A Comparative study on clustering of data using Improved K-means Algorithms

#1Abhilash C B
Assistant professor, Department of CSE
JSSATE, Bangalore

#2Sharana basavanagowda
Assistant professor, Department of CSE
JSSATE, Bangalore

Abstract — There exist many algorithms for clustering, and most widely used is K-means algorithm as it is easy to understand and simulate on different datasets. In our paper work we have used K-means algorithm for clustering of yeast dataset and iris datasets, in which clustering resulted in less accuracy with more number of iterations. We are simulating an improved version of K-means algorithm for clustering of these datasets, the Improved K-means algorithm use the technique of minimum spanning tree. An undirected graph is generated for all the input data points and then shortest distance is calculated which intern results in better accuracy and also with less number of iterations.

Both algorithms have been simulated using java programming language; the results obtained from both algorithms are been compared and analysed. Algorithms have been run for several times under different clustering groups and the analysis results showed that the Improved K-means algorithm has provided a better performance as compared to K-means algorithm; also Improved K-means algorithm showed that, as the number of cluster values increases the accuracy of the algorithm also increases. Also we have inferred from the results that at a particular value of K (cluster groups) the accuracy of Improved K-means algorithm is optimal.

Index Terms—***K-Means, MST, Improved K-Means, Yeast dataset, iris dataset.***

1. INTRODUCTION

1.1. Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, resulted in the introduction of bioinformatics, computational genomics and proteomics, large-scale analysis of complete genomes.

Bioinformatics is an interdisciplinary field involving biology, computer science, mathematics and statistics to analyse biological sequence data, genome content and arrangement, and to predict the function and structure of macromolecules. It can be viewed as the use of computational methods to make biological discoveries.

Gene expression is the process by which the information from gene is used in the synthesis of a functional gene product.

Clustering is the process of partitioning a set of objects (patterns) into a set of disjoined groups (clusters). Its goal is to reduce the amount of data by categorizing or grouping similar data items together and obtain useful information.

II. K-MEANS ALGORITHM

K-mean algorithm is modified in such a way that the efficiency and accuracy of the algorithm is increased. To achieve the same, the concept of Kruskal's algorithm that is a part of graph theory and uses the technique of minimum spanning tree (MST) and both the versions of algorithm have been simulated and analyzed in the scope of this master thesis paper.

Traditional K-means algorithm and the improved version of the algorithm are implemented in Java programming Language. Development of the improved version of the algorithm is to improve the algorithm in order to increase the accuracy in less number of iterations and results in reasonable time. The properties of algorithms which are to be modified are very important for improvement task. As improvement is a strategy for performing better accuracy and time consuming tasks faster, here minimum spanning tree (MST) is considered and it is obvious that the algorithm efficiency is increased and it can achieve the task in less number of iterations. Some algorithms are very suitable for improvement but some of them are not.

Properties of Improved K-means algorithm are also very important for the success of improvement. Minimum spanning tree is the shortest spanning tree in which length of a tree is equal to the sum of the length of the arcs on the tree. So this technique is used in the K-means algorithm which

leads to the improvement of the algorithm with better accuracy in less number of iteration.

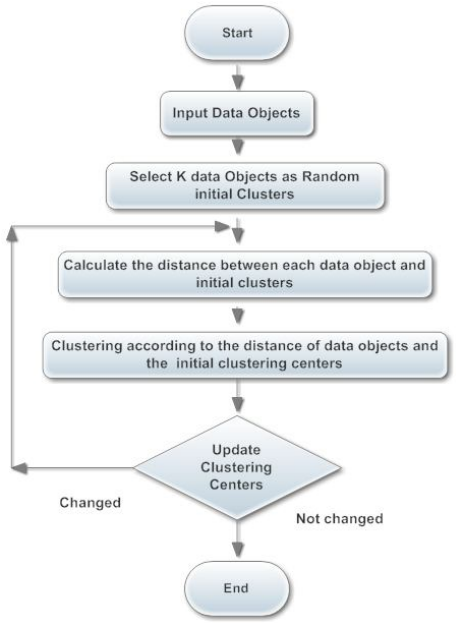


Fig Steps of K-means Algorithm in Schematic Representation

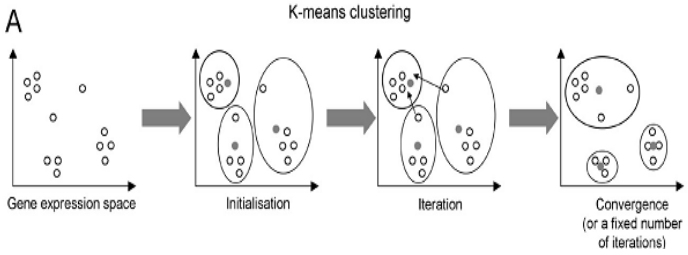


Fig. Clustering in different iterations.

III. K-means Clustering Problems

K-means clustering algorithm works on the assumption that the initial centers are provided. The search for the final clusters or centers starts from these initial centers. Without a proper initialization the algorithm may generate a set of poor final centers and this problem can become serious if the data are clustered using an on-line k-means clustering algorithm. In general, there are three basic problems that normally arise during clustering namely dead centers, local minima and centre redundancy.

Dead centers are centers that have no members or associated data. These centers are normally located between two active centers or outside the data range. The problem may arise due to bad initial centers, possibly because the centers have been initialized too far away from the data. Therefore, it is a good idea to select the initial centers randomly from the training data or to set them to some random values within the data range. However, this does not guarantee that all the centers are equally active. Some centers may have too many members and be frequently updated during the clustering process whereas some other centers may have only a few members and are hardly ever updated.

1V. Benefits of Improvement to K-Means Algorithm

The traditional K-means algorithm has a problem in choosing the initial clustering centres and the result obtained by traditional K-means algorithm varies with choice of initial clustering centres. So, to overcome this problem, improvement of traditional K-means is done in this paper. The graph theory concept, minimum spanning tree which relays on Kruskal's algorithm is introduced in the traditional K-means algorithm. By using Kruskal's algorithm concept the problem of traditional K mean is solved.

The benefit of improvement to traditional k-mean results in improvement of accuracy and better objective function value. And the same is achieved in less number of iteration. However, time consumed in improved k mean algorithm is nearly 5 times than the traditional k mean algorithm. This is because the process of obtaining the minimum spanning tree is relatively time consuming.

V. Kruskal's Algorithm for Improved K-means Algorithm

Sort edges in E based on cost.
T is empty (* T will store edges of a MST *).
Each vertex u is placed in a set by itself.
While E is not empty
Pick e = (u, v) 2 E of minimum cost
If u and v belong to different sets
Add e to T
Merge the sets containing u and v
Return the set T.

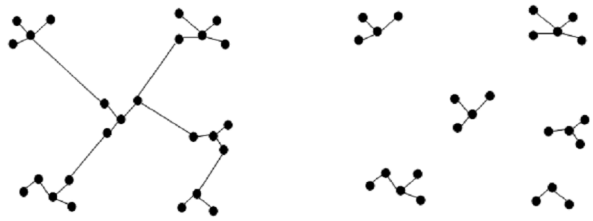


Fig Minimum spanning tree representation and six-group clustering

VI. Improved K-means Algorithm

Improved K-means algorithm has been designed implemented in this paper for the aim of achieving

performance increase when compared with K-means. Improvement of K-means algorithm has is the solution for the need of a faster K-means algorithm in order to cluster large amounts of datasets in reasonable durations. And, by using improved K-means it has been aimed to gather better accuracy clustering results than the traditional algorithm.

Traditional K-means algorithm is improved using the minimum spanning tree (MST) and Kruskal's algorithms. The Kruskal's algorithm which is based on the concept of minimum spanning tree (MST), improves the algorithm for better accuracy in less number of iterations. Since, the algorithm has to go through minimum spanning tree (MST) it takes time to finish the process of clustering.

The procedure flow of the improvement algorithm is it start with input of n number of data objects then, It calculate distance between any two data objects using Euclidean distance, To weights assigned to the edges and make all data object to generate undirected graph by which the minimum spanning tree have been generated to the clustered object using Kruskal's algorithm. Then the k-1 edges are deleted based on weights in descending order, the average value of the object contained by k connected graph are regarded as initial clusters, by which the problem of traditional k means problem is overridden**.**

The improved algorithm have been tested for the synthetic data as an example analysis of results with data, and then with gene expression data (yeast data) and also with iris dataset for comparative study. Main objective of paper is clustering of gene expression data and for checking the clustering accuracy we have used iris and synthetic data. The clustering results obtained for all the 3 datasets has been compared and discussed in the next chapter. The improved algorithm has resulted in achieving better accuracy in less number of iteration.

Start with input of n number of data objects then it calculate distance between any two data objects using Euclidean distance, To weights assigned to the edges and make all data object to generate undirected graph by which the minimum spanning tree have been generated to the clustered object using Kruskal's algorithm. Then the k-1 edges are deleted based on weights in descending order, the average value of the object contained by k connected graph are regarded as initial clusters, by which the problem of traditional k means problem is overridden

In the traditional K- means algorithm, in which it calculates the distance between any 2 input data objects as weight assigned to the edge, and also make all data objects (point) to generate undirected weighted graph. Next it calculates minimum spanning tree of the clustered objects using Kruskal's algorithm, which is discussed in the next chapter. Then K-1 edges are deleted based on the weights in descending order, then the average value the object contained by the K-connected graph are regarded as the ancient clusters.

By using Kruskal's algorithm concept we have solved the problem of traditional K mean, i.e. the result obtained by k mean algorithm varies with choice of initial clusters.

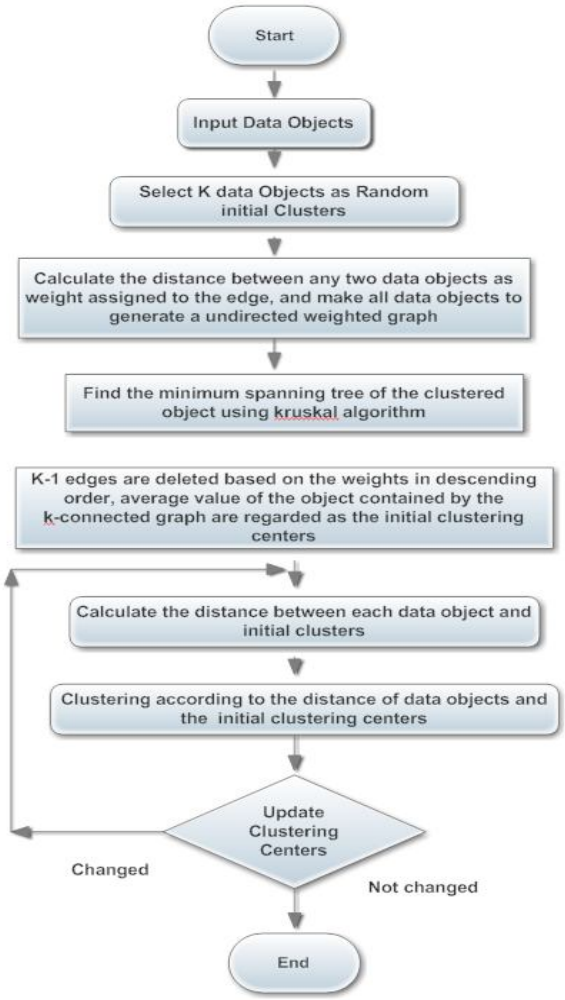


Fig Flow graph of Improved k-means

VII. Expectations from the Improved Algorithm

Aim of improvement in K-means algorithm was to improve the algorithm such that it will produce the best accuracy results with traditional algorithm. As no change has been performed on the K-means algorithm during clustering, it is expected that the algorithm will produce exactly same clustering results with traditional one. This can also be seen by the sample convergences of both the traditional algorithm and the improved algorithm. In the design of a improved algorithm, minimum spanning tree have been implemented to overcome the problem of choosing the initial clustering centers. In this paper, K-means algorithm has been designed carefully to run for all 3 datasets and also the Improved K-means algorithm also designed to run for the same 3 datasets and the results are been compared and analyzed, Results

have proved that Improved K-means algorithm has better accuracy in less number of iterations.

EXPERIMENTS AND RESULTS

Windows 7 operating system, Netbeans IDE , java JDK
Datasets of yeast and iris for inputs.

Execution Strategy for Testing the Algorithms

In order to compare two different versions (traditional and Improved) of K-means algorithm, both algorithms have been executed on three different sets of data (synthetic data, Yeast Dataset and Iris Dataset) mainly concerned with gene expression data (Yeast dataset) which are mentioned previously. Each execution is repeated for 5 times by using these datasets with K value 3 and 4, and so on for cluster centers with better results. Finally all the results of the three datasets will be gathered and analyzed and shown that the better accuracy have been achieved with Improved K-means algorithm.

The improved K-means algorithm is tested for the all the 3 datasets discussed above, for efficiency check the Improved K-means algorithm has been analyzed with different K values (2,3,4,5,10,13,15,30), in order to understand the relation of performance gain with the improved K-means. Trial of improved K-means algorithm with only one run is not considered as efficient so repeated run are been done as a part of testing and conclude that it has a better accuracy in less number of iterations.

So it is observed that the improved algorithm has improvement but the execution time increases almost proportional, this is because the flow control of the algorithm is gone through the minimum spanning tree which takes time is generating initial clusters. These results have been gathered as expected and will be presented in the following parts of the document.

It is important that, corresponding runs of each version should be checked for different values of K in order to compare the two algorithms in their accuracy and number of iterations with the execution time. The repeated run of the algorithm with different values of K will give us the efficient clustering results and makes easy to analyze between traditional and improved algorithms.

Data sets used:

Dataset Name	Narrative
Dataset1	Synthetic dataset
Dataset2	Yeast data
Dataset3	Iris data

Table Naming of Datasets

K values	Narrative
2	initial points for the dataset's 1st run
3	initial points for the dataset's 2nd run
4	initial points for the dataset's 3rd run
5	initial points for the dataset's 4th run
10	initial points for the dataset's 5th run
13	initial points for the dataset's 6th run
15	initial points for the dataset's 7th run
30	initial points for the dataset's 8th run

Table Naming of K values Points

Different initial numbers for K have been produced in different ranges for Dataset1, Dataset2 and Dataset3, because data stored in datasets have different ranges, these ranges have been selected randomly to check for datasets in order to have better clustering.

It will be better to name the algorithms also in order to refer them clearly in the following parts. As mentioned previously improved algorithm is executed on all the three different datasets and these executions can be considered for analysis of K-means and improved K-means algorithm. Naming of the mentioned versions of algorithms is presented in the table below.

Algorithm Name	Narrative
K-means algorithm	Traditional version
Improved K-means algorithm	Improved version

Table Naming of K-means Algorithm Versions

Finally, execution sequence of algorithms can be presented by using the naming standard listed above tables. Both algorithms have been executed by using each dataset (3 datasets) and by using each initial point set (different K initial points).

Result Name	Execution	Used datasets with K value
Result1	Run1	Dataset1 , K=2
Result2	Run2	Dataset 2,3 , K=3
Result3	Run3	Dataset2,K=4
Result4	Run4	Dataset 2, K=5
Result5	Run5	Dataset 2, K=10
Result6	Run6	Dataset2, K=13
Result7	Run7	Dataset2, K=15
Result8	Run8	Dataset2, K=30

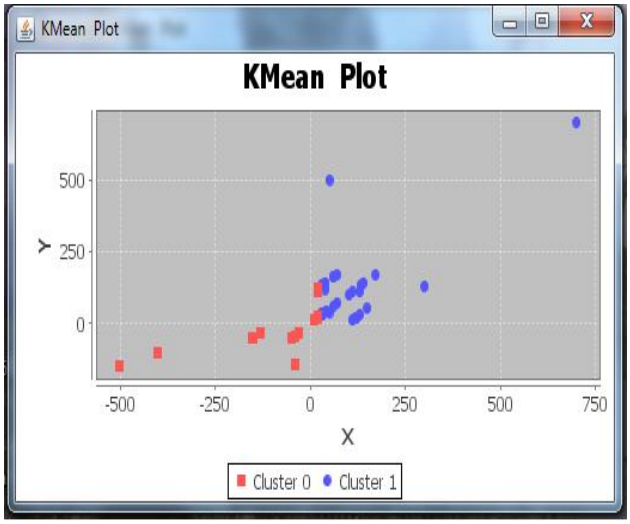
Table Expected Results of Executions

K value	K-means algorithm	Improved K-means algorithm
K=2	106	91
K=3	89	73

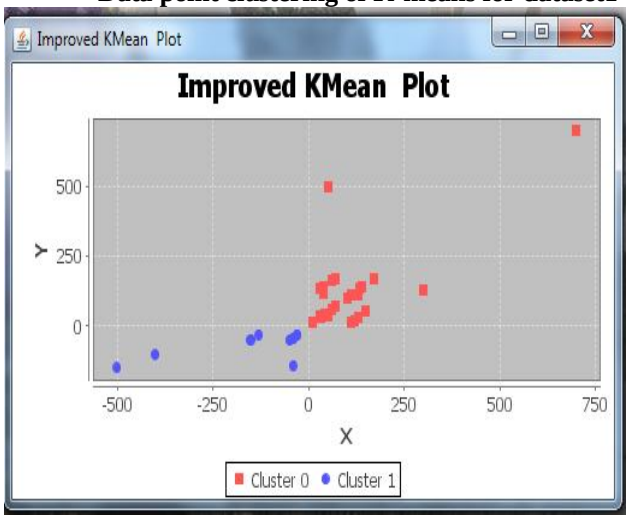
Results of Executions for Dataset1

Algorithm	Accuracy	Objective function	Iterations	Running time
K-means	78.987	106	6	58
Improved K-means	86.765	91	3	124

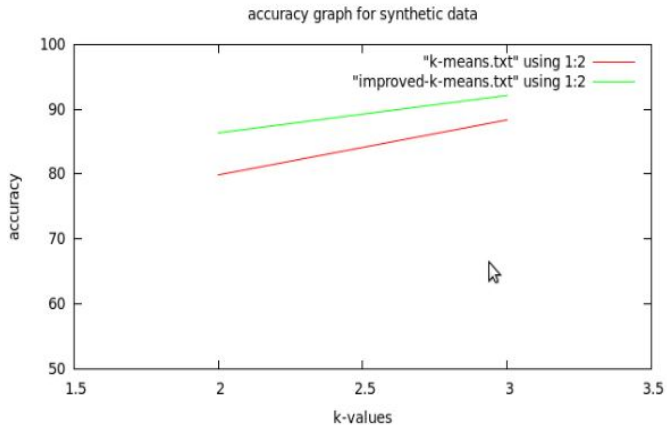
Analysis results for dataset1.



Data point clustering of K-means for dataset1



Data point clustering of improved K-means for dataset1



Accuracy graph for dataset1

Execution Results of Dataset2

In this section, execution results of Dataset2 (Yeast dataset) which is main aim of our paper, will be examined. This dataset will be clustered into different values of K , The cluster by using K value 2, 3, 4, 5, 10, 13, 15, 30 of different initial point sets. For this dataset, initial numbers of clusters are produced for all K values, with different values of K cluster groups. The attributes of dataset mostly occur between 1 and 36 as discrete values. Here we have used 8 attribute data of yeast cycle data and have clustered by both the version of K-means algorithms.
In the following tables the 100 yeast data points are been clustered by both the traditional and improved K-means algorithm with different initial cluster values and the results are been tabulated and proved that the improved version of K-means has better accuracy with less number of iteration but the time taken by improved K-means is more compared to traditional version.

K value	K-means algorithm	Improved K-means algorithm
K=2	104	98
K=3	94	91
K=4	79	64
K=5	75	62
K=10	69	58
K=13	67	63
K=15	76	73
K=30	59	54

Table Results of Executions for Dataset2.

K value	Algo	Accuracy	Objective function	Iterations	Running time
2	K-M	67.231	104	4	69
2	I K-M	73.012	98	2	141
3	K-M	87.987	94	5	89
3	I K-M	90.012	91	4	93
4	K-M	64.231	79	7	87
4	I K-M	79.012	64	3	99
5	K-M	57.231	75	9	78
5	I K-M	69.012	62	4	102

Table Analysis results for Dataset2

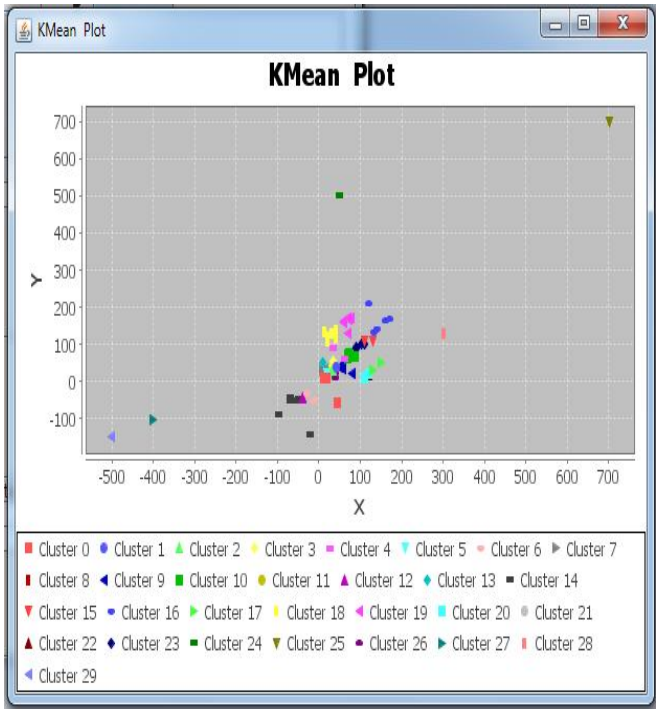


Fig Yeast data point clustering for K-means with k value 30

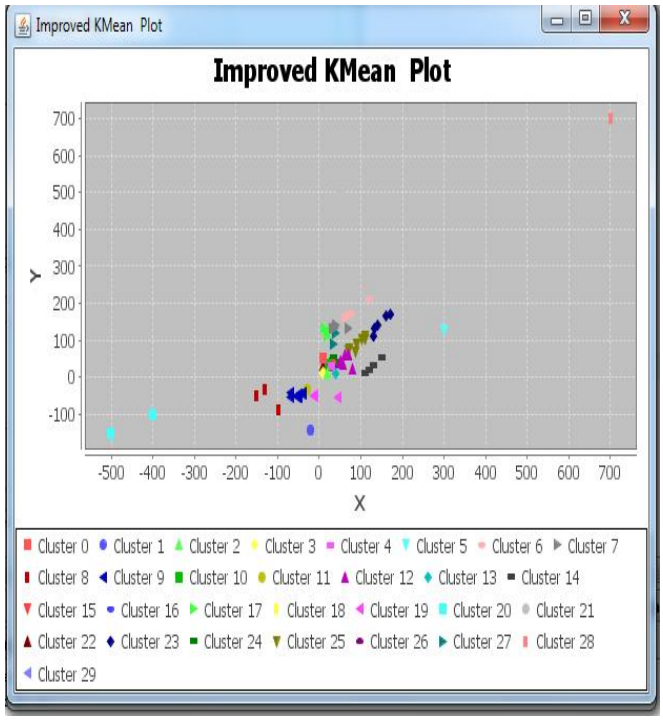


Fig Yeast data point clustering for Improved K-means with k value 30

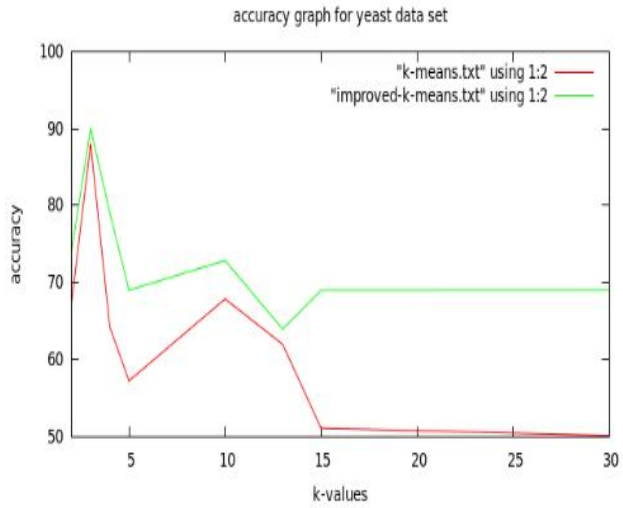


Fig Accuracy graph for dataset2

Execution Result's of Dataset3

In this section we have clustered iris dataset with 30 data points with 2 attributes and the clustering is done by both the versions of the algorithm and results are compared as shown in the below table.

K value	K-means algorithm	Improved K-means algorithm
K=3	112	98
K=4	98	76

Table 5.9 Results of Executions for Dataset3

Clustering algorithm	Accuracy	Objective function	Iterations	Running time
K-means	66.075	112	7	89
Improved K-means	79.543	98	3	97

Table Analysis results for Dataset3

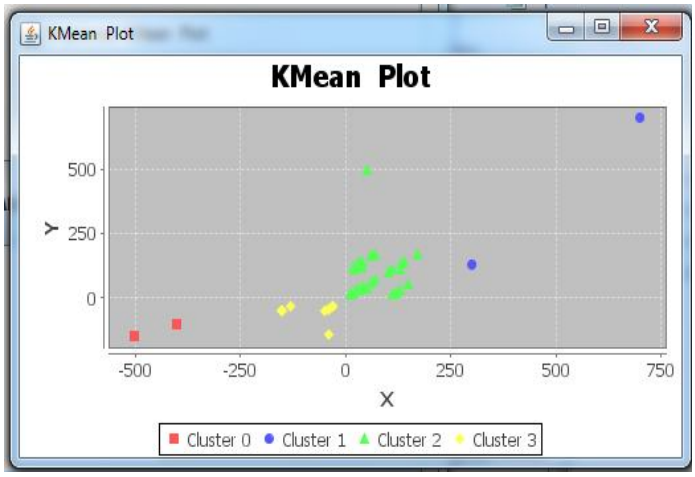


Fig for Iris data point clustering of K-means with k value4

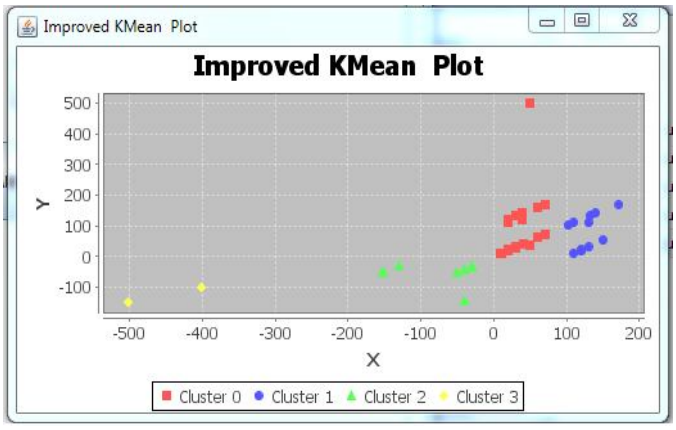


Fig for Iris data point clustering of Improved K-means with k value 4

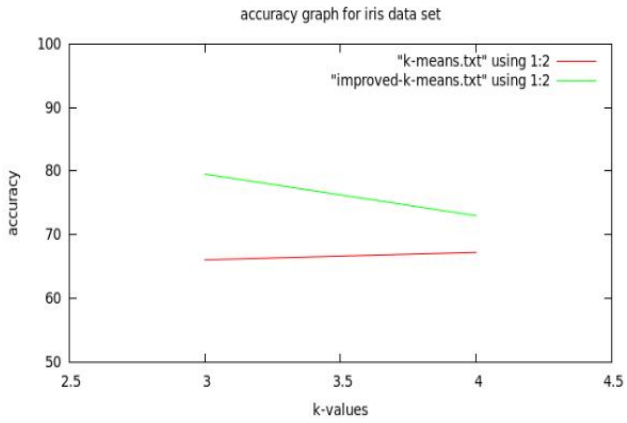


Fig for Accuracy graph for dataset3

VIII. CONCLUSION

In this paper, K-means algorithm and Improved K-means algorithm have been used for gene expression data (yeast dataset), and analysis is made with both the algorithms with different cluster groups. It has been stated by "Qian Ren and Xingjian Zhou" that, improvement version of the algorithm will produce better accuracy in less number of iteration with the traditional algorithm. Improved version of the algorithm has been designed and implemented by using Java language and traditional algorithm has also been implemented by Java language for the purpose of comparison with improved version. Simulation of the improved algorithm has been performed carefully, such that algorithm uses the concept of minimum spanning tree(MST), to the K-means algorithm, which transmit objects in whole dataset between processes in each iteration of the clustering algorithm.

Main aspect of this paper is to simulate K-means algorithm and improved K-means algorithm for different datasets, so that clustering with better accuracy is analysed with different cluster groups. When considering datasets of humans and different species, this improvement becomes very useful, because analysis of huge amount of gene expression has been made easy, as the datasets of all organisms have large amounts of data or sample. When examining traditional K-means algorithm, traditional K-means algorithm lacks in performance or in choice of initial clustering centre. A clustering algorithm based on the minimum spanning tree has been presented. The algorithm has been shown to be very effective in clustering multidimensional data sets. The algorithm has been tested on synthetic data and yeast cell cycle data from UCI, repository and iris datasets.

The results with all three datasets have been examined and compared and shown that improved version of algorithm has better accuracy and mainly it has proved that as the cluster groups increases the accuracy of the algorithm

also increases as there will be less number of misclassifications with more number of cluster groups. For the same we have conducted experiments with the different cluster groups and with different datasets and concluded.

REFERENCES

[1] A.K. Jain and R.C. Dubes, Algorithms for Clustering, prentice Hall, 1988.

[2] Webster, Two Crows Corporation 1999 Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery", 1999.

[3] Kiri Wagsta and Claire Cardie, Department of Computer Science, Cornell University, Ithaca, "Constrained K-means Clustering with Background Knowledge" USA, 2001.

[4] Kantabutra 1999 S. Kantabutra, "Parallel K-means Clustering Algorithm on NOWs", Department of Computer Science, Tufts University, 1999.

[5] Bashar Al-Shboul, and Sung-Hyon Myaeng "Initializing K-Means using Genetic Algorithms" World Academy of Science, Engineering and Technology. 2009.

[6] Min Feng College of Information Engineering. Taishan Medical University Taian 271016, China. "A Genetic K-means Clustering Algorithm Based on the Optimized Initial Centers" E-mail:fmxxsc@126.com. May 2011.

[7] Refining Initial Points for K-Means Clustering P. S. Bradley Microsoft Research Redmond, WA 98052, USA bradley@microsoft.com. May 1998.

[8] Eisen MB, Spellman P T, Brown PO, a1.Cluster analysis and display of genome-wide expression patterns [J]. Proc National Accad of Science, USA, 1998, 95:14863-14868.

[9] Y. Xu, V. Olman and D. Xu, Clustering gene expression data using a graph-theriotic approach: An application of minimum spanning trees, Bioinformatics, 18(2002) 536-545.

[10] Tamayo P,Slonim D,Mesirov Jet a1.Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation[J].Proc Natl Acad Sci USA,1999,96:2907-2912.

[11] Nikos Vlassis, JakobJ. Verbeek, The global k-means clustering algorithm, Department of Computer Science, University of Ioannina, 45110 Ioannina, Greece, 4 March 2002.

[12] Roy Kwang Yang Chang, Chu Kiong Loo and M.V.C. Rao, A Global k-means Approach for Autonomous Cluster Initialization of Probabilistic Neural Network, May 14, 2007.

Links to websites used in our research work

[19] http://archive.ics.uci.edu/ml/datasets/Yeast

[20] http://archive.ics.uci.edu/ml/datasets/Iris/.