

Implementation of K-Means Clustering Algorithm Using Java

Prof. S.China Venkateswarlu¹, Prof. M.Arya Bhanu² and Prof.Yudhaveer Katta³

¹ JNTUH, HYDERABAD

Received: 3 August 2011 Accepted: 1 September 2011 Published: 11 September 2011

Abstract

Emergence of modern techniques for scientific data collection has resulted in large scale accumulation of data pertaining to diverse fields. Conventional database querying methods are inadequate to extract useful information from huge data analysis. Cluster analysis is one of the major data analysis methods and k-means clustering algorithm Emergence of modern techniques for scientific data collection has resulted in large scale accumulation of data pertaining diverse felids. Conventional Data base methods are inadequate to extract useful information from huge data banks. Cluster analysis is one of the major data analysis methods and the k-means clustering algorithm is widely used for many practical applications. But the original k-means algorithm is computationally expensive and the quality of the resulting clusters heavily depends on the selection of initial cancroids. Several methods have been proposed in the literature for improving the performance of the k-means clustering algorithm. The k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations. This papert proposes a method for making the algorithm more effective and efficient.

Index terms— About four key words or phrases in alphabetical order, separated by commas.

1 INTRODUCTION

a) Module design and organization i. Fixed Transmission n this module the fixed transmission is computed by retrieving each link which is limited to maximum signal ratio. Then the fixed routes are computed which give the maximum transmitted power per node and is limited to hardware constraint. Then the distance between the nodes is compute to calculate the end-to-end reliability. The end-to end reliability is a decreasing function, which can be treated as the cost metric for route selection.

ii. End to End Reliability This module focuses on the problem of optimizing transmission power levels and route selection on an end-to end basis. This module minimizes the endto end power for a fixed route. The end-to-end route reliability under the optimal power allocation scheme is represented as any fixed route, different power allocation schemes result in different end-to-end reliability and power consumption. The next step is to retrieve the total bandwidth then the distance between nodes are minimized and reliability routes are computed.

iii. Outage Diversity

In the module the case of a point-to-point link, is considered and the trade-off between route outage and consumed power in a network setting. This type of analysis gives insight to how fast the end-to-end outage decreases as more power is spent on the transmission. First, we look at the case that the maximum transmitted power at each link is fixed. It is observed that the route selection does not have any effect on the form of this tradeoff. By selecting the optimal route, we minimized the end to end outage probability by minimizing.

This shows that as long as we limit our approach to a single transmitter and a single receiver per link, even under optimal power allocation and route selection, the trade-off maintains the same form as in the single link case. Phase 1 of the heuristic algorithm requires a time complexity of $O(nkp)$ for finding the initial centroids, as the maximum time required here is for computing the distances between each data point and all other data-points in the set D . In the original k-means algorithm, before the algorithm converges, the centroids are calculated

1 INTRODUCTION

many times and the data points are assigned to their nearest centroids. Since complete redistribution of the data points takes place according to the new centroids, this takes $O(nkl)$, where n is the number of data-points, k is the number of clusters and l is the number of iterations. To obtain the initial clusters, algorithm 4 requires $O(nk)$. Here, some data points remain in its cluster while the others move to other clusters depending on their relative distance from the new centroid and the old centroid. This requires $O(1)$ if a data-point stays in its cluster, and $O(k)$ otherwise. As the algorithm converges, the number of data points moving away from their cluster decreases with each iteration. Assuming that half the data points move from their clusters, this requires $O(nk/2)$. Hence the total cost of this phase2 of the heuristic algorithm is $O(nk)$, not $O(nkl)$. Thus the overall time complexity of the heuristic algorithm becomes $O(nkp)$.

In an Enhanced k-means approach [??4] we are some of the data points may become more closer to a different cluster and such points are redistributed accordingly. The entire process is repeated until no more data points cross cluster boundaries. The enhanced method is described as Algorithm 4. This method involves keeping track of the distance between each data point and the centroid of its nearest cluster.

During the subsequent iteration, instead of computing the distance of the data point from all cluster centroids, its distance from the previous nearest cluster alone is determined. If that distance is less than or equal to the previous nearest distance, the data point.

b) Algorithm: finding the initial centroids i. Finding Initial Centroids Input: $D=\{d_1, d_2, \dots, d_n\}$ // set of n data items K // number of desired cluster Output: A set of k initial centroids.

Steps:

1. Set $m=1$: 2. Compute the distance between each data point and all other data-points in the set D : 3. Find the closest pair of data points from the set D and form a data-point set $A_m (1 \leq m \leq k)$ which contains these two data-points. Delete these two data points from the set D ; 4. Find the data point in D that is closest to the data point set A_m . Add it to A_m and delete it from D : 5. Repeat step 4 until the number of data points in A_m reaches $0.75 * (n/k)$; 6. If $m < k$, then $m=m+1$, find another pair of data points from D between which the distance is the shortest, from another data -point set A_m and delete them from D , Go to step 4: 7. For each data-point set $A_m (1 \leq m \leq k)$ find the arithmetic mean of the vectors of data points in A_m , these means will be the initial centroids. Algorithm 3 describes the method for finding initial centroids of the clusters [??12]. Initially, compute the distances between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A_1 consisting of these two data points, and delete them from the data point set D . Then determine the data point which is closest to the set A_1 , add it to A_1 and delete it from D . Repeat this procedure until the number of elements in the set A_1 reaches a threshold. At that point go back to the second step and form another data-point set A_2 . Repeat this till 'k' such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids. In the first phase, the initial centroids are determined systematically so as to produce clusters with better accuracy [??12]. The second phase makes use of a variant of the clustering method discussed in [??4]. It starts by forming the initial clusters based on the relative distance of each data-point from the initial centroids. These clusters are subsequently fine-tuned by using a heuristic approach, thereby improving the efficiency. The two phases of the enhanced method are described below as Algorithm 3 and Algorithm 4.

The first step in Phase 2 is to determine the distance between each data-point and the initial centroids of all the clusters. The data-points are then assigned to the clusters having the closest centroids. This results in an initial grouping of the data-points. For each data-point, the cluster to which it is assigned (ClusterId) and its distance from the centroid of the nearest cluster (Nearest_Dist) are noted. Inclusion of data-points in various clusters may lead to a change in the values of the cluster centroids. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. Up to this step, the procedure is almost similar to the original k-means algorithm except that the initial centroids are computed systematically. The next stage is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the datapoints may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data-point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. This results in the saving of time required to compute the distances to $k-1$ cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the datapoint than its previous centroid, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. Identify the new nearest cluster and record the new value of the nearest distance. The loop is repeated until no more data-points cross cluster boundaries, which indicates the convergence criterion. The heuristic method described above results in significant reduction in the number of computations and thus improves the efficiency. c) Algorithm: assigning data-points to clusters [??1] Assigning DataPoints to Centroids Input: manner. Once the initial centroids are thus determined, the distance between each data point and centroids of all the clusters are determined and the data points are included in the nearest cluster. Cluster means are then recalculated to find the new centroids. As a result of this, centroid. The algorithm proceeds in the following Phase 1 of 2 the enhanced algorithm requires a time complexity of $O(n)$ for finding the initial centroids, as the

max-imum time required here is for computing the distances between each data point and all other data-points in the set D. In the original k-means algorithm, before the algorithm converges the centroids are calculated many times and the data points are assigned to their nearest centroids. Since complete redistribution of the data points takes place according to the new centroids, this takes $O(nkl)$, where n is the number of data-points, k is the number of clusters and l is the number of iterations. To obtain the initial clusters, Algorithm 4 requires $O(nk)$. Here, some data points remain in its cluster while the others move to other clusters depending on their relative distance from the new centroid and the old centroid. This requires $O(1)$ if a data-point stays in its cluster, and $O(k)$ otherwise. As the algorithm converges, the number of data points moving away from their cluster decreases with each iteration. Assuming that half the data points move from their clusters, this requires $O(nk/2)$. Hence the total cost of this phase of the algorithm is $O(nk)$, not $O(nkl)$. Thus the overall time complexity of the enhanced algorithm (Algorithm 2) becomes $O(n)$, since k is much less than n .

II.

2 CONCLUSION

In this section we have shown how testing is performed and different test cases are designed to test the system for its performance as well as debugging process. The validation of the test cases is also shown. The k-means algorithm is widely used for clustering large sets of data. But the standard algorithm does not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. Moreover, the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times, during every iteration of the loop. This Project presents an enhanced k-means algorithm which combines a systematic method for finding initial centroids and efficient way for assigning data points to clusters.

3 III. IMPLEMENTATION AND RESULTS

4 a) Introduction

In the module the case of a point-to-point link, is considered and the trade-off between route outage and consumed power in a network setting. This type of analysis gives insight to how fast the end-to-end outage decreases as more power is spent on the transmission. First, we look at the case that the maximum transmitted power at each link is fixed. It is observed that the route selection does not have any effect on the form of this tradeoff. By selecting the optimal route, we minimized the end to end outage probability by minimizing

5 IV. IMPLEMENTATION AND RESULTS

Algorithm 3 describes the method for finding initial centroids of the clusters [??12]. Initially, compute the distances between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data-point set A2. Repeat this till 'k' such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids. In the first phase, the initial centroids are determined systematically so as to produce clusters with better accuracy [??12]. The second phase makes use of a variant of the clustering method discussed in [??4]. It starts by forming the initial clusters based on the relative distance of each data-point from the initial centroids. These clusters are subsequently fine-tuned by using a heuristic approach, thereby improving the efficiency. The two phases of the VI.

6 CONCLUSION

In this section we have shown how testing is performed and different test cases are designed to test the system for its performance as well as debugging process. The validation of the test cases is also shown. The k-means algorithm is widely used for clustering large sets of data. But the standard algorithm does not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. Moreover, the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times, during every iteration of the loop. This Project presents an enhanced k-means algorithm which combines a systematic method for finding initial centroids and

¹© 2011 Global Journals Inc. (US)

²© 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XVII Version



Figure 1: ©

```

root@localhost:~#
1 94 97 99 100 107
No of elements in the cluster is: 27

elements in the 2 th cluster is:
1 2 3 4 7 9 10 12 13 14 23 25 26 30 31 33 36 38 39 42 43 46 48
50
No of elements in the cluster is: 23

elements in the 4 th cluster is:
1 5 6 8 11 15 16 17 18 19 20 21 22 24 27 28 29 32 33 34 37 40
41 44 45 47 49
No of elements in the cluster is: 27

*****PROGRAM END*****

real    0m0.073s
user    0m0.116s
sys     0m0.024s
[root@localhost ~]# time java dummy
*****
A Heuristic Method to improve the efficiency of the Kmeans Clustering Alg
*****

Enter desired no of clusters k: 1
the value of k is:1
Obtained cluster centroid are:
new 0 th cluster centroid is:
5.84031303130313
3.0540000000000007
3.798666666666667
1.1906666666666672

elements in the 0 th cluster is:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66
67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88
89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110
111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132
133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
No of elements in the cluster is: 150

*****PROGRAM END*****

real    0m0.764s
user    0m0.073s
sys     0m0.021s
[root@localhost ~]#

```

Figure 2:

```

root@localhost:~#
No of elements in the cluster is: 150

*****PROGRAM EN*****

real    0m0.862s
user    0m0.077s
sys     0m0.017s
[root@localhost ~]# time java dummy
*****
A Heuristic Method to improve the efficiency of the kmeans Clustering Alg
*****

Enter desired no of clusters k: 2
the value of k is:2
Obtained cluster centroid are:
new 0 th cluster centroid is:
4.30181082780385
1.8845979381441103
4.588742884597816
1.4584741884597816

new 1 th cluster centroid is:
5.008460377384941
3.340377384903167
1.541244180943184
0.188479145180188

elements in the 0 th cluster is:
  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149  150
No of elements in the cluster is: 97

elements in the 1 th cluster is:
  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99
No of elements in the cluster is: 53

*****PROGRAM EN*****

real    0m0.174s
user    0m0.126s
sys     0m0.013s
[root@localhost ~]#

```

Figure 3:

```

root@localhost:~#
ayy Om0.022s
[root@localhost ~]# time java dummy
*****
A Heuristic Method to Improve the Efficiency of the Kmeans Clustering Alg
*****

Enter desired no of clusters k: 3
the value of K is:3
Desired cluster centroids are:
new 0 th cluster centroid is:
5.945144497477419
2.769170762110372
4.482486949119212
1.472347011304412

new 1 th cluster centroid is:
5.00205412112783
0.188413290829429
1.5081807340021504
0.241805815973274

new 2 th cluster centroid is:
4.83170092881963
1.0652077121204904
4.483787779808356
2.0505462056086777

elements in the 0 th cluster is:
  51  52  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73
  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95
  96  97  98  99 100 102 107 114 115 120 122 124 127 128 134 139 140 147 150
No of elements in the cluster is: 62

elements in the 1 th cluster is:
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
 45 46 47 48 49 50
No of elements in the cluster is: 50

elements in the 2 th cluster is:
 83  78 101 103 104 105 106 108 109 120 111 112 113 114 117 118 119 121 123 125 126
130 131 132 133 135 136 137 138 140 141 142 144 145 146 148 149
No of elements in the cluster is: 38

*****PROGRAM END*****
*****

real    Om0.012s
user    Om0.118s
sys      Om0.022s
[root@localhost ~]#

```

Figure 4:

```

root@kali:~#
1.50077403434004
0.0000000000000000

new 2 th cluster centroid is:
5.90761217816009
1.7514011181677047
0.4112463355154035
1.410380266747770

new 3 th cluster centroid is:
4.882449407114525
1.2412874047741535
1.840174819104081
0.7196111144490002

elements in the 0 th cluster is:
  50   78   101  120   124   105   124   108   100   1
10   111  121   113   116   117   118   110   121   1
16   124   129   130   131   132   133   135   136   137   1
30   140   141   142   144   145   146   148   149

No of elements in the cluster is: 38

elements in the 1 th cluster is:
  1    5    6   11   15   16   17   18   19   2
3    21   22   24   27   28   29   31   32   34   3
7    40   41   44   45   47   48

No of elements in the cluster is: 33

elements in the 2 th cluster is:
  31   32   34   35   36   37   38   39   40   4
1    42   43   44   45   46   47   48   49   50   5
1    52   53   54   55   56   57   58   59   60   6
2    63   64   65   66   67   68   69   70   71   7
2    80   84   85   86   87   88   89   90   91   9
07   104   105   110   112   124   127   118   114   119   1
41   147   150

No of elements in the cluster is: 62

elements in the 3 th cluster is:
  2    3    4    7    8    9   10   11   13   3
4    20   21   24   25   32   35   36   38   39   4
1    42   46   48   52

No of elements in the cluster is: 24

*****PROGRAM END*****

real    0m0.715s
user    0m0.116s
sys     0m0.011s
[root@kali:~]#

```

Figure 5:

```

root@localhost:~#
=====
*****PROGRAM END*****
=====

real    0m0.715s
user    0m0.128s
sys     0m0.021s
(root@localhost)~# time java dummy
=====
A heuristic method to improve the efficiency of the kmeans Clustering Alg
=====

Enter desired no of clusters k: 5
the value of k is: 5
Obtained cluster centroids are:
new 0 th cluster centroid is:
6.10791894840716
1.814305287048151
1.66561768771543
1.5072136309431104

new 1 th cluster centroid is:
6.912441724941728
1.081866997914187
6.821883601478222
3.107188637841661

new 2 th cluster centroid is:
6.440008946117288
1.568586018018915
6.786873214610076
1.173891608380637

new 3 th cluster centroid is:
4.719913010109215
1.14297871899312417
1.4145407451089455
0.201310876501670

new 4 th cluster centroid is:
0.221543170019003
1.647409575416544
1.488443881643434
0.1727993710016473

elements in the 0 th cluster is:
  11   32   51   59   71   79   84   96   97   98   71   72   74   75   76   77   78   79   84   85   87   8
  8   92   98   102   113   122   114   115   120   121   124   127   128   129   135   138   143   147   148   150
No of elements in the cluster is: 41

elements in the 1 th cluster is:
  101   103   104   106   104   108   109   110   113   114   117   118   119   121   123   125   124   129   133   131   132   1

```

Figure 6:


```

root@localhost:~#
dev 1 th cluster centroid is:
0.442008146117259
1.563386401891901
0.7868702204510071
1.173393608380517

dev 2 th cluster centroid is:
4.79981000089215
1.1419787199012417
1.4143407491089481
0.10311187682124795

dev 4 th cluster centroid is:
0.221369234055003
1.447499870414644
1.4984682531843834
0.17179971709386673

elements in the 0 th cluster is:
  51  52  53  54  57  59  64  66  67  68  71  73  74  75  76  77  78  79  84  86  87  8
8  92  96  102  111  112  114  115  116  117  118  119  124  125  128  143  147  148  150
No of elements in the cluster is: 41

elements in the 1 th cluster is:
  101  103  104  105  106  108  109  110  113  116  117  118  119  121  123  125  126  129  130  131  132  1
23  136  137  138  140  141  142  144  145  146  149
No of elements in the cluster is: 32

elements in the 2 th cluster is:
  54  56  58  60  61  62  63  65  66  68  70  72  80  81  82  83  85  89  90  91  93  94  9
6  96  97  99  100  107
No of elements in the cluster is: 27

elements in the 3 th cluster is:
  1  3  4  7  9  10  12  13  14  15  16  18  19  20  21  22  24  25  26  28  29  30  31  32  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50
No of elements in the cluster is: 33

elements in the 4 th cluster is:
  1  3  6  8  11  13  16  17  18  19  20  21  22  24  25  26  29  31  32  34  37  4
0  41  44  45  47  49
No of elements in the cluster is: 27

*****PROGRAM END*****

real    0m0.073s
user    0m0.126s
sys     0m0.026s
[root@localhost ~]#

```

Figure 7:

```

root@localhost:~#
login as: root
root@192.168.1.100's password:
Last login: Sat Nov 6 15:38:14 2010 from 192.168.1.76
[root@localhost ~]# time java dummy
*****
A Heuristic Method to improve the efficiency of the kmeans Clustering Alg
*****

Enter desired no of clusters k: 5
the value of k is:5
^C
real    0m11.020s
user    0m0.150s
sys     0m0.014s
[root@localhost ~]# time java dummy
*****
A Heuristic Method to improve the efficiency of the kmeans Clustering Alg
*****

Enter desired no of clusters k: 5
the value of k is:5
Obtained cluster centroid are:
new 0 th cluster centroid is:
6.173180401171683
1.843168461296551
4.707579256282695
1.5840710246071621

new 1 th cluster centroid is:
6.127011119211989
0.512700068040763
1.5047582505215784
0.47454205120190715

new 2 th cluster centroid is:
6.44281943185274
0.81461704653775
1.499047138471519
0.2830486194177253

new 3 th cluster centroid is:
6.469791254317411
0.557645917881807
1.1901640021984184
0.2086834114732774

new 4 th cluster centroid is:
6.462007846242422
0.5780487197133668
1.2529671548740468
1.1955487821707004

```

Figure 8:

```

Implementation of K-Means Clustering Algorithm Using Java
c1,c2,...,ck} // set of k centroids
Steps:
1. Compute the distance of each data point di
(1<=i<=n) to all the centroids cf (1<=j<=k) as
d(di, cj);
2. For each data-point di, find the closest centroid
cf and assign di to cluster j.
3. Set CllusterId[i]=j;                                     //
                                                             j:
                                                             Id
                                                             of
                                                             the
                                                             clos-
                                                             est
2014cluster 4. Set Nearest __Dist[i] = d(di, cj) ;
Oct 5. For each cluster j ( 1<= j<= k) troids: 6. Repeat 7. For each data-point di.
8. Compute its distance is less than or equal to
64 the present nearest cluster:
9. If this distance is < or = to the present nearest
distance , the data-point satays in the cluster:
Else
10. For every centroid cj (1 <=j<= k) Compute the
distance d(di, cj); Endfor;
11. Assign the data-point di to the cluster with the
nearest centroid cj
12. Set Cluster ID[i];
13. Set Nearest__Dist[i] = d(di, cf) ;
14. Endfor;
15. For each cluster j ( 1<= j<=k), recalculate the
centroids:
16. Until the convergence criterin is met.

```

Figure 9:

enhanced method are described below as Algorithm 3 and Algorithm 4.

The first step in Phase 2 is to determine the distance between each data-point and the initial centroids of all the clusters. The data-points are then assigned to the clusters having the closest centroids. This results in an initial grouping of the data-points. For each data-point, the cluster to which it is assigned (ClusterId) and its distance from the centroid of the nearest cluster (Nearest_Dist) are noted. Inclusion of data-points in various clusters may lead to a change in the values of the cluster centroids. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. Up to this step, the procedure is almost similar to the original k-means algorithm except that the initial centroids are computed systematically. The next stage is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the datapoints may get redistributed to different clusters.

The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data-point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. This results in the saving of time required to compute the distances to k-1 cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the data-point than its previous centroid, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. Identify the new nearest cluster and record the new value of the nearest distance. The loop is repeated until no more data-points cross cluster boundaries, which indicates the convergence criterion. The heuristic method described above results in significant reduction in the number of computations and thus improves the efficiency.

.1 V. OUTPUT

The modified algorithm is applied to multidimensional gene expression data taken from the UCI(university of california irvine) repository [??]. The input data are the iris data[10], the breast cancer data [11], the e coli data[9], the echo cardiogram data[12], the yeast data[13] and the height-weight data obtained from the web site of disabled-world [18]. The results are compared with that of the original k-means algorithm as well as Enhanced k-means algorithm. Tables 6.1 to 6.6 show the performance comparison of the three algorithms. Figures 6.1 to 6.6 illustrate that the modified algorithm provide better accuracy and efficiency compared to the k-means and enhanced k-means methods.