorithms can be implemented to image data bases for segmentation of an image.

REFERENCES

[1] Fahim.A.M., Salem.A.M., et al."An Efficient enchanced k-means clustering algorithm," Journal of Zhejiang University,10(7): 1626-1633,2006.

[2] Fahim.A.M., Salem.A.M., Torkey.A.,et al."An Efficient k-means with good initial starting points," Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19,pp. 47-57,2009.

[3] Frigui.H and Krishnapuram.R. "A Robust Competitive Clustering Algorithm with Applications in Computer Vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, pp. 340-355, 2001.

[4] Jain.A.K., Duin.P.W and Mao.J"Statistical Pattern Recognition: A Review," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, no. 1, pp. 4-37, Jan. 2000.

[5] Jung W Lee, Seo Son.H. and Soon.H."Advanced Mountain Clustering Method," Kwon 0-7803-7078-3/01/$l0.0(C) IEEE 200l.

[6] Junnian Wang., Jianxun Liu. and Lanxia Liu. "A mountain means clustering algorithm," Proceedings of the 7th World Congress on Intelligent Control and Automation June 25 - 27, 2008, Chongqing, China, IEEE 2008.

[7] Leung.Y, Zhang.J and Xu.Z."Clustering by Space-Space Filtering," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, no.12, pp. 1396-1410, 2000.

[8] Madhu Yedla. et al. "Enhancing K-means clustering algorithm with improved initial centers," International Journal of Computer Science and information Technologies. Vol.1(2), 2010.

[9] Nazeer.K. A., Abdul and Sebastian.M.P."Improving the accuracy and efficiency of the k-means clustering algorithm," Proceedings of the World Congress on Engineering, Vol. 1, pp. 308-312, 2009.

[10] Rajashree Dash, Debahuti Mishra,Amiya Kumar Rath, Milu Acharya. "A hybridized k-means clustering approach for high dimensional dataset," Int. Journal of Engineering, Science and Tech., Vol. 2, no 2,pp. 59-66, 2010.

[11] Ursem.R.K."Multinational evolutionary algorithms," in: Proceedings of Congress of Evolutionary Computation, Vol. 3, Washington.

[12] Yager.R.R. and Filev.D.P."Approximate clustering via the mountain method," IEEE Transactions on Systems, Man, and Cybernetics, 24(8):1279-1284, 1994.

AUTHORS PROFILE

Prof. G.V.S.N.R.V.Prasad did his MS Software Systems, BITS Pilani and M.Tech in Computer Science and Technology in Andhra University .He has 25 years of teaching experience. Published 7 Research Papers in various National and International Conferences and 3 Research papers in National and International Journals. He is a member in various Professional Bodies . Presently working as Professor in CSE at Gudlavalleru Engineering College , Gudlavalleru ,A.P. His area of interest is Data Mining, Network Security and Image Processing.

Prof..V Venkata Krishna received the B.Tech. (ECE) degree from Sri Venkateswara University. He completed his M. Tech. (Computer Science)from JNT University. He received his Ph.D in Computer Science from JNTUniversity in 2004. He worked as Professor and Head for ten years in Mahatma Gandhi Institute of Technology, Hyderabad. After that, he worked as a principal for Vidya Vikas College of Engineering, Hyderabad, for two years, and he worked as Principal for Chaitanya Institute of Science and Technology, Kakinada from past one year. In addition, presently he is working as Principal for Chaitanya Institute of Science and Technology, Kakinada from past one year. He is an advisory member for many Engineering colleges. He has published 20 research articles. Presently he is guiding 10 research scholars. He is a life member of ISTE and CSI.

Prof..Vijaya Kumar did his MS Engineering in Computer Science [USSR –TASHKENT STATE UNIVERSITY] and Ph.D in Computer Science . Worked as Associate Professor in Department of CSE and School of Information Technology (SIT) at Jawaharlal Nehru Technological university (JNTU)Hyderabad . Having a total of 13 years of experience. He Published 60 Research Papers in various National and International Conferences /Journals. Guiding 10 Research scholars . He is a Member for various National and Inter National Professional Bodies .Presently working as Dean for CSE & IT at GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY Rajamundry