Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

An Approach for Selecting Optimal Initial Centroids to Enhance the Performance of K-means

Md. Mostafizer Rahman[1]
Department of Computer Science & Engineering
Dhaka University of Engineering & Technology
Gazipur, Bangladesh

Md. Sohrab Mahmud[2]
Department of Computer Science & Engineering
Dhaka University of Engineering & Technology
Gazipur, Bangladesh

Md.Nasim Akhtar[3]
Department of Computer Science & Engineering
Dhaka University of Engineering & Technology
Gazipur, Bangladesh

Abstract— **Clustering is the process of grouping data into a set of disjoint classes called cluster. It is an effective technique used to classify collection of data into groups of related objects. K-means clustering algorithm is one of the most widely used clustering techniques. The main puzzle of K-means is initialization of centroids. Clustering performance of the K-means totally depends upon the correctness of the initial centroids. In general, K-means randomly selects initial centroids which often show in poor clustering results. This paper has proposed a new approach to optimizing the designation of initial centroids for K-means clustering. We propose a new approach for selecting initial centroids of K-means based on the weighted score of the dataset. According to our experimental results the new approach of K-means clustering algorithm reduces the total number of iterations, improve the time complexity and also it has the higher accuracy than the standard k-means clustering algorithm.**

Keywords- Clustering, K-means algorithm,Weighted Score, Data analysis, Initial centroids, Improved K-means.

I. INTRODUCTION

Cluster analysis is one of the major data analysis tools to identify the behavior of data in a set of data items. The main purpose of this technique is to group data with maximum similarities into same clusters and separate data with dissimilarities into different clusters [1] [11]. It is a process of a set of data objects into disjoint clusters. Clustering is an example of unsupervised classification. It is similar to classification as it groups data but unlike classification groups are not predefined [2].

Clustering has been used in many application domains, including biology, medicine, anthropology, sensor networks, marketing and economics [2]. Clustering applications include plant and animal classification, disease classification, image processing, pattern processing and document retrieval. Medical taxonomy is one of the first domains in which clustering was used. In recent times classifying web log data to detect usage patterns is another important application [2].
In literature, a number of clustering methods are introduced. K-means is one of the major clustering techniques that is widely used and most popular in this category. The K-means algorithm is an effective one in computing clusters for huge practical applications in current researches such as bioinformatics, biomedical data analysis, pattern recognition etc [8,9]. Due to high computational complexity of the basic K-means algorithm, especially for large dataset, it is not time efficient and does not scale well. Moreover, it results in different types of clusters depending on the random choice of initial centroids. Researchers made a number of attempts to improve the performance of k-means algorithm. In this paper we propose a new method that improve the time complexity, cluster accuracy of k-means as well as improve the efficiency of K-means Clustering algorithm.

II. RELATED WORKS

Although K-means is a very simple and widely used clustering algorithm for variety of data types. Performance of K-means clustering algorithm strongly depends upon the selection of initial centroids. Therefore, it is quite important for K-means clustering to select initial centroids. In this paper, some proposals are reviewed.

Likas et al. [3] proposed the global k-means clustering algorithm that constructs initial centers by recursively partitioning data space into disjoints subspaces using a k-d trees method. The cutting hyper plane used in the method is defined as the plane that is perpendicular to the highest variance axis derived by principal component analysis. The partitioning is performed until each of the leaf nodes (bucket) contains less than a predefined number of data instances (bucket size) or the predefined number of buckets has been created. The centroids of data in the final buckets are then used as initial centers for K-means.

S.S. Khan and A. Ahmad [4] proposed cluster center initialization algorithm (CCIA) based on considering values for each attribute of the given data set. This can provide some information leading to a good initial cluster center.

Fang Yuan et al. [5] proposed a systematic method for finding the initial centroids. The centroids obtained by this method are consistent with the distribution of data. Hence it produced

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

clusters with better accuracy, compared to the original k-means algorithm. However, Yuan's method does not suggest any improvement to the time complexity of the k-means algorithm.

Xu et al. [6] specify a novel initialization scheme to select initial cluster centers based on reverse nearest neighbor search. Nazeer et al. [7] proposed an enhanced K-means algorithm, which combines a systematic method for finding initial centroids and an efficient way for assigning data points to cluster. This method ensures the entire process of clustering in $O(n^2)$ time without sacrificing the accuracy of clusters.

III. OVERVIEW OF STANDARD K-MEANS CLUSTERING ALGORITHM

In this section, we briefly describe the K-means algorithm. The basic idea of K-means algorithm is to classify the dataset D into k different clusters where D is the dataset of n data; k is the number of desired clusters. The algorithm runs in two basic phases[10]. The first phase is to select the initial centroids for each cluster arbitrary. In the second and final phase calculate distance of each data point with every centroid and assign data points to a cluster with nearest distance with centroids [10]. To measure the distance between data points and centroids Euclidean Distance method is used. When a new point is assigned to a cluster the cluster mean is immediately updated by calculating the average of all the points in that cluster [2]. The process of assigning a data point to a cluster and updating cluster centroids continues until the convergence criteria is met or the centroids don't differ between two consecutive iterations. Once a situation is met where centroids don't move anymore the algorithm ends. The Pseudo code for k-means clustering algorithm is given below [2].

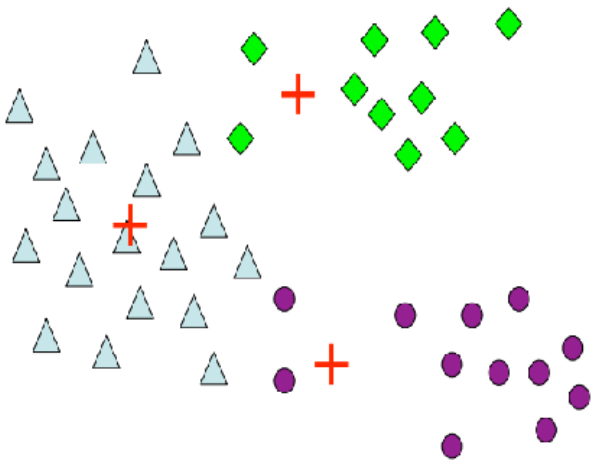

Figure 1: Random selection of initial centroids in K-means

A. *Standard K-means Algorithm*

Input:
D = {t_1,t_2,t_3,……,t_n} // set of elements
k // number of desired clusters;
Output:
K // set of clusters
Outer:
K-means algorithm:

Assign initial values for means $\mathbf{m_1}$,$\mathbf{m_2}$,….,$\mathbf{m_n}$;
Repeat
Assign each item $\mathbf{t_i}$ to the cluster which
has the closest Euclidean distance with mean;

Calculate new mean for each cluster.
Until
Convergence criterion is met;

IV. PROPOSED K-MEANS ALGORITHM

This paper proposed a novel approach to find the optimal initial centroids.To find initial centroids we calculate Weighted Score (WS) of each data point. Therefore, in a data point each attribute divided by its maximum value and finally sum all these attributes. Result of these Sums of attributes is called Weighted Score (WS) of a data point.

A. *Weighted Score (WS) Calculation:*

Attributes $=x_1, x_2, x_3,\ldots\ldots\ldots\ldots\ldots,x_m$
Data Points $=T_1, T_2, T_3,\ldots\ldots\ldots\ldots\ldots\ldots..,T_n$
Weighted Score (WS) $T_n = \sum_{j=1}^{m} \frac{x_j}{x_{j(max)}}$

$$= \frac{x_1}{x_{1(max)}} + \frac{x_2}{x_{2(max)}} + \frac{x_3}{x_{3(max)}} + \frac{x_4}{x_{4(max)}} + \ldots\ldots + \frac{x_j}{x_{j(max)}}$$

If we consider the max value of attributes $x_1(max)=1$, $x_2(max)=2$, $x_3(max)=1.5$, $x_4(max)=1.6$, $x_5(max)=1.2$, $x_6(max)=1.1$, $x_7(max)=1.0$

Hence, Weighted Socre (WS) of
datapoint T_1 =2.07
datapoint T_2 =2.00

TABLE 1: Calculation of Weighted Score (WS)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Weighted Score (WS)
T_1	**0.1**	**0.4**	**0.3**	**0.1**	**0.3**	**0.5**	**0.8**	**2.07**
T_2	**0.2**	**0.3**	**0.2**	**0.3**	**0.4**	**0.1**	**0.9**	**2.00**

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

A sorting algorithm is applied to sort the data points based on Weighted Score (WS) and sorted data divided into k subsets where k is the number of desired clusters. Calculate mean of each subset and finally choose an initial centroids whose Weighted Score (WS) is closest to the mean value.

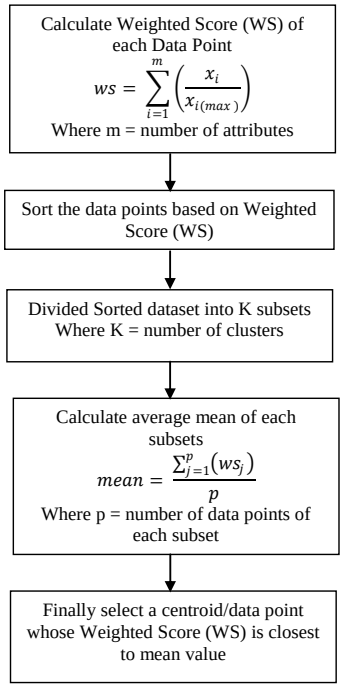

Figure 2: Flow Chart of proposed method.

For example a data set D is consists of n data points such as $T_1,T_2,T_3,\ldots\ldots,T_n$. Each data point of this set may contain multiple attributes such as T_n may contain attributes $x_1,x_2,x_3,.....,x_m$, where m is the number of attributes.

B. Pseudo code of Proposed Method

Input:
D = {t_1, t_2, t_3,……,t_n } // set of elements
k = number of desired clusters;
Output:
Set initial centroids K.
Steps:
1.Calculate the Weighted Score(WS) of each data point;

- $T_n=x_1,x_2,x_3,..,x_m$
- Weighted Score(WS) of $T_n = \sum_{j=1}^{m} \frac{x_j}{x_{j(max)}}$

// where x = the attributes value, m = number of attributes, $x_{j(max)}$=Maximum value of attributes x_j,T= Data points.
2. Sort the data points based on Weighted Score (WS);
3. Divide the datasets into k subsets;
4. Calculate the mean value of the each subset;
5. Select an initial centroids whose Weighted Score (WS) is closest to the mean value of subsets;

The method described above to find initial centroids of the clusters is more significant than the standard k-means where centroids are selected randomly. The algorithm meets the convergence criteria faster than the standard k-means.

V. COMPLEXITY ANALYSIS

In basic K-means algorithm, the initial centroids are randomly calculated. For that, the cluster centroids are tuned many times before the convergence criterion of the algorithm is met and the data points are assigned to their nearest centroids. Since, complete reassignment of data points takes place according the new centroids, this method takes time O(nkl) where n is the number of data points, k is the number of clusters and l is the number of iterations. The proposed algorithm discussed in this paper works in two phases. In the first phase of the algorithm the time required to calculate the Weighted Score (WS) of all the data points is O(n) where n is the number of data points. The algorithm then proposes to sort the data in ascending order. Sorting the data points based on the Weighted Score (WS) of each data point can be done in $O(n\log_n)$ time using Merge Sort. Finally in the first phase of the proposed, the overall time complexity is $O(n\log_n)$.

The second phase of the proposed algorithm follows that of the original k-means algorithm. Distribution of the data points to the nearest cluster and the consequent tuning of centroids are conducted repeatedly until the convergence criteria reached. This process concluded with a time complexity of O(nkl) where the symbols represent the meaning mentioned above. The experimental data shows that the algorithm converges in less number of iterations as the initial centroids are calculated in a strategic way rather than randomly. Thus the overall complexity of the proposed algorithm is of $O(n(kl + \log_n))$.

VI. EXPERIMENTAL RESULTS

The multivariate data sets, taken from the UCI repository of machine learning databases that are used to test the accuracy and efficiency of the modified k-means algorithm. The same data sets are given as input to the standard k-means algorithm and the modified k-means algorithm. The value of k, the number of clusters, is taken as 3. We have applied our proposed algorithm IV (A), IV(B) on several datasets and compared our results with standard k-means algorithm in terms of the accuracy of cluster and total execution time.

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

TABLE 2 : PIMA Indians Diabetes Data Set

Number of times pregnant	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable (0 or 1)
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

TABLE 3. Categorical Partitioning of PIMA diabetic data set.

Attributes name	Partitioned data with max-min Weight of Attributes
Number of times of pregnancy (# Preg)	{low, medium, high} { <3, 3-5 ,>6}
Plasma glucose concentration every 2 hours in an oral glucose tolerance test -- (Plasma)	{low, medium, high } {< 95 , 95-150 , >150}
diastolic blood pressure (mm Hg) --- (Diast BP)	{low, normal, high } { <70, 70-100, >100}
Triceps skin fold thickness (mm) --(skin)	{low, medium, high } { <21, 21-40, > 40}
2-Hour serum insulin (mu U/ml) --- (insulin)	{normal, medium, high} {<140, 140-200, > 200}
body mass index (weight in kg/(height in (mm)2) ----BM	{normal,obese,overweight} {<23 , 23-29, > 29}
diabetes pedigree function --- Pedigree	{low, medium, high } { <0.4, 0.4- 0.8, > 0.8}
Age in years ---- Age	{young, middle aged, senior}
Class	{Tested Positive, Tested Negative}

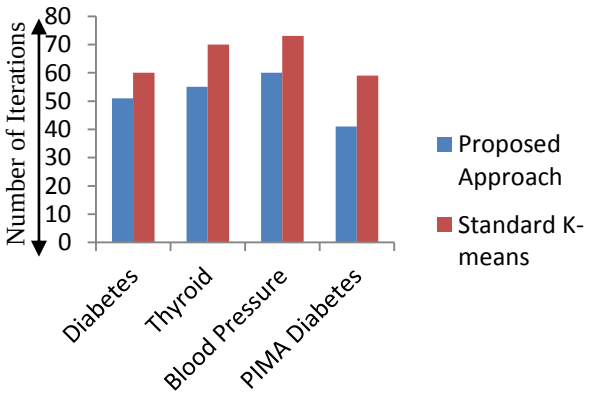

Figure 3: Iteration comparison between Proposed and Standard k-means Clustering Algorithm

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

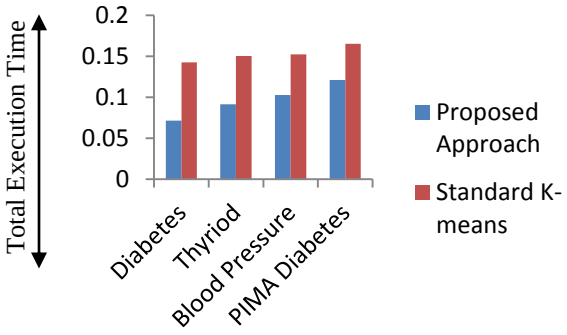

Figure 4: Time comparison between Proposed and Standard k-means Clustering Algorithm

The experimental results are shown in Table 4.

TABLE 4 : Performance analysis of proposed Algorithm

Data Sets	Number of Clusters	Total number of Execution	**Algorithm**	
			K-Means	Proposed Algorithm
			Avg. Time Taken (ms)	Avg. Time Taken (ms)
Diabetes	3	20	0.1423	0.0712
Thyroid	3	25	0.1493	0.0981
Blood Pressure	3	30	0.1523	0.1024
PIMA Diabetes	3	35	0.1713	0.1203

In standard k-means algorithm centroids are taken randomly but in proposed algorithm the dataset and the value of k are the only inputs needed since the initial centroids are computed automatically and find optimal cantroids by the program. In experiments proposed k-means algorithm shows better performance than standard k-means algorithm.

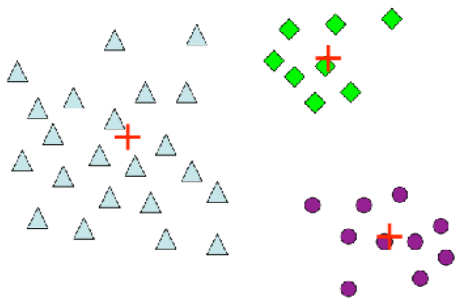

Figure 5: Optimal selection of initial centroids in proposed method

VII. CONCLUSIONS AND FUTURE DIRECTIONS

K-means algorithm is a common and widely used technique for clustering. In recent, due to incredible growth of multi-dimensional data, conventional k-means technique is inadequate to efficiently classify the distribution of data. Conventional k-means algorithm doesn't uses any technique to select initial centroids whereas accuracy of a cluster mostly depends on selection of initial centroids. So researchers nowadays are emphasized to develop new techniques to meet the raised requirements.

In this paper we proposed a new technique to select initial centroids that increase the cluster accuracy as well as decrease the time complexity also reduce the iteration time. Automating the value of k is suggested as a future work.

ACKNOWLEDGMENT

This research work is funded by Dhaka University of Engineering & Technology (DUET),Gazipur,Bangladesh

REFERENCES

[1] Kathiresan V., Dr. O Sumanthi, "An Efficient Clustering Algorithm based on Z-Score Ranking method", International Conference on Communication and Informatics (ICCCI-2012), Jan. 10-12, 2012, Coimbatore, INDIA.

[2] Margaret H. Dunham, S. Sridhar, "Data Mining, Introductory and Advanced Topics", Pearson Education, ISBN: 0130888923.

[3] Likas, N. Vlassis and J.J. Verbeek, "The Global k-means Clustering algorithm", Pattern Recognition , Volume 36, Issue 2, 2003, pp. 451-461.

[4] S. S. Khan and A. Ahmad,"Cluster Center Initialization for K-mean Clustering", Pattern Recognition Letters, Volume 25, Issue 11, 2004, pp. 1293-1302.

[5] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[6] Xu Junling, Xu Baowen, Zhang Weifeng, Zhang Wei and Hou Jun, 2009. Stable initialization scheme for K-means clustering, Wuhan University Journal of National Sciences, Vol. 14, No. 1, pp. 24-28.

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

[8] Amir Ben-Dor, Ron Shamir and Zohar Yakini, "Clustering Gene Expression Patterns," Journal of Computational Biology, 6(3/4): 281-297, 1999

[9] Daxin Jiang, Chum Tong and Aidong Zhang, "Cluster Analysis for Gene Expression Data," IEEE ransactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

[10] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Diego, 2001.

Fang Yuan, Zeng-Hui Meng, Hong-Xia Zhangz and Chun-Ru Dong ,"A NEW ALGORITHM TO GET THE INITIAL CENTROIDS" , Proceedings of the Third International Conference on Machine Laming and Cybernetics, Shanghai, 26-29 August 2004