Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7 July 2021: 12673 – 12686

Research Article

Gene Expression Data Clustering Using Improved K-Means Algorithm

S. Ranathive[1], Nelson Kennedy Babu C[2], Miretab Tesfayohanis[3], S.Sivakumar[4]

Abstract

Several K-means algorithm are available for clustering using various datasets of simulation. Yeast dataset and iris dataset are used for clustering by using K-means algorithm with numerous iteration and lower accuracy. For clustering, K-means algorithm of these datasets are simulated by enhanced version, Minimum spanning tree approach is used by Improved K-means algorithm. Each and every input data points are produced by an undirected graph and later shortest distance is computed by intern outcomes with increased accuracy with lower number of iterations. Java programming language was used for the simulation by these algorithms. Analysis and comparison of both the algorithms are resulted. Algorithms are run many times below various groups of clusters. From the outcomes, it is observed that better performance are acquired by improved K-means algorithm and contrasted with L-means algorithm. Accuracy is increased by the values of number of clusters. It is inferred that k's specific value increases the accuracy of this algorithm with optimal values.

Keywords: K-Means, Minimum Spanning Tree, Improved K-Means, Yeast dataset, iris dataset, accuracy, spanning tree.

1. Introduction

In recent years, Molecular biology field advancement and geometric technologies are coupled together for the extensive development of information from biology produced by the community of science. This genomic information's deluge was obtained as bioinformatics's introduction, genomics of computation and proteomics, entire genomes are analysed at large scale [1-3].Interdisciplinary study incorporating biology,computer science, mathematics and statistics for analysing the data of biological sequence, content of genomes and organization, and for the function prediction and macromolecule's construction is known as Bioinformatics. The methods

[1]Department of Computer Science and Engineering, VelTechRangarajanDr.Sagunthala R&D Institute of Science and Technology, Chennai, India.
[2]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai.
[3]Department of Information Technology, Dambi Dollo University, Ethiopia.
[4]Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, India
Corresponding Author:Email:miremsc2011@gmail.com

of computation are viewed for making the discoveries in biology.Functional gene product synthesis uses the information obtained from the gene and this process is known as Gene expression. Dividing the objects set(patterns) into disjoined groups set (clusters) is defined as the process of clustering. Data amount is decreased by categorization or groping the items of alike data for obtaining the necessary information is the major objective of clustering [4-6]. The clustering plot is indicated in figure 1.

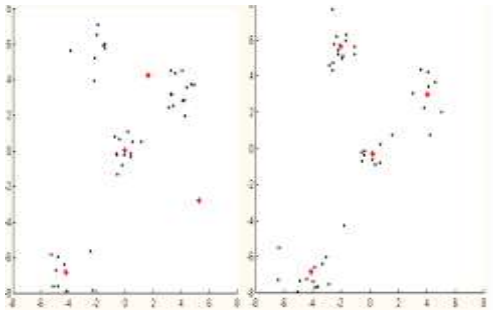

Fig.1. clustering plot

2. Related Works

In recent times, most widely utilized algorithm for clustering is K-means clustering algorithm. Partitioning idea are adopted by this algorithm and utilizing the error criteria of least mean square for dividing the information into various clusters. Unsupervised benefits are present with it and effective performance and superior effects of clustering are obtained when the data obeyed normal distribution. For data mining in big data applications, it is difficult to adopt the requirements when this K-means clustering algorithm is applied [7]. Applications of K-means clustering problems are improved by some process, they are: Alteration of datasets and center of clustering in processing Kmeans algorithm in Hadoop. The sequential and parallel execution are compared by keeping same value of other factors. Results demonstrate that this algorithm efficiently handles huge data sets in Hadoop environment. Artificial intellegience data DBSCAN clustering research are discussed in [9]. For Large-scale Artificial Intelligence data DBSCAN clustering algorithm depending on Hadoop platform in combined Hadoop platform. MapReduce parallel computing framework was used in algorithm of DBSCAN clustering that completely utilizes the advantages of Hadoop in large data processing by MapReduce distributed computing package and HDFS distributed storage in which algorithm's efficiency is improved greatly. The problem of selecting initial clustering center randomly is solved by the algorithm proposed in [10]. Underlying data structure's correlation are combined for the development of this algorithm producing the result of unreasonable clustering and by this process, strategy of initial clustering choices are improved by enhancing the clustering's convergence speed and efficiency of clustering. OClustR's two parallel version are presented in [11], particularly used by multi-core CPUs and GPUs for OClustR's efficiency enhancement in dealing the problems with numerous documents. Still, the efficiency of clustering is lower in many applications of data mining.

Similarity averages among each and every clusters are measured by Davies-Bouldin index (DB) [12] and its furthermost similar cluster. Cluster distances are maximized and distance among the cluster centroid and some data objects are minimized by attempts in the validity index of DB. Comparison of own cluster and other cluster are similar in the measurement of the Silhouette

value [13] and this range of this value lies from -1 to +1, in which the own cluster is matched well to the object is indicated by the higher value and the neighboring clusters are matched poorly. Indices of clustering validity are gathered into two main classification they are internal and external [14]. The results of clustering are evaluated by external indices for the comparing the memberships of clusters allocated by a algorithm of clustering with the prior identified information like externally supplied class label [15, 16]. Cluster goodness are evaluated by the structure of cluster through internal indices to focus the data with intrinsic information [17] because of this internal indices are considered. In literature, so many works are available to enhance the accuracy of data clustering, though it is efficient, proposed data clustering in gene expression data used to enhance the accuracy by improved K-means clustering algorithm

3. K-Means Algorithm

In order to increase the algorithm's effectiveness and accuracy, modification is required for K-mean algorithm. Minimum spanning tree (MST) technique and a portion of graph theory called Kruskal's algorithm is used for achieving the increment in accuracy and efficiency. These algorithms are analysed and simulated in this paper.Java programming Language is used as an implementation tool for the algorithm's enhanced version and conventional K-means algorithm. The enhancement in these algorithms are made for increasing accuracy with reduced iteration number and produce outcomes in specific duration. For the enhancement process, modification in algorithm properties are very significant to achieve improved accuracy and reduce the computation time for performing task and MST is used here for increasing accuracy with lowered iteration number. Few algorithms are appropriate to improve the accuracy and efficiency [18-24]. Improvement in success of algorithm is achieved by Improved K-means algorithm's properties. Sum of arc length on the tree with shortest spanning tree is defined as Minimum spanning tree (MST). Accuracy improvement in K-means algorithm is achieved by MST with reduced iteration number that are represented in figure 1 and figure 2 shows the Clustering in different iterations.

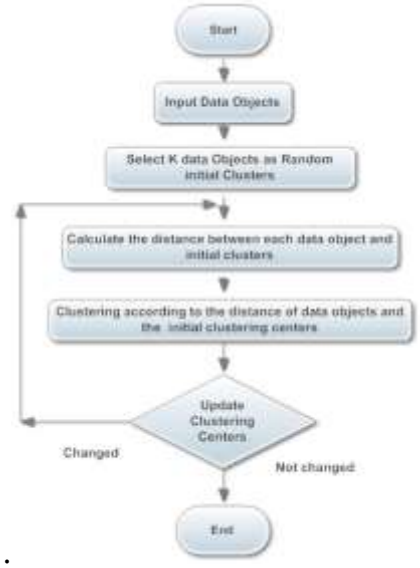

Fig.2. K-means Algorithm steps in Chart Representation

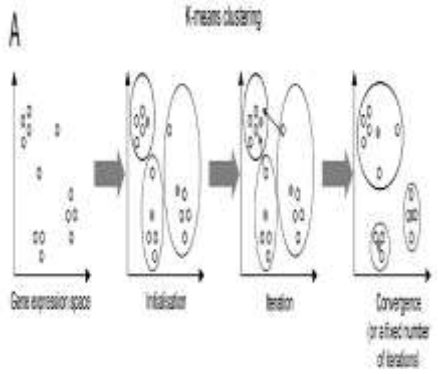

Fig.3. Different iterations present in Clustering.

4. Problems in K-means Clustering

Consideration for K-means clustering algorithm is that intial centers are given and from these initial centers, final cluster search or starting of center occurred. With no perfect initialization, poor final center set was generated by this algorithm and this issue will be thoughtful when data clustered utilizing an on-line k-means clustering algorithm. Generally, three fundamental issues arose in the process of clustering are dead centers, redundancy of centre and local minima. Members or associated data are not present in dead centers. Location of these Dead centers are present in amongst two active centers or exterior to the range of data. Bad initial centers arises problems probably because of the initialization of center that is very distant from the data. Thus this is the best plan to choose the random initial centers from training data or assigning few values randomly to it, which is inside the range of data. Also, the equal active state of centers are not guaranteed. Several members will be present in few centers and this will be updated often in the process of clustering, at the same time few centers will have few members and that are not updated hardly.

5. Advantages of Improved K-Means Algorithm

Problem of selecting clustering centre initially was faced by conventional K-means algorithm and its results are varied by the selection of initial clustering centres. In order to rectify this issue. In this paper conventional K-means algorithm is enhanced. Concept of graph theory along with minimum spanning tree are depending on Kruskal's algorithm is applied to the conventional scheme of K-means and because of this traditional K- mean algorithm' problems are solve.This improvement in conventional algorithm increases the accuracy with reduced iteration number which is the major motivation of this paper. Computation time required for this method is higher because MST process is time consuming task. Figure 3 represents representation of Minimum spanning tree and clustering in six-group.

Improved K-means Algorithm with Kruskal's Algorithm

Cost based E with Sort edges.

Empty T (*MST's edges are stored by T*).

Placement of every vertex is done by its own.

While Values present in E

Prefer e = (u, v) 2 least cost in E

If dissimilar sets are associated for u and v

T is added to e, u and v sets are merged, T set is returned.

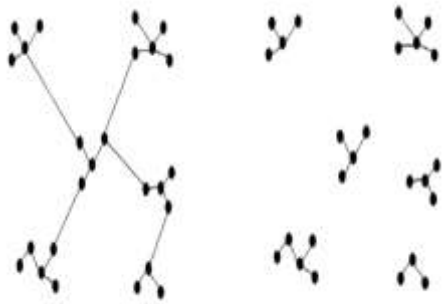

Fig.4. Minimum spanning tree representation and six-group clustering

6. Improved K-means Algorithm

Performance improvement in the K-means algorithm is achieved by improved K-means algorithm was premeditated and executed. Huge datasets are clustered by enhancement in K-means algorithm with sensible durations. Accuracy is improved by improved K-and improvement in results as compared to the conventional algorithm. Number of data objects are n and r provided as starting input in this improved K-means algorithm.

Later, distane among any two objects of data are computed by Euclidean distance. For assigning weights to the edges and making each and every data objects capable to produce undirected graph through it generation of MST are obtained for object clustering by utilizing Kruskal's algorithm. Depending on descending order of weights, it is needed to delete k-1 edges. Initial clusters are defined by the object's average value restricted by k connected graph which rectifies the issues occurred in conventional k means algorithm.

For the comparative study, this improved algorithm are applied with data of gene expression (yeast data) as well as iris dataset. Major motivation of this paper is data clustering of gene expression and to check the accuracy of clustering by using synthetic data and iris. All three datasets are compared based on the results of clustering and are described in the upcoming chapter. Better accuracy with lower iteration number are obtained by the improved K-means algorithm.

The problem associated with traditional K mean was solved by the Kruskal's algorithm's concept, i.e. the k mean algorithm results produced are varied by the selection of initial clusters.Figure 4 shows the Improved k-means Flow graph.

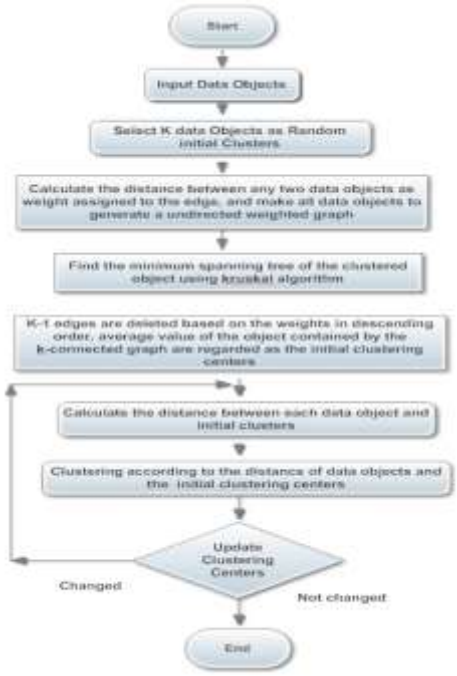

Fig.5. Improved k-mean's Flow graph

7. Expectations from the Improved Algorithm

Motivation towards improving K-means algorithm for producing best results of accuracy as compared to the conventional algorithm. If the clustering is performed without any alteration in the improved algorithm, then it will generate the exact results same as the results of conventional algorithm which is viewed through the convergences of samples in both traditional and improved algorithm. Initial clustering centre chosing problem in conventional approach is rectified by MST implementation. Thus, improved K-means algorithm is designed prudently for running each and every 3 datasets as like the conventional approach and the outcomes are analysed and compared. From the results it is observed that improved K-means algorithm produced improved accuracy with reduced iteration number.

8. Experiments And Results

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

Execution Approach for Algorithms Testing

Two various versions of algorithms i.e conventional K-means algorithm and Improved K-means algorithm are compared and in 3 various dataset types are used execution they are: synthetic data, Yeast Dataset and Iris Dataset. Major concentration is provided to gene expression data (Yeast Dataset). Value assigned for K are 3,4 and etc. are used in the dataset and are repeated for 5 times for every execution with enhanced results. At last, outcomes of these 3 datasets are clustered and evaluated for getting accuracy improvement by Improved K-means

algorithm.For various values of K (2,3,4,5,10,13,15,30), evaluation of improved K-means algorithm are performed in 3 datasets to prove the efficiency and to recognize the gain of performance through this algorithm. This execution is repeated until good accuracy with reduced iteration number is concluded. Since there is improvement in accuracy, time of execution is increased because of MST, in which time taken to generate initial clusters are increased and are described in the upcoming section of this paper.It is significant to check the respective repetitions of every version with various K values for the purpose of comparing these two algorithms with respect to accuracy, number of iterations and time of execution. Repetitions of this algorithm with various k values produces effective results of clustering which are simple for analysing the outcomes of conventional and improved algorithms.

Data sets used:

Table 1
Dataset Naming

Dataset Name	Narrative
1st Dataset	Synthetic dataset
2nd Dataset	Yeast data
3rd Dataset	Iris data

Table 2
K values Points Naming

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

Various initial K numbers are generated in various ranges for Dataset 1,2 and 3 as the stored data in these datasets consumes various ranges and these ranges are randomly chosen for checking datasets for achieving best clustering results.For convenience, the algorithms should be named for getting clear results in the upcoming sections. As stated earlier, 3 datasets are used for execution and analysis of this algorithm, table presented below shows the algorithm's version.

Table 3
K-means Algorithm Version's Naming

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

At last, sequence of execution for these algorithms was evaluated by utilizing the naming the list of standard in the aforementioned tables. Utilizing every 3 dataset and every initial point set (various K initial points) these two algorithms were executed.Figure 5 represents the Data point

clustering of K-means for dataset1 represented in figure 6 and Figure 7indicates dataset1's accuracy graph

Table 4
Expected outcomes of Executions

Execution	K value utilized in datasets	Result Name
Run1	K is assigned as 2 for 1st dataset	Result1
Run2	K is assigned as 3 for 2nd and 3rd dataset	Result2
Run3	K is assigned as 4 for 2nd dataset	Result3
Run4	K is assigned as 5 for 2nd dataset	Result4
Run5	K is assigned as 10 for 2nd dataset	Result5
Run6	K is assigned as 10 for 2nd 3dataset	Result6
Run7	K is assigned as 15 for 2nd dataset	Result7
Run8	K is assigned as 30 for 2nd dataset	Result8

Table 5
Results of Executions for Dataset1

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

Table 6
Analysis results for dataset1

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

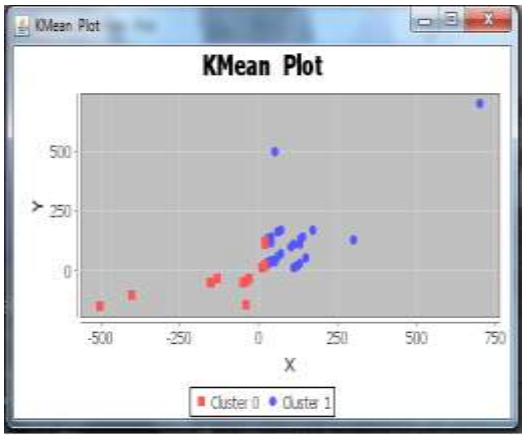

Fig.6. Data point clustering of K-means for dataset1

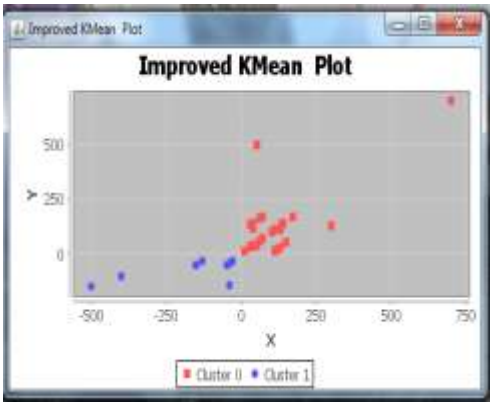

Fig.7. 1st dataset with Data point clustering of improved K-means

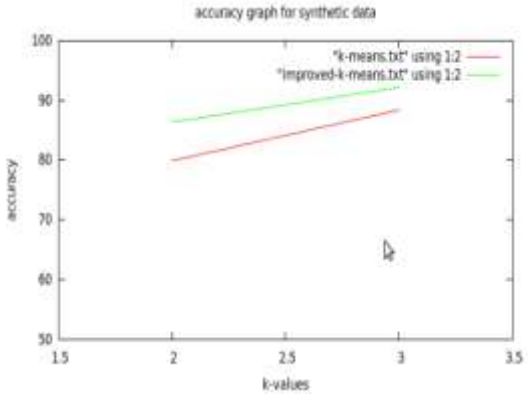

Fig.8. 1stdataset's accuracy graph

Execution Results of Dataset2

Results of execution for dataset 2 (Yeast dataset) is described in this section. Various K values are clustered in the dataset, The K value 2, 3, 4, 5, 10, 13, 15, 30 of various set of initial point are used in clustering. Cluster with initial number are generated for all values of K in this dataset, having various K cluster group's values. The dataset attributes are occurred mostly in-between the discrete values of 1 and 36. Yeast cycle data's 8 data attributes are used here and are clustered through these two K-means algorithm's version.

Clustering of 100 yeast data are performed by these two K-means algorithm with various values of initial cluster are provided in the following table and the tabulation also shows results of both the versions in which improved K-means produces best accuracy with reduced iteration number with increased execution time as it is contrasted with conventional one.Figure 8 For k value is 30, K-means result of Yeast data point clustering, Figure 9 shows For k value is 30, improved K-means result of Yeast data point clustering Figure 10 represents 2nd dataset accuracy graph

Table 7
Results of Executions for Dataset2

K value	Improved K-means algorithm	K-means algorithm
2	98	104

3	91	94
4	64	79
5	62	75
10	58	69
13	63	67
15	73	76
30	54	59

Table 8
Analysis results for Dataset2

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

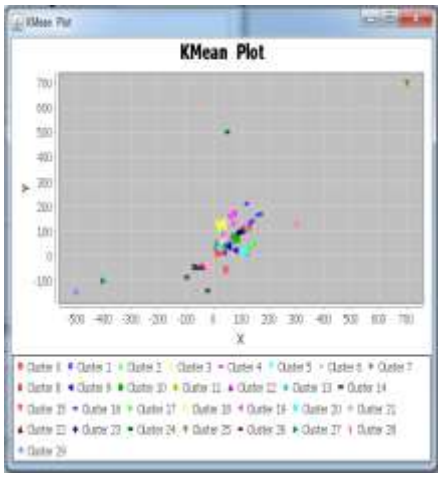

Fig.8. For k value is 30, K-means result of Yeast data point clustering

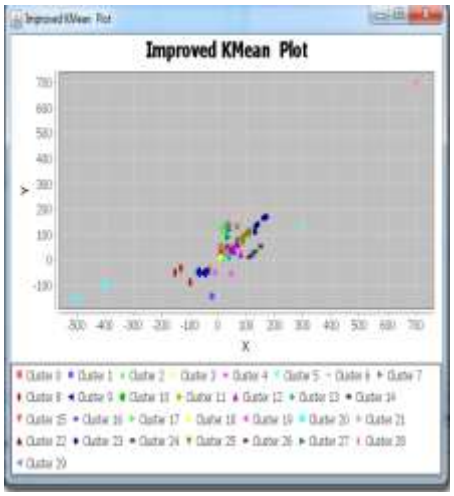

Fig.9. For k value as 30, Improved K-means results of Yeast data point clustering

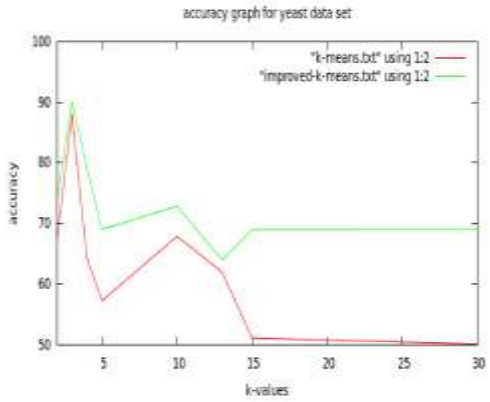

Fig.10. 2nd dataset accuracy graph

Execution Results of Dataset

Iris dataset are clustered in this section with 30 data points having two characteristics and Improved K-means and K-means algorithm performs clustering and outcomes of these two algorithms are compared and it is represented in the following table. Figure 11represents for k value 4, K-means Iris data point clustering, Figure 12 indicates k value 4, improved K-means Iris data point clustering. Figure 13 shows for 3rd dataset's Accuracy graph

Table 9
3rd Dataset's executions results

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

Table 10

Results Analysis for 3rd Dataset

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

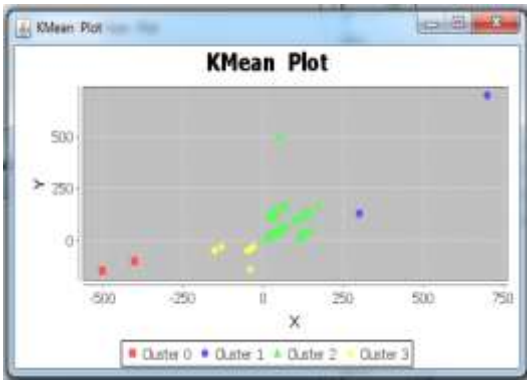

Fig.11. for k value 4, K-means Iris data point clustering

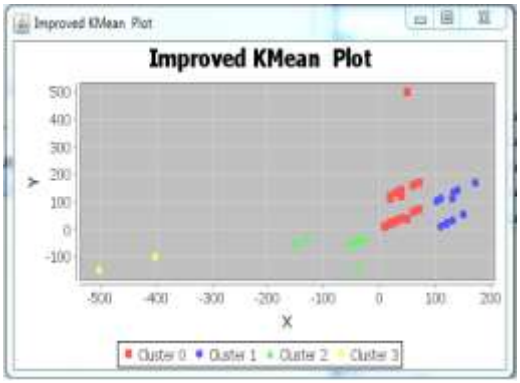

Fig.12. for k value 4, improved K-means Iris data point clustering

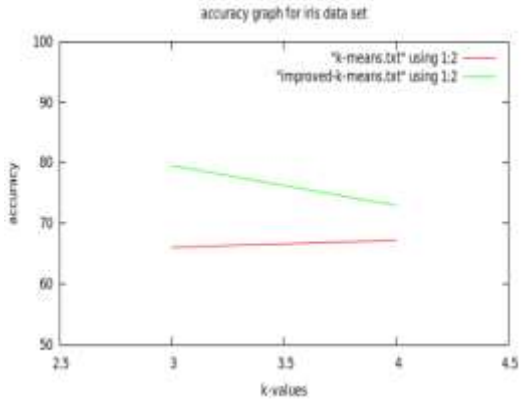

Fig.13. For Accuracy graph for dataset3

9. Conclusion

Analysis of gene expression data by K-means algorithm and Improved K-means algorithm with various groups of cluster are evaluated in this paper. Improved version of K-means algorithm produces the results with increased accuracy in lower iteration number as contrasted with conventional algorithm is stated by "Qian Ren and Xingjian Zhou". Java language is utilized for the design and implementation of improved K-means algorithm. Careful simulation are carried out for this improved K-means algorithm with the MST concept whereas traditional K-means algorithm utilizes the objects transmission in entire dataset during the every iteration process of this algorithm. Improved and traditional version of algorithm are simulated in three various datasets for achieving good results of accuracy having various groups of cluster. Traditional K-means algorithm performance lack due to selection of initial clustering centre. Minimum Spanning Tree forms the basis for clustering algorithm. Because of this MST, efficiency of algorithm is high for multidimensional clustering datasets. The yeast cell cycle data and synthetic data are used for testing these algorithms from UCI, repository and iris datasets.All 3 datasets are evaluated and are contrasted and are proved that this improved K-means produces better results and it is established that algorithm's accuracy is improved by increasing the groups of cluster as well as with reduced misclassification number. Experiments are conducted for various groups of cluster with various datasets and are concluded.

Future Enhancement

As the future work, the improvement in the K-means algorithm can be made by using the optimal minimum spanning tree technique, instead of using normal minimum spanning tree, we use Optimal minimum spanning tree which will do the same function as minimum spanning tree in less amount of time, which is efficient in improving the run time and make improved K-means more efficient and effective.

References

1. A.K. Jain and R.C. Dubes, Algorithms for Clustering, prentice Hall (1988).
2. Webster, Two Crows Corporation, 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 (2001).
4. Kantabutra S. Kantabutra, Parallel K-means Clustering Algorithm on NOWs, Department of Computer Science, Tufts University(1999).
5. Bashar Al-Shboul, and Sung-HyonMyaeng, Initializing K-Means using Genetic Algorithms, World Academy of Science, Engineering and Technology (2009).
6. Min Feng College of Information Engineering. TaishanMedical University Taian 271016, China. A Genetic K-means Clustering Algorithm Based on the Optimized Initial Centers (2011).
7. Salabat, K., Amir, K., Muazzam, M., Optimized Gabor feature extraction for mass classification using cuckoo search for big data E-healthcare, J. Grid Comput. 17(2) (2019)239–254
8. Bandyopadhyay, S.S., Halder, A.K., Chatterjee, P., HdK-means: Hadoop based parallel K-means clustering for big data IEEE Calcutta Conference, (2018) 452–456.
9. Chen, Z., Guo, J., Liu, Q, DBSCAN algorithm clustering for massive AIS data based on the Hadoop platform 2017 International Conference on Industrial Informatics - Computing

Technology, Intelligent Technology, Industrial Information Integration (ICIICII) (2017) 25–28
10. Ye, K., Jiang, X., He, Y.,Hadoop: a scalable Hadoop virtual cluster platform for mapreduce-based parallel machine learning with performance consideration, IEEE International Conference on Cluster Computing Workshops (2012)152–160
11. Soler, L.J.G., Suárez, A.P., Chang, L, Efficient overlapping document clustering using GPUs and Multi-core systems, Iberoamerican Congress on Pattern Recognition Ciarp (2014) 264–271
12. D. Davies and D. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence (1979) 224–227.
13. P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20(1987)53-65.
14. E. Rendon, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus external cluster validation indexes, Int. J. Computers and Communications 5(2011) 27- 34.
15. Y. Lei, J.C. Bezdek, S. Romani, N.X. Vinh, J. Chan, J. Bailey, Ground truth bias in external cluster validity indices, Pattern Recognition 65(2017)58-70.
16. J. Wu, J. Chen, H.Xiong, M. Sie, External validation measures for k-means clustering: a data distribution perspective, Expert Syst. Appl., 36(2009)6050- 6061.
17. L.J. Deborah, R. Baskaran, A. Kannan, A survey on internal validity measure for cluster validation, Int. J. Comput. & Eng. Surv. 1 (2010)85-102.
18. Sampathkumar, A., Murugan, S., Sivaram, M., Sharma, V., Venkatachalam, K., Kalimuthu, M, Advanced Energy Management System for Smart City Application Using the IoT, Internet of Things in Smart Technologies for Sustainable Urban Development (2020)185-194.
19. Sampathkumar, A., Murugan, S., Rastogi, R., Mishra, M. K., Malathy, S., & Manikandan, R, Energy efficient ACPI and JEHDO mechanism for IoT device energy management in healthcare, Internet of things in smart technologies for sustainable urban development(2020)131-140.
20. Sampathkumar, A., Maheswar, R., Harshavardhanan, P., Murugan, S., Jayarajan, P., &Sivasankaran, V, Majority voting based hybrid ensemble classification approach for predicting parking availability in smart city based on IoT. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2020) 1-8.
21. Sampathkumar, A., Murugan, S., Elngar, A. A., Garg, L., Kanmani, R., & Malar, A, A novel scheme for an IoT-based weather monitoring system using a wireless sensor network. In Integration of WSN and IoT for Smart Cities (2020) 181-191.
22. Sampathkumar, A., Rastogi, R., Arukonda, S., Shankar, A., Kautish, S., &Sivaram, M, An efficient hybrid methodology for detection of cancer-causing gene using CSC for micro array data, Journal of Ambient Intelligence and Humanized Computing 11(11) (2020) 4743-4751.
23. Sampathkumar, A., Mulerikkal, J., &Sivaram, M, Glowworm swarm optimization for effectual load balancing and routing strategies in wireless sensor networks, Wireless Networks 26(6) (2020)4227-4238.
24. Manikandan, V., Gowsic, K., Prince, T., Umamaheswari, R., Ibrahim, B. F., &Sampathkumar, A, DRCNN-IDS Approach for Intelligent Intrusion Detection System, In 2020 International Conference on Computing and Information Technology (ICCIT-1441) (2020) 1-4.