CHAPTER 5

A FAST QUAD-TREE BASED TWO DIMENSIONAL HIERARCHICAL CLUSTERING

5.1 INTRODUCTION

The DNA microarray has emerged as the latest breakthrough in molecular biology; it helps researchers to monitor the genome-wide expression systematically (Xue et al. 2004). The microarrays are used to study the gene expression profiles in biological samples (Fadiel and Naftolin 2003). A Microarray or microchip is a chip made with glass or other solid material, with an array of tiny DNA spots placed on it. Each of the spots contains fragments of DNA or RNA molecule whose sequence is predefined, and corresponds to portions of a particular gene (Liang and Kachalo 2002). Microarrays do not provide full information about genes; they denote genes indirectly through their expressions. Also, the expressions obtained can be inaccurate depending on the applied microarray technology (Gruzdz et al. 2006).

Microarrays have emerged as the standard for the simultaneous evaluation of the expression level of thousands of genes (Urska Cvek et al. 2009). The clustering technique plays a significant role through the discovery of a set of objects with identical functions from huge quantities of data (Kim and Choi 2005). A good clustering algorithm should be able to identify genes that have similar expression profiles, which include time-shifted or inverted ones, and provide the information known as phase (Wu et al. 2004). There are many popular techniques for clustering gene expression data with
different functional roles in genes, that play an important role in the biological process (Matthew and Praveen 2006). A major area of concern in the clustering analysis of gene expression data is the sensitiveness and vulnerability of the results with noise and over-fitting respectively, due to their excessive dependence on limited biological and medical information (Young et al 2005). Most of the clustering algorithms are distance-based focussed, whereas some are based on Hierarchical clustering, K-means clustering and Self Organizing Map (Zhaohui Qin 2006; Lee et al 2007).

In the microarray data analysis, hierarchical clustering techniques are mainly used. They combine all the data points closest to each other into a single set in the feature space until eventually all data points are merged into a single group (Wang et al 2003). At present, hierarchical clustering is the most often used method for discovering the data as a group (Kalocsai and Shams 1999; Van der Laan and Pollard 2003). The main objective of hierarchical clustering is to obtain the best clustering, that will signify a set of patterns in the background of a given distance metric. This method is special for biologists, because it will permit users to visualize global expression patterns in the DNA microarray data through graphic representation, by using hierarchical clustering (Jin Hanwan Do and Dong-Kug Choi 2007).

The hierarchical clustering method is classified into two distinct types: agglomerative and divisive (Trepalin and Yarkov 2008), (Jin Hanwan Do and Dong-Kug Choi 2007). This method is based on a similarity or distance measure of the data like correlation, Euclidean, squared Euclidean, or Manhattan distance, etc (Kalocsai and Shams 1999). Normally, clusters are constructed using a hierarchical tree which is created after calculating the distance between pairs of objects in the correlation matrix (Tuncbag et al 2005). A Quad Tree (QT) is a tree data structure in which each internal node has upto four children. The Quad tree and its different
derivatives are being considered as the backbone for the storage, retrieval and analysis of spatial data. An efficient searching technique depends mainly on the height of the tree; an arbitrary insertion of the point features makes the tree unbalanced, and increases the time of searching.

From the review it is obvious that clustering algorithms used in previous research have partitioned the data, where each gene belongs to only one cluster. The K-means algorithm, hierarchical clustering algorithm, biclustering algorithm, fuzzy k-means algorithm and SOM are some of the clustering algorithms that can create only one cluster for a gene. These methods have some disadvantages, while working with microarray gene expression data that gives high complexity in the biological process. The nature of proteins and their interaction is the major reason for this. The genes that generate proteins are expected to express in more than one group of genes, because proteins generally perform unique biological roles by interacting with other groups of proteins. This explains the inclusion of a gene in more than one cluster of the microarray gene expression data. The QT focuses on the clustering gene expression data with the Expectation Maximization (EM) algorithm, which estimates the incomplete data samples. It also validates the cluster and measures the representation of the objects in the clusters. The algorithm is compared with the k-means algorithm for performance validation (Leela Rani and Rajalakshmi 2012).

5.2 QUAD TREE

A Quad Tree (QT) shown in Figure 5.1 is a tree like data structure, where each inner node has exactly four children. QT and its different derivatives are mainly used for the purpose of storage, recovery and analysis of spatial data (Chakraborty et al 2011). QT is used to index the data records in a spatial database, in accordance with the spatial location. Most of the business and government agencies may desire to query the database using
spatial constraints; in such cases, the QT facilitates the storage and querying of spatial data. QT can also be used to index data records in multiple spatial databases. The QT index values are used to perform searching across the different databases by extracting the associations, patterns, clusters, outliers or nearest neighbours, etc.

Figure 5.1 Sample structure of the quad tree

The point QT is constructed step by step, by adding the data points one by one. Initially, a point search is performed to add a point. In the quad tree, if there is no point related to the target point, the target point is inserted into the leaf node where the search has to be ended.

The QTs are mainly used to perform searching in various applications and the performance depends on the distribution of data in the domain. The height and shape of the point quad tree relies greatly on the insertion sequence. Until date, no significant effort has been taken to overcome this height balancing problem as well as to make a more efficient search operation on the quad tree (Lee et al 2007).
5.3 A FAST TWO DIMENSIONAL HIERARCHICAL CLUSTERING

The recent arrival of microarray technologies has permitted biologists to monitor simultaneously the behavior of genes, which produces a large number of complex data. A large amount of gene expression data has been generated continuously from the microarray experiments. It has been proved that such gene expression data that contain vital information, are very useful in medical diagnosis, therapy, and drug design. The objective of this technique is to examine such data to obtain the essential information. A fast two dimensional hierarchical clustering developed, is shown in Figure 5.2.

An innovative high-speed two dimensional hierarchical clustering technique, to handle the micro array genes that exist in more than one cluster, is shown in Figure 5.2. Genes are effectively expressed by this technique, which avoids biological complexities. This technique is evaluated on clustering the ground truth data of the cancer classes, namely, AML and ALL using precision, recall and F-measure. Subsequently, these values are compared with the precision, recall and F_measure values of hierarchical clustering without the quad tree.

This technique includes a method called self clustering of each gene type of clustering elements in the vertical dimension, and hierarchical clustering of the gene type in the horizontal dimension. The clustering is performed at different levels, and the gene values in every gene type are clustered internally into four clusters, and every gene type is again clustered horizontally for analysis. From the resultant clusters, the best clusters are found by using the fitness evaluation test. Subsequently, the closest index of all the best clusters is calculated, and used to obtain the next set of clustering elements from the database. This process is repeated ‘r’ times until the optimum cluster is found.
Figure 5.2 Quad-tree based Two Dimensional hierarchical Clustering
In this hierarchical clustering process, the creation of the closest pair data structure at each level is an important time factor, that provides the processing time of clustering. This quad tree based data structure is used for identifying the closest pair elements which reduces the processing time successfully and offers a better analysis of the gene expression data.

In order to find the closest pair, the developed two dimensional hierarchical clustering with the quad tree system performs faster than the conventional hierarchical clustering without the quad tree.

Cluster analysis has played a key role in analyzing gene expression data. The main aim of clustering is to partition a set of objects into clusters, so that objects in a group are more analogous to one another, than patterns in diverse clusters. Several clustering algorithms have been used to find the co-expressed genes, but the processing time of these techniques is very high. Hence, our aim is to reduce the processing time of the clustering analysis. In order to solve the aforesaid problem, in this research, a fast, semi-supervised two dimensional quad tree based hierarchical clustering technique is developed. Here, the clustering elements are selected from the microarray gene expression database by means of the index matrix, and these elements are clustered using the quad-tree based two dimensional clustering technique. After the clustering process, the best ‘K’, where ‘K’ is the output clusters is identified using the fitness evaluation. Then, the closest index of all the best ‘K’ clusters is calculated, and this is used to obtain the next set of clustering elements from the database. This process is repeated until an optimum cluster is found.

Using the index ‘Y’ the gene expression data ‘x’ is selected randomly from the database \( X_{ab} \), which contains ‘a’ gene representation of ‘b’ clustering elements, given in Equation 5.1.
\[x = \left\{ x_{ij} \mid x_{ij} \in X \right\} \quad 1 < i \leq q; \quad 1 < j \leq Q \quad (5.1)\]

where ‘\(x_{ij}\)’ is the distance between gene ‘\(i\)’ and gene ‘\(j\)’

‘\(x\)’ is the randomly selected dataset

‘\(q\)’ is the number of gene samples

‘\(Q\)’ is the number of genes

‘\(X\)’ is the microarray gene expression database

The gene data ‘\(x\)’ can be represented as the form given in Equation 4.2, where \(q\) and \(Q\) are the no of rows and columns in the dataset.

\[
X_j = \begin{bmatrix}
x_{(1, 1)} & x_{(1, 2)} & \ldots & x_{(1, Q)} \\
x_{(2, 1)} & x_{(2, 2)} & \ldots & x_{(2, Q)} \\
\vdots & \vdots & \ddots & \vdots \\
x_{(Q, 1)} & x_{(Q, 2)} & \ldots & x_{(Q, Q)}
\end{bmatrix} \quad (5.2)
\]

\[Y = \left\{ Y_{ij} \mid Y < A \iff i, j \right\} \quad 1 < i \leq m; \quad 1 < j \leq n \quad (5.3)\]

where ‘\(m\)’ and ‘\(m\)’ are the number of rows and columns in the

‘\(Y_{ij}\)’ dataset.

‘\(Y\)’ is the index matrix.
Each value in the ‘Y’, which represents the row index value of the Database Xab, must be unique and less than the maximum number of gene representations ‘A’ in the database Xab.

5.4 QUAD TREE BASED TWO DIMENSIONAL CLUSTERING

In the cluster analysis, groups of similar objects are identified by maximizing inter-group similarity and minimizing intra-group similarity. Cluster analysis partitions the data into significant or useful groups (clusters). It also plays a key role in evaluating the gene expression data. Clustering techniques are very useful for understanding the basic biological processes. There are many popular and robust techniques for clustering of the gene expression data, with the aim of describing the different functional roles of genes, that play an important role in the biological process. The cluster analysis also describes the genes in a cluster with similar expression (co-expressed genes), which serve a similar functional role in a process.

In this analysis, a quad-tree based technique is used to expedite the two dimensional hierarchical clustering. Figure 5.3 illustrates the overall process of a two dimension clustering data technique in a single iteration. The clustering elements are selected from the gene expression database, using the random values of the index matrix. Each ‘q’ type gene representations of the ‘Q’ elements are clustered into four clusters, by means of the quad tree based vertical dimensional clustering, and then clustered into two clusters, by using the quad tree based horizontal clustering.
5.5 GENES CLUSTERING VERTICAL DIMENSION

The pseudo code for the quad tree based vertical dimensional clustering techniques shown in Figure 5.4 explains how the genes are clustered by using Quad-tree based vertical dimensional clustering. Initially, the value of the root node of the quad tree is set to ‘Null’ and four values are selected randomly from the first type gene expression elements, and inserted as a child of the Root node. In the function Rchild(n, Root, P), the ‘n’ is the randomly selected data to be inserted, ‘Root’ is the parent node, and ‘P’ is the position of the child in the root node. In the clustering process ‘Nchild’ is the number of the child node, ‘L’ is the depth of the quad tree, and ‘l’ is the level of the tree. Then, the element to be inserted is selected from the first type gene expression data, and to find its insertion position in the quad tree, its minimum Euclidean distance from the first level nodes, that is, children of the
root node, is calculated, and subsequently, the node is inserted as the child node of the closest node having the minimum distance value. The validity of the parent node is then evaluated while inserting a child node, to check whether the weight of the parent node is less than four or not. If the above condition is satisfied, then the element is inserted as a child node, or it will find the closest node at the subsequent levels of the quad tree by calculating the minimum distance. While inserting a node, if the parent node value is less than the child node to be inserted, then both the values are swapped. Thus, every element in the first type gene expression values is inserted into the quad tree, and hence, clustered into four groups. Similarly, all the gene type elements are clustered into four groups, using the quad tree structure.

1. Set Root=NULL
2. c=1;
3. For i = 1 to Q
4. For j = 1 to q
5. For P = 1 to 4
6. Set n = Random (x )
7. Insert Rchild (n, Root, P)
8. End For
9. While c < L
10. l=1;
11. For k = 1 to 4
12. Set \( E_k = \sqrt{(x_{ji} - R_{lk})^2} \)
13. Set s = findmin (E_k)
14. End For
15. If Nchild (s) < 4
16. Insert child (x_{ji}, R_{lk} (s), P, l);
17. Exit While
18. Else
19. Movedown (l, R_{lk} (s));
20. End If
21. End While
22. End for
23. End for

Figure 5.4 Pseudo code for the quad tree based vertical dimensional clustering
5.6 GENES CLUSTERING HORIZONTAL DIMENSION

Every gene value is clustered inner wise into four clusters according to their distance, and then these clusters are again clustered, using the horizontal dimension for analyzing the gene values. The distance is calculated based on the following equation.

\[ E_k = \sqrt{(x_{ji} - R_{lk})^2} \]  

where \( E_k \) is the Euclidean distance

\( x_{ji} \) is the clustering elements in ‘x’

\( R_{lk} \) is the clustered element at the \( k^{th} \) level of the quad tree.

The distance between the two elements \( x_{ji} \) and \( R_{lk} \) is the root of, sum of, square of, and deviation among \( x_{ji} \) and \( R_{lk} \). The first level of horizontal wise hierarchical clustering starts with the selection of two elements having the greatest distance, and inserted into the root node. Using the Euclidean distance, the closest pair node of the next gene representation element at the \( k^{th} \) level is found, and inserted into the corresponding node, if the weight of the node is less than four. Otherwise, it will find out the closest pair element at the subsequent levels and insert the node. At the final level, every gene type is clustered, based on the quad tree.

5.6.1 Fitness Evaluation

From the resultant output cluster ‘C’, the fitness of C is calculated as follows:

\[ C = \frac{1}{0.1 + \sum w(C)} \]  

(5.5)
where $C[i]$ is the clustering element in the resultant cluster

$w(C) = \begin{cases} 
1 & \text{if } C[i] = R^{\text{def}} \\
0 & \text{otherwise}
\end{cases}$

‘$w$’ is the weight of each clustered element

$R^{\text{def}}$ is the defined cluster dataset

If an element in the consequential cluster is in the defined cluster $R^{\text{def}}$, at that moment the weight of the gene cluster is assigned the value 1, and 0 otherwise. The two dimensional clustering progression and fitness evaluation are processed, for every row of the index ‘$Y$’ and $C_i$ is the resultant cluster. From the resultant cluster set ‘$K$’, the best output clusters having the highest fitness value are selected and the contiguous index of all clusters is computed, which is then used to get the next set of clustering elements from the database. The clustering is performed repeatedly until an optimum cluster is produced by using this cluster, and then the gene is analyzed.

5.7 EXPERIMENTAL RESULTS

The two dimensional hierarchical clustering technique is implemented in the working platform of Matrix Laboratory 7.11, and it is evaluated using the microarray gene expression data of human acute leukemia. The two datasets of leukemia for training and testing are obtained from http://www.broadinstitute.org/cancer/software/genepattern/datasets/ and the performance of the developed technique on clustering the ground truth data of the cancer classes, namely, AML and ALL is demonstrated.
The two training leukemia datasets are again partitioned, and turned to four sets, datasets 1, 2, 3 and 4, each having N values of 20, 18, 18, 17 respectively. This high dimensional training dataset is subjected to clustering for analyzing the occurrence of the microarray genes in more than one cluster. In this clustering technique, an adaptive approach is used to represent the number of clusters that are dynamically generated from the microarray gene expression dataset.

The two dimensional hierarchical clustering technique is a multi-stage clustering technique, which performs clustering at diverse levels. Initially, the gene values in every gene type are clustered vertically into four groups, and then these gene types are clustered horizontally for further analysis. In the inner gene clustering, initially the root node is created, and then the four clustering elements that are selected randomly from the first type gene expression data, are inserted into the root node. Then, the insertion position of the next element to be inserted is found, by calculating its minimum Euclidean distance from the first level nodes, and the element is inserted into the corresponding node, as children whose weight is less than four. If the weight of the node is greater than or equal to four, then the closest node at the subsequent levels of the quad tree is found, by calculating the minimum distance. This process is repeated until all the elements get clustered. Thus all the gene representation values are clustered into four clusters. The clustering output of the ALL/AML datasets 1, 2, 3 and 4 are shown in the following Figure 5.5. The x-axis represents the number of clusters, while the y-axis represents the hierarchical level.
Figure 5.5 (a) Clustering outputs of the ALL/AML dataset_1

Figure 5.5 (a) shows the clustering output of the ALL/AML dataset_1. The result of the dataset_1 having a N value of 20 indicates, that both AML and ALL are crossing each other; that means AML has ALL and vice versa.

Figure 5.5 (b) Clustering outputs of the ALL/AML dataset_2
Figure 5.5 (b) shows the clustering output of the ALL/AML dataset_2. The result of the dataset_2 having a N value of 18 indicates, that both AML and ALL are crossing each other; that means AML has ALL and vice versa.

Figure 5.5 (c) Clustering outputs of the ALL/AML dataset_3

Figure 5.5 (c) shows the clustering output of the ALL/AML dataset_3. The result of the dataset_3 having a N value 18 indicates, that both AML and ALL are crossing each other; that means AML has ALL and vice versa.
Figure 5.5 (d) Clustering outputs of the ALL/AML dataset_4

Figure 5.5 (d) shows the clustering output of the ALL/AML dataset_4. The result of the dataset_4 having a N value of 17 indicates, that both AML and ALL are crossing each other; that means AML has ALL and vice versa. The AML and ALL data in the graph are resided in more than one cluster; so this technique proves that the developed algorithm is more efficient than the existing ones.

5.8 COMPARISON OF THE EXISTING ALGORITHMS

The performance of the SOM is less, whenever the number of clusters goes up. The hierarchical and SOM clustering algorithms provide better results for small data sets. The expectation of maximization and K-Means algorithms are very suitable for large datasets. These algorithms may suffer a lot over the noise in the data sets (Abbas 2008). The algorithms provide the clustered output, but lack in several aspects.
Table 5.1 Comparison of the Algorithms

<table>
<thead>
<tr>
<th>No of Clusters(K)</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SOM</td>
</tr>
<tr>
<td>8</td>
<td>59</td>
</tr>
<tr>
<td>16</td>
<td>67</td>
</tr>
<tr>
<td>32</td>
<td>78</td>
</tr>
<tr>
<td>64</td>
<td>85</td>
</tr>
</tbody>
</table>

Each algorithm is analyzed, based on the input data sets as well as the number of clusters (i.e, 8, 16, 32 and 64), taken for the observation as shown in Table 5.1. The hierarchical clustering algorithm has a performance higher than the other algorithms, but it is inefficient for a large data set. So, this two dimensional quad tree based hierarchical clustering has been developed.

![Performance comparison of the existing algorithms](image)

Figure 5.6 Performance comparison of the existing algorithms
A comparison of algorithms in terms of performance, is shown in Figure 5.6. In the following graph, X-axis represents the various clustering algorithms. The Y-axis represents the performance of these algorithms with many parameters. In this, the two dimensional quad tree based clustering technique is developed for grouping a gene expression data, which gives a better performance than the other clustering algorithms. It also finds the genes in more than one cluster, using this technique for achieving high accuracy.

5.9 PERFORMANCE EVALUATION

The performance of the two dimensional hierarchical data clustering technique is evaluated, on clustering the ground truth data of the cancer classes, namely, ALL and AML, using precision, recall and F-measure, and subsequently, these values are compared with the precision, recall and F-measure values of conventional hierarchical clustering. Precision, recall and F-measure were used as described by (Larsen and Aone 1999; Karypis et al 1999) to evaluate the performance of the proposed incremental text clustering approach.

\[
\text{Precision} (i, j) = \frac{M_{ij}}{M_j}
\]  

\[
\text{Recall} (i, j) = \frac{M_{ij}}{M_i}
\]  

\[
\text{F-measure} (i, j) = \frac{2 \times \text{Recall} (i, j) \times \text{Precision} (i, j)}{\text{Recall} (i, j) + \text{Precision} (i, j)}
\]

where ‘\(M_{ij}\)’ is the number of members of genes \(i\) in cluster \(j\)

‘\(M_j\)’ is the number of members of cluster \(j\)

‘\(M_i\)’ is the number of members of genes \(i\).
The precision and recall values of the conventional method of hierarchical clustering are compared with those of the quad tree based hierarchical technique.

![Figure 5.7 Precision, recall evaluation of the quad tree technique](image)

**Figure 5.7 Precision, recall evaluation of the quad tree technique**

Figure 5.7 describes the precision and recall evaluation of the quad tree based hierarchical technique. The X-axis represents the number of clusters, whereas the Y-axis represents the evaluation metrics, the values of precision and recall.
Figure 5.8 Precision, recall evaluation of the conventional technique

Figure 5.8 describes the conventional method of hierarchical clustering. The precision and recall values of the conventional method of hierarchical clustering are less, when compared to those of the quad tree based hierarchical technique. For example, in the conventional method, cluster C1 has the precision and recall values of 0.5 and 0.2, respectively. In the quad tree based hierarchical technique, cluster C1 has the precision and recall values of 0.8667 and 0.9286, respectively. All the clusters have a higher value than the conventional method, indicating that the quad tree based hierarchical clustering technique is faster.
Table 5.2 Precision, Recall and F-measure values of the clusters_ quad tree technique

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cluster</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset_1</td>
<td>C1</td>
<td>0.8667</td>
<td>0.9286</td>
<td>0.8966</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>0.8000</td>
<td>0.6667</td>
<td>0.7273</td>
</tr>
<tr>
<td>Dataset_2</td>
<td>C3</td>
<td>0.9091</td>
<td>0.7692</td>
<td>0.8333</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>0.5714</td>
<td>0.8000</td>
<td>0.6667</td>
</tr>
<tr>
<td>Dataset_3</td>
<td>C5</td>
<td>0.9091</td>
<td>0.9091</td>
<td>0.9091</td>
</tr>
<tr>
<td></td>
<td>C6</td>
<td>0.8571</td>
<td>0.8571</td>
<td>0.8571</td>
</tr>
<tr>
<td>Dataset_4</td>
<td>C7</td>
<td>0.8000</td>
<td>0.8000</td>
<td>0.8000</td>
</tr>
<tr>
<td></td>
<td>C8</td>
<td>0.7143</td>
<td>0.7143</td>
<td>0.7143</td>
</tr>
</tbody>
</table>

The precision and recall values of the clusters obtained by the quad tree based two dimensional hierarchical clustering technique are given in Table 5.2.

Table 5.3 Precision, Recall and F-measure values of the clusters_ conventional hierarchical clustering

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cluster</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset_1</td>
<td>C1</td>
<td>0.5000</td>
<td>0.2143</td>
<td>0.3000</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>0.2143</td>
<td>0.5000</td>
<td>0.3000</td>
</tr>
<tr>
<td>Dataset_2</td>
<td>C3</td>
<td>0.7000</td>
<td>0.5385</td>
<td>0.6087</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>0.2500</td>
<td>0.4000</td>
<td>0.3077</td>
</tr>
<tr>
<td>Dataset_3</td>
<td>C5</td>
<td>0.8182</td>
<td>0.8182</td>
<td>0.8182</td>
</tr>
<tr>
<td></td>
<td>C6</td>
<td>0.7143</td>
<td>0.7143</td>
<td>0.7143</td>
</tr>
<tr>
<td>Dataset_4</td>
<td>C7</td>
<td>0.5000</td>
<td>0.3000</td>
<td>0.3750</td>
</tr>
<tr>
<td></td>
<td>C8</td>
<td>0.3636</td>
<td>0.5714</td>
<td>0.4444</td>
</tr>
</tbody>
</table>
The precision and recall values of the clusters obtained by the conventional hierarchical clustering technique are given in Table 5.3.

The performance of the two dimensional hierarchical data clustering approach is also evaluated by computing its processing time. From Figures 5.9 and 5.10, it is obvious that the quad tree technique has performed clustering faster than the conventional hierarchical clustering.

Figure 5.9 Time comparison graph for conventional clustering and quad tree technique

Figure 5.9 shows the comparison between the conventional and QT based hierarchical clustering techniques. It is obvious that the quad tree technique has performed clustering faster than the conventional hierarchical clustering.
Figure 5.10 Time comparison graph for two dimensional hierarchical clustering without the quad tree and with the quad tree

Figure 5.10 illustrates the comparison between the two dimensional hierarchical clustering and the fast two dimensional hierarchical clustering, based on the QT. The performance of the two dimensional hierarchical clustering without the Quad tree and with the Quad tree, is also evaluated by comparing its processing time. In dataset_1 the processing time of the two dimensional hierarchical clustering without the Quad tree and with the Quad tree is 2, 0.2; likewise, for datasets 2, 3 and 4 it is (1.2, 0.1), (1.3, 0.1), (1.2, 0.1); the processing time difference is 1.8, 1.1, 1.2 and 1.1 respectively.

5.10 SUMMARY OF THIS WORK

- A fast two dimensional hierarchical clustering technique has been developed, to deal with the microarray genes that are available in more than one cluster.
• A set of clustering elements is selected randomly from the microarray gene expression data, by using the index; then they are clustered, to evaluate the fitness for selecting the best ‘k’ clusters from the various clusters.

• The next set of clustering elements is selected by finding the closest two cluster indices among the best ‘k’ clusters. These clusters are combined until the best clusters are found. The resultant genes are expressed in an efficient manner, by eliminating the biological complexities during the clustering process.

• In the existing techniques, each and every gene expression database was analyzed for finding the closest pair, whereas the quad-tree based data structure uses some specific set of gene expression database.

• The average processing time of the two dimensional hierarchical clustering without the quad tree, is 1.3 times longer than that of the two dimensional quad tree clustering method.

• The two dimensional quad tree technique is faster when compared to the existing clustering techniques, in terms of performance.

• The experimental results on real life datasets have proved, that the two dimensional quad tree technique was more robust and efficient than the traditional hierarchical clustering.

• The two dimensional quad tree algorithm can be implemented in various application areas, with better performance in future.