Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.191 – 203

Implementing & Improvisation of K-means Clustering Algorithm

Unnati R. Raval[1], Chaita Jani[2]
[1]Department of Computer Science & KIRC (GTU), India
[2]Department of Computer Science & KIRC (GTU), India
[1] Unnatiraval31@gmail.com, [2] Jani.chaita@gmail.com

Abstract

The clustering techniques are the most important part of the data analysis and k-means is the oldest and popular clustering technique used. The paper discusses the traditional K-means algorithm with advantages and disadvantages of it. It also includes researched on enhanced k-means proposed by various authors and it also includes the techniques to improve traditional K-means for better accuracy and efficiency. There are two area of concern for improving K-means; 1) is to select initial centroids and 2) by assigning data points to nearest cluster by using equations for calculating mean and distance between two data points. The time complexity of the proposed K-means technique will be lesser that then the traditional one with increase in accuracy and efficiency.

The main purpose of the article is to proposed techniques to enhance the techniques for deriving initial centroids and the assigning of the data points to its nearest clusters. The clustering technique proposed in this paper is enhancing the accuracy and time complexity but it still needs some further improvements and in future it is also viable to include efficient techniques for selecting value for initial clusters(k). Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the k-means.

Introduction

Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, but data belonging to different cluster differ. A cluster is a collection of data object that are similar to one another are in same cluster and dissimilar to the objects are in other clusters [k-4]. At present the applications of computer technology in increasing rapidly which

created high volume and high dimensional data sets [10]. These data is stored digitally in electronic media, thus providing potential for the development of automatic data analysis, classification and data retrieval [10]. The clustering is important part of the data analysis which partitioned given dataset in to subset of similar data points in each subset and dissimilar to data from other clusters [1]. The clustering analysis is very useful with increasing in digital data to draw meaningful information or drawing interesting patters from the data sets hence it finds applications in many fields like bioinformatics, pattern recognition, image processing, data mining, marketing and economics etc [4].

There have been many clustering techniques proposed but K-means is one of the oldest and most popular clustering techniques. In this method the number of cluster (k) is predefined prior to analysis and then the selection of the initial centroids will be made randomly and it followed by iterative process of assigning each data point to its nearest centroid. This process will keep repeating until convergence criteria met. However, there are shortcomings of K-means, it is important to proposed techniques that enhance the final result of analysis. This article includes researched on papers [1,2,3,4,5,6] which made some very important improvements towards the accuracy and efficiency of the clustering technique.

Basic K-means Algorithm : A centroid-based Clustering technique

According to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters[k-1]. This partitioning clustering is most popular and fundamental technique [1]. It is vastly used clustering technique which requires user specified parameters like number of clusters k, cluster initialisation and cluster metric [2]. First it needs to define initial clusters which makes subsets (or groups) of nearest points (from centroid) inside the data set and these subsets (or groups) called clusters [1]. Secondly, it finds means value for each cluster and define new centroid to allocate data points to this new centroid and this iterative process will goes on until centroid [3] does not changes. The simplest algorithm for the traditional K-means [2] is as follows;

Input: D = {d1, d2, d3,.....dn} // set of n numbers of data points

K // The number of desire Clusters

Output: A set of k clusters

1. *Select k points as initial centroids.*
2. *Repeat*
3. *From K clusters by assigning each data point to its nearest centroid.*
4. *Recompute the centroid for each cluster until centroid does not change* [2].

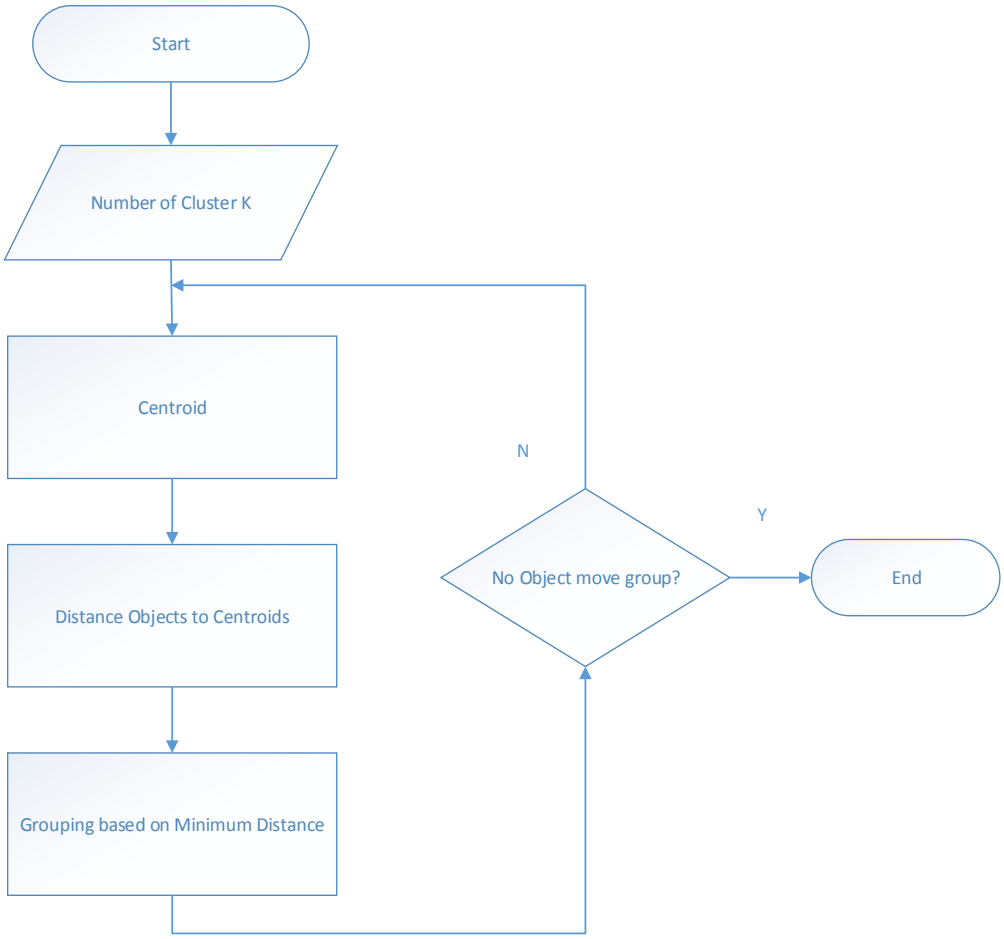

Fig: Basic Flowchart of K-means algorithm

However, the algorithm has its own pros and cons, which is as follows;

PROs:

1. It is relatively faster clustering technique [1].
2. It works fast with the Large data set since the time complexity is O(nkl) where n is the number of patterns, k is the number of clusters and l is the number of the iterations.
3. It relies on Euclidian distance which makes it works well with numeric values with interesting geometrical and statistic meaning [4].

CONs:

1. The initial assumption of value for K is very important but there isn't any proper description on assuming value for K and hence for different values for K will generate the different numbers of clusters [4].
2. The initial cluster centroids are very important but if the centroid is far from the cluster center of the data itself then it results into infinite iterations which sometimes lead to incorrect clustering [4].
3. The K-means clustering is not good enough with clustering data set with noise [4].

Improved K-means Clustering Algorithm

Proposed Methodology

On the based on survey that have been carried –out on some proven enhanced K-means algorithms, there have been some areas which could be improved to get better accuracy and efficiency from altering traditional K-means. These areas have been discussed in this section with reference to proven theorems and methods.
As per research it is clear that we needs to make better assumption or find method for determining initial centroids with assigning data points to closed centroid clusters after iteration to enhance the results of traditional K-means.

Important Equations

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters.

1. The Euclidean distance is given by:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Improved Technique

Part1: Determine initial centroids

Step1.1: Input Dataset

Step1.2: Check the Each attributes of the Records

Step1.3: Find the mean value for the given Dataset.

Step1.4: Find the distance for each data point from mean value using Equation (Equ).

IF

The Distance between the mean value is minimum then it will be stored in

Then Divide datasets into k cluster points don't needs to move to other clusters.

ESLE

Recalculate distance for each data point from mean value using Equation (Equ) until divide datasets into k cluster

Part2: Assigning data points to nearest centroids

Step2.1: Calculate Distance from each data point to centroids and assign data points to its nearest centroid to form clusters and stored values for each data.

Step2.2: Calculate new centroids for these clusters.

Step2.3: Calculate distance from all centroids to each data point for all data points.

IF

The Distance stored previously is equal to or less then Distance stored in Step2.1

Then Those Data points don't needs to move to other clusters.

ESLE

From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids.

Step2.5: Calculate centroids for these new clusters again.

Until

The convergence criterion met.

OUTPUT

A Set of K clusters.

4.2 Flow of Proposed Methodology

The research has been done on clustering algorithms that improves accuracy of the results with better accuracy lesser time complexity. But there are still many departments that needs to be improved that enhance the accuracy and time complexity with the larger data sets. There are many ways to improve K-means for the larger data set hence the further research will be done to improve it.

- An improved K-means be based on the two phases;
  - ✓ Deriving Initial Centroids
  - ✓ Assigning data-points to nearest clusters

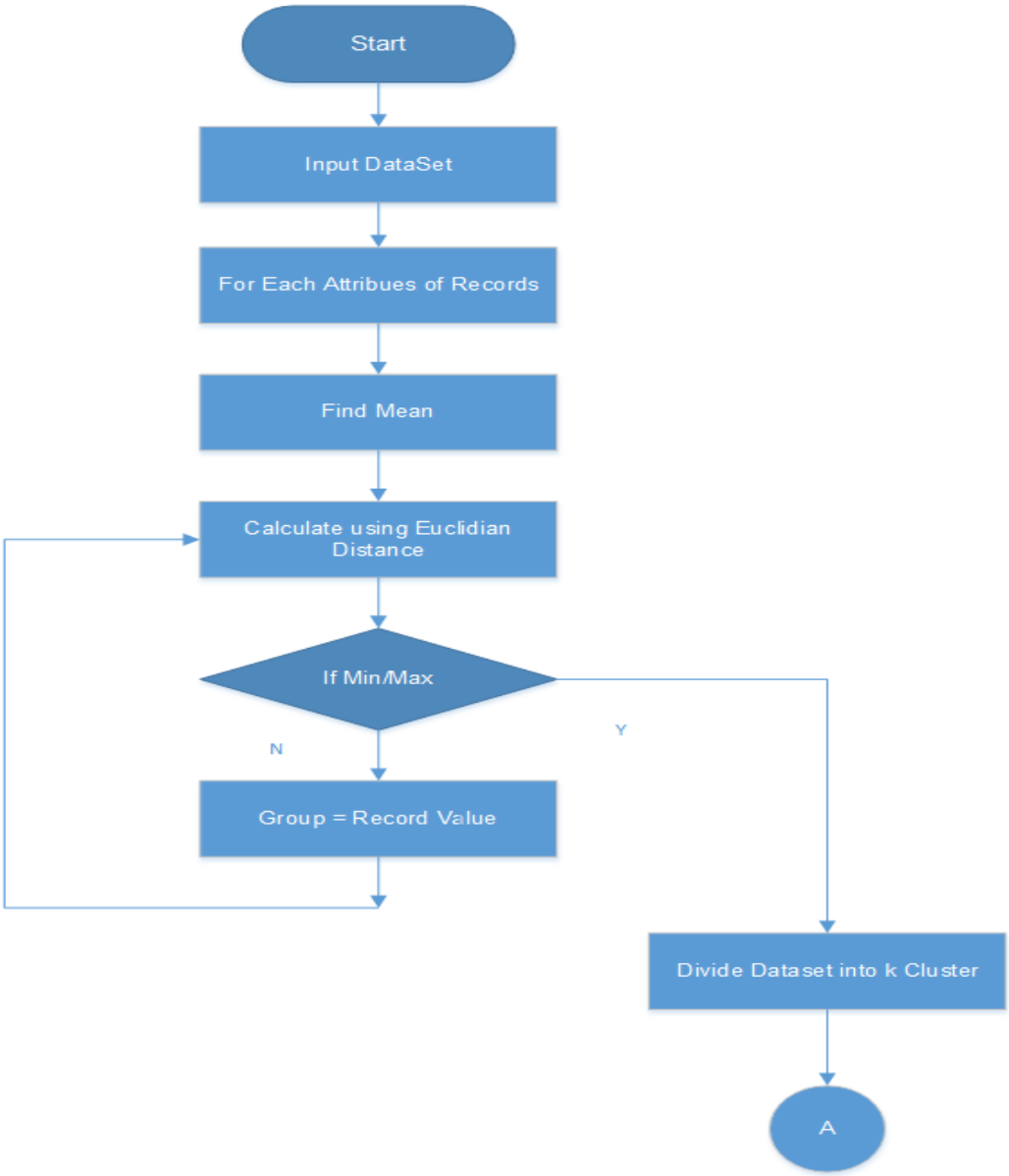

Fig: Flowchart of proposed k-means algorithm

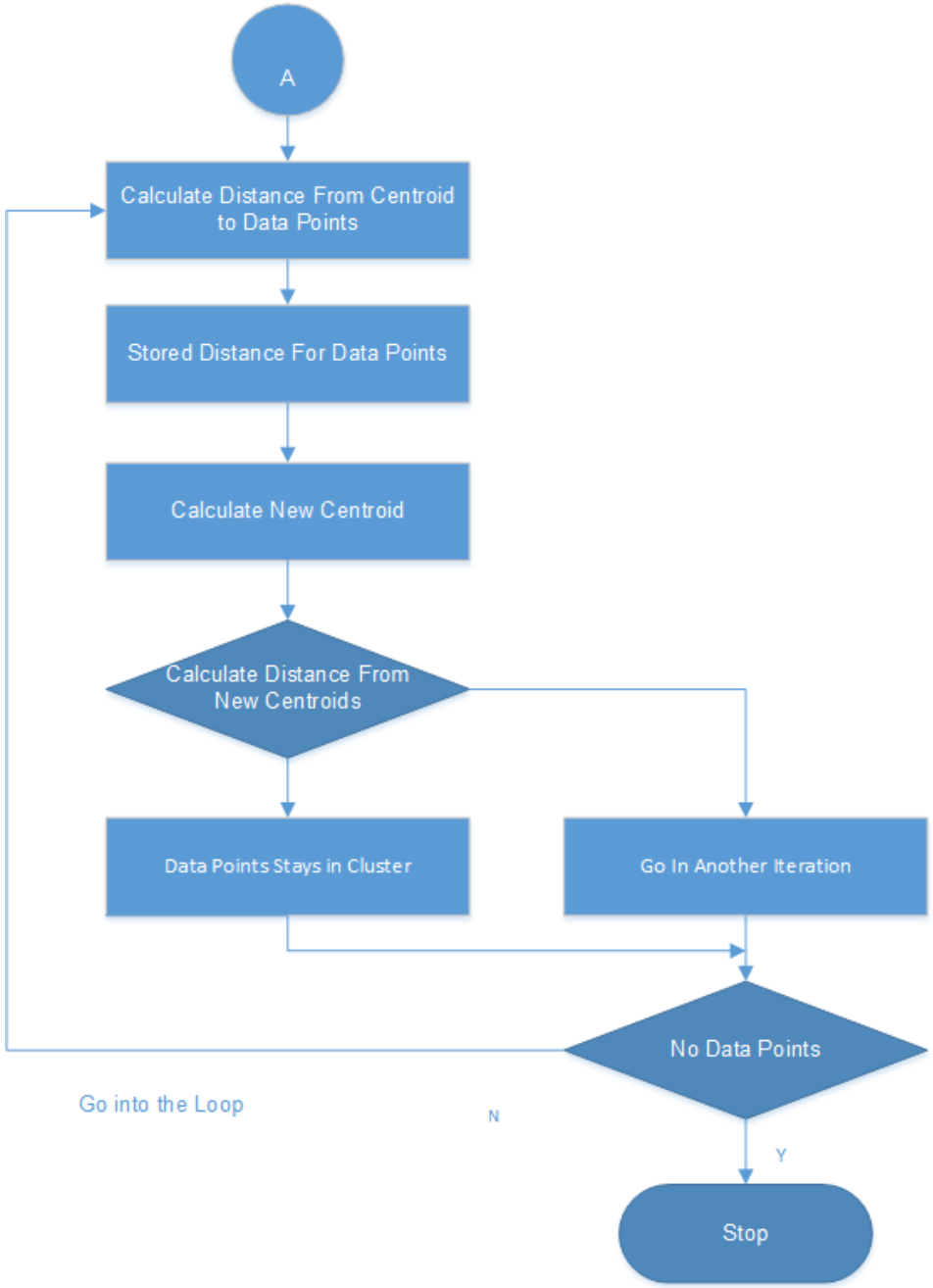

Fig: Flowchart of Proposed k means algorithm

The K-means is very old and most used clustering algorithm hence many experiments and techniques have been proposed to enhance the efficiency accuracy for clustering. Let's discuss some of the improved K-means clustering proposed by different authors.

The first paper [4], proposing algorithm by improving the methods of finding of initial centroids and assigning data points to appropriate clusters [5]. Initially, it starts with the checking for negative value attributes in given data set and if there is some, then it a transom all negative value attributes to positive by subtracting it for the minimum positive value of the data set itself [5]. Next, calculate the distance from center to each data point then the original data points are sorted in accordance with the sorted distance and partitioned it with K equal cluster to find better initial centers [5]. In next step, all data points will be assigned to new centroids by using heuristic approach to reduce

computational time. The time complexity of the proposed algorithm in this case will be O(nk) [5] instead of O(nkl) [1] for traditional K-means, is much faster.

In this proposal [6], instead of calculating distance from centroids to data points in all iterations it goes for, in traditional technique, it calculates the distance from new centroid once and if the distance is less than or equal to the previously calculated distance then it stays in cluster itself and no further calculation will be carry on for this particular data point otherwise it goes for same procedure again until all data points are assigned to its closest pre tagged centroids [6]. The time complexity of the proposed algorithm will be O(nk) [6] which will be faster than the traditional K-means.

The enhanced K-means [7] divides the algorithm in two phases and uses different algorithms to make proposed algorithm more efficient and accurate. In the first phase the initial centroids are determined systematically to produce the clusters with better accuracy and in second phase it uses various methods described in the algorithm to assigned data points to the appropriate clusters. There have been discussed algorithms for both phases in the paper. The time complexity of the algorithm, as claimed in paper, is O(n) where n is number of data point[7] with makes the algorithm much faster than the others.

In the other paper [4], initial centroids are optimised to find a set of data to reflect the characteristics of data distribution as the initial cluster centers and then optimising the calculation of distance between data points to the cluster centers and make it more match with the goal of clustering [4]. The test has been carried-out using IRIS and WINE data set and results are clearly showing improvements in accuracy for proposed techniques.

Finally, this technique proposed in [8], is base on the assumptions that the value of k (clusters) known in priori and then divide all data set into k clusters to calculate the new centroids. After that decide the membership of the patterns according to minimum distance from cluster center criterion and then do iterations using formula discussed in the paper [8]. Then the iteration will start with calculating distance from cluster center criterion and assigning of data points to cluster until no changes found in the cluster centers [8].

Experimental Result

We tested both the algorithms for the data sets with known clustering, Iris and Glass. And Also tested algorithm for the large Dataset With known as Bank, Credit and Audit. After this, the proposed algorithm will take no. of clusters as input. As the computation is complete, clustering result will shown in window. This clustering result also contain the total computation time.

The results of the experiments are tabulated in The Improved k-means algorithm requires the values of the initial centroids also as input, apart from the input data values and the value of *k*. For the enhanced algorithm, the data values and the value of k are the only inputs required since the initial centroids are computed automatically by the program. The percentage accuracy and the time taken in the case of this algorithm are also computed and tabulated. Here show the Experimental results for the large dataset.

Dataset	Traditional K-means Running Time(s)	Improved K-means Running Time(s)	Traditional K-Means Accuracy%	Improved K-means Accuracy%
Iris	0.0586	0.0457	84.3	90.6
Glass	0.0814	0.0778	78.3	84.9

Table 1: Comparison Table of K- Mean Clustering Algorithm and Our Proposed Algorithm

After these Experiments we got the results and it is shows that the An Improved K means Algorithm is a better then the traditional k- means for the large data sets. As below Table 2 show the tabular form for the results for the apply the proposed k means for the different Datasets like Bank, Credit and Audit. The Results show that the time complexity of the proposed K-means technique will be lesser that then the traditional one with increase in accuracy and efficiency.

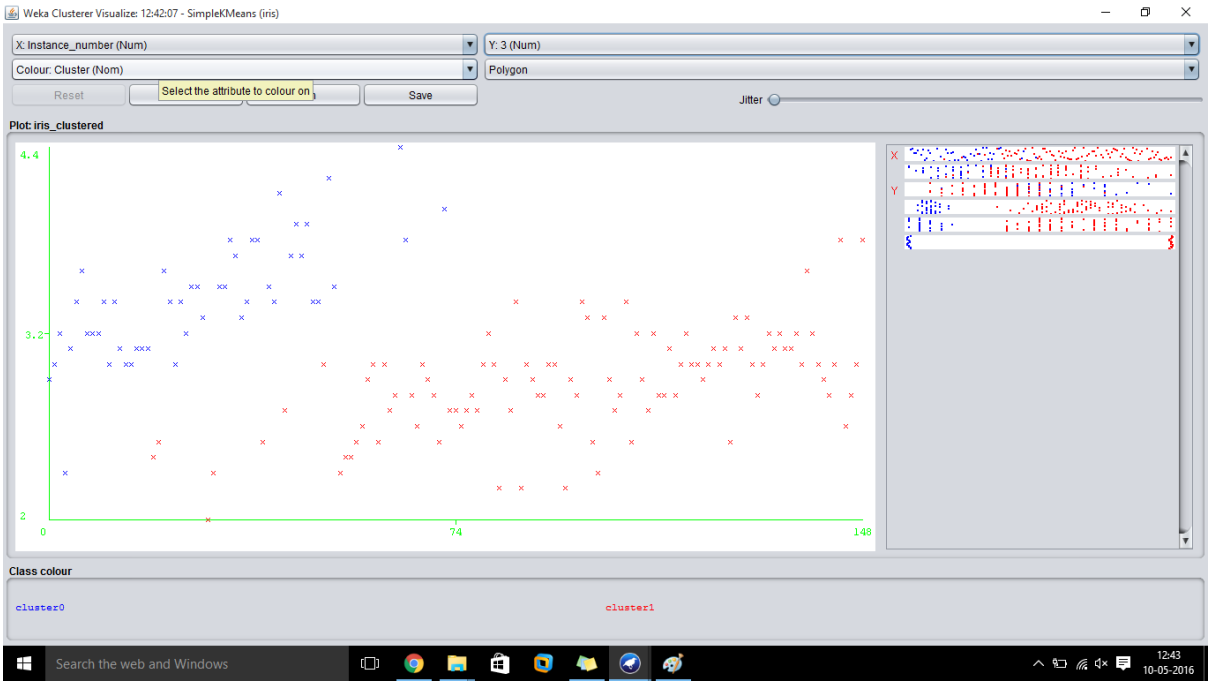

Fig: Performance chart of Simple k-means clustering on Iris Dataset

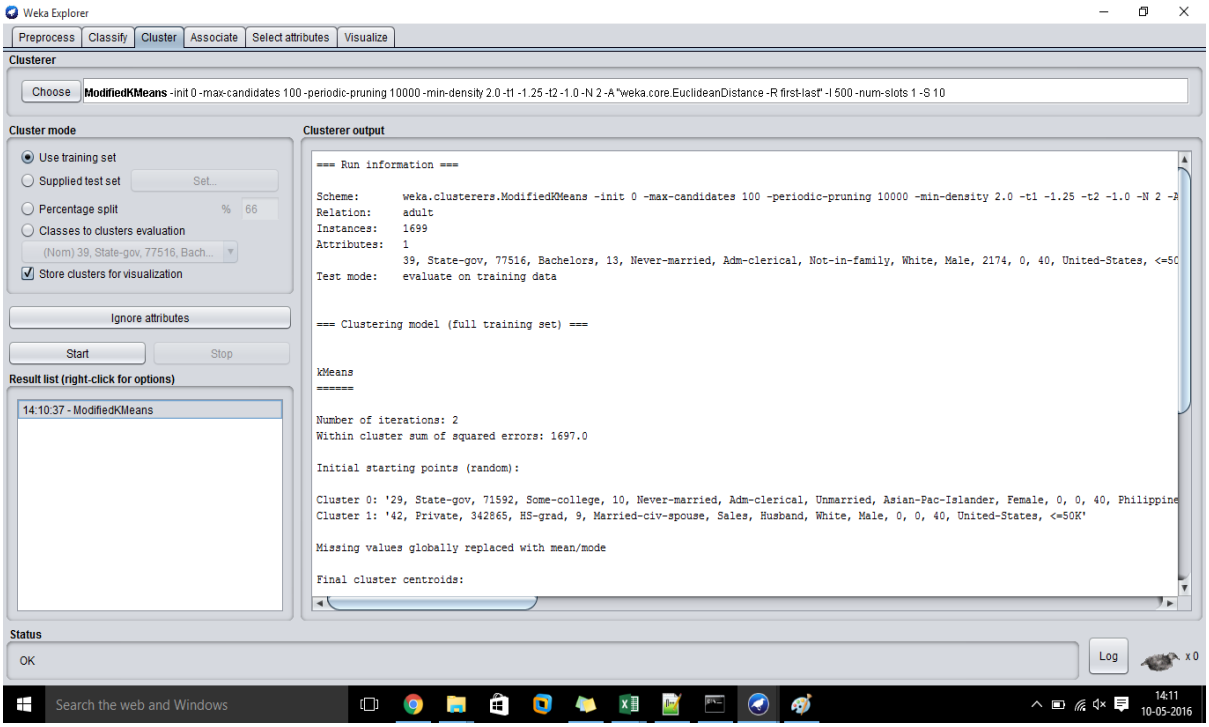

Fig: Modified k means clustering in weka tool

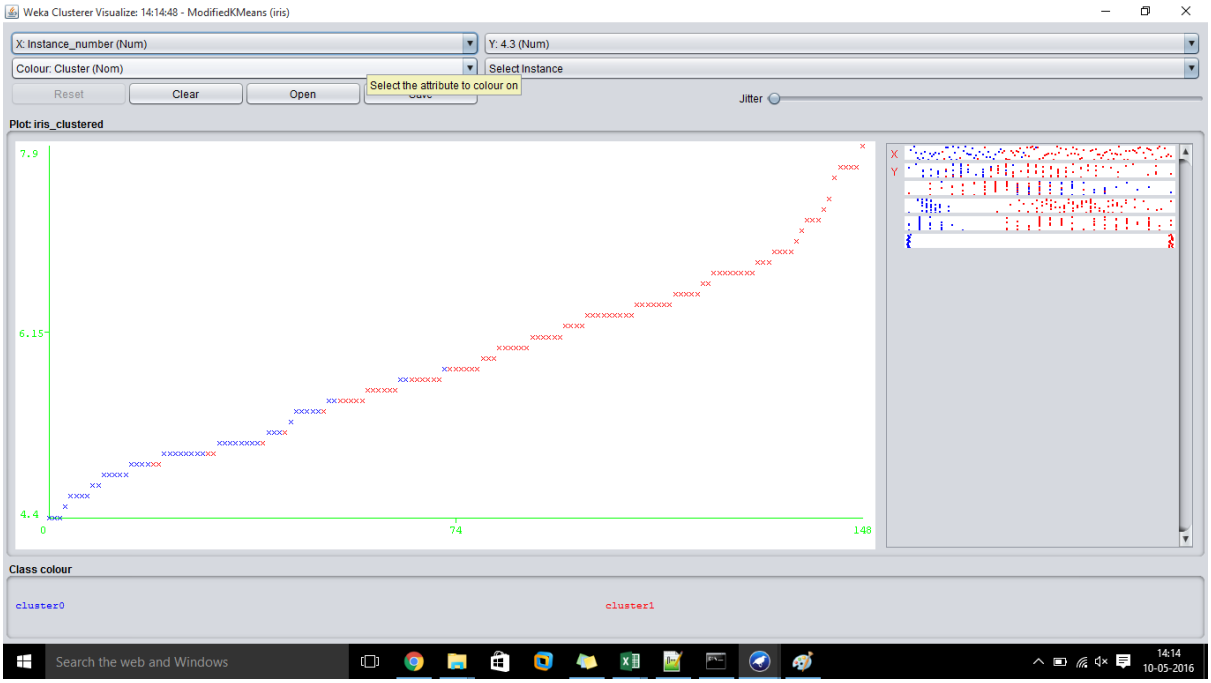

Fig: Performance chart of Improved K-means Clustering on iris dataset

Dataset	Traditional K-means Running Time(s)	Improved K-means Running Time(s)	Traditional K-Means Accuracy%	Improved K-means Accuracy%
Bank	10.9684	10.882	80.19	84.66
Credit	12.6680	11.7794	70.82	78.90
Adult	14.99	12.806	66.64	81.33

Table 2: Comparison Table of K- Mean Clustering Algorithm and Our Proposed Algorithm For the large Datasets

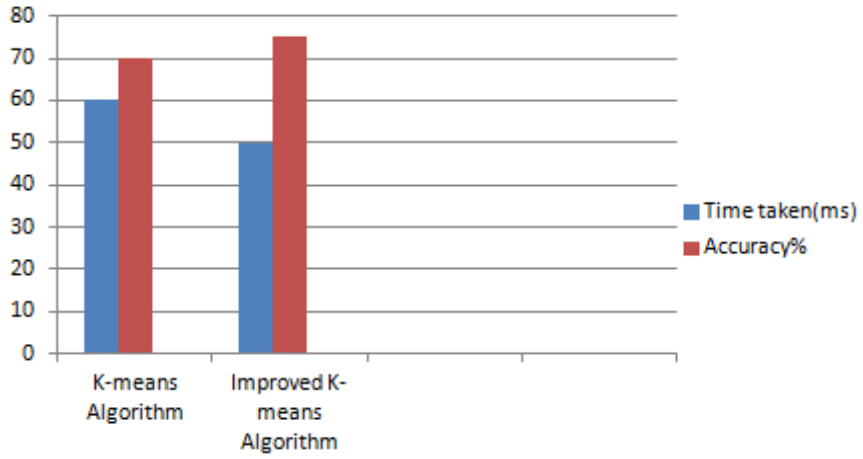

Fig.: Performance Comparison chart for Efficiency and Accuracy of the algorithms

Time Complexity

As the concept is drawn from the for deriving initial centroids the time it will take for first phase will be O(nlogn) where n is the number of data points . However these technique needs to go through lot of sorting hence the overall rime complexity becomes O(nlongn) in both and worst case. In the second phase of clustering, if the data point remains in the clusters itself then the time complexity becomes the O(1) and for others it else O(K)[6]. If half of the data points retains its clusters then time complexity will become O(nK/2) hence the total time complexity becomes O(nk). Hence the total time complexity for the Improved K-means clustering is O(n) which has less time complexity than the traditional k-means which runs with time complexity of $O(n^2)$.

Total time required by improved algorithm is O(n) while total time required by standard k-mean algorithm is $O(n^2)$. So the improved algorithm improves clustering speed and reduces the time complexity.

Conclusion and Future Work

The traditional K-means clustering is most used technique but it depends on selecting initial centroids and assigning of data points to nearest clusters. There are more advantages than disadvantages of the k-means clustering but it still need some improvements. This paper explains the techniques that improves the techniques for determining initial centroids and assigning data points to its nearest clusters with more accuracy with time complexity of O(n) which is faster than the traditional k-means. The initial value for the K (number of clusters) is still area of concern because it can improve accuracy of the clustering, which will be improved by enhancing the traditional way in future. However, researching on the improvement of K-means clustering algorithms are still not solved completely. And the further attempt and explore will be needed.

References

[1] H. Jiawei,M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, (San Francisco California, Morgan Kaufmann Publishers, 2012).

[2]P. Rai, and S. Sing, A survey of clustering techniques, *International Journal of computer Applications, 7(12)*, 2010, 1-5.

[3]T. Soni Madhulatha, An overview on clustering methods, *IOSR Journal of engineering, 2(4)*, 2012, 719-725.

[4] C. Zhang, and Z. Fang, An improved k-means clustering algorithm, *Journal of Information & Computational Science, 10(1)*, 2013, 193-199.

[5] M. Yedla, S.R. Pathakota, and T.M. Srinivasa, Enhancing K-means Clustering algorithm with Improved Initial Center, *International Journal of Computer Science and Information Technologies, 1 (2)*, 2010, 121-125.

[6] S. Na, G. Yong, and L. Xumin, Research on k-means clustering algorithms, *IEEE Computer society, 74*, 2010, 63-67.

[7] K.A.Abdul Nazeer, and M.P. Sebastian, Improving the accuracy and efficiency of the K-means clustering algorithm, *The World Congress on Engineering, 1*, 2009.

[8] M.A. Dalal, N.D. Harale, and U.L. Kulkarni, An iterative improved k-means clustering, *ACEEE International Journal on Network Security, 2(3)*, 2011, 45-48.

[9] D.T. Pham, S.S. Dimov, and C.D. Nguyen, Selection of K in K-means clustering, *IMechE 2005, 219*, 2004, 103-119.

[10]A.K. Jain, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters, Elsevier, 31*, 2010, 651-666.

[11] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity, Middle-East Journal of Scientific Research 12 (7): 959-963, 2012

[12] Malay K. Pakhira, A Modified *k*-means Algorithm to Avoid Empty Clusters, International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009

[13] Jyoti Yadav, Monika Sharma, A Review of K-mean Algorithm, International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013

[14] Dr. Aishwarya Batra, Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms

[15] 1G. Sathiya and P. Kavitha, An Efficient Enhanced K-Means Approach with Improved Initial Cluster Centers, Middle-East Journal of Scientific Research 20 (4): 485-491, 2014

[16] D.Napoleon, P.Ganga Lakshmi, An Enhanced k-means algorithm to improve the Efficiency Using Normal Distribution Data Points, International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010, 2409-2413

[17] Kahkashan Kouser, Sunita, A comparative study of K Means Algorithm by Different Distance Measures, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 9, November 2013

[18] Sudesh Kumar, Nancy, Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10

