CHAPTER 4

A SEMI-SUPERVISED HIERARCHICAL APPROACH

4.1 INTRODUCTION

Microarrays based global gene expression profiling is developed as a vital technology, for comprehending the basic biology of the gene function, its development, finding out new classes of diseases like cancer, and interpreting their molecular pharmacology (Liang and Kachalo 2002). Microarray technologies are widely used all over the world to study the gene expression of several tissues (Khlopova et al 2009). The technique of microarrays, which offers high output capability of simultaneously examining the RNA expression of the entire genomes, began with the simultaneous gene expression analysis of 45 genes in one experiment (Chintanu et al 2010). Instead of providing the full information about genes, microarrays indirectly represent genes through their expressions. Furthermore, these expressions can be inaccurate, based on the employed microarray technology (Gruzdz et al 2006).

Microarrays emerged as the standard for simultaneous evaluation of the expression levels of thousands of genes (Cvek et al 2009). Clustering techniques play a significant role in discovering sets of objects with identical functions from huge quantities of data (Kim and Choi 2006) Grouping genes with identical biological functions or categorizing samples with identical gene expression profiles is the usual objective of performing clustering on microarray gene expression data (Wu et al 2004). The major concerns in the
clustering analysis of gene expression data are the sensitiveness and vulnerability of the results of the clustering analysis to noise and over-fitting, respectively, due to their excessive dependence on limited biological and medical information. Several clustering algorithms have been used for the analysis of microarray gene expression data (Chen et al 2002). Most of the clustering algorithms available these days are distance-based; some examples are hierarchical clustering, K-means clustering and Self Organizing Map (Qin 2006; Lee et al 2007).

Hierarchical clustering can be classified based on the employed comparative method, as agglomerative and divisive (Trepalin and Yarkov 2008; Hwan Do and Choi 2007). Clusters are constructed by any of the above-mentioned methods to form the hierarchical tree, which is created by calculating the distance between pairs of objects in the correlation matrix (Tuncbag et al 2005).

4.2 MICRO ARRAY GENE EXPRESSION DATA CLUSTERING THROUGH TWO DIMENSIONAL HIERARCHICAL CLUSTERING

In this research, a semi supervised two dimensional hierarchical clustering technique has been used, which includes the self clustering of each gene type in a vertical dimension, and bottom up hierarchical clustering in the horizontal dimension, with overlapping clustering for merging. The conceptual view of the semi supervised hierarchical clustering technique is illustrated in Figure 4.1. Sets of clustering elements are selected randomly from the microarray gene expression database using the index, and are clustered repeatedly using the two dimensional clustering technique. From the resultant clusters, the best ‘k’ output clusters are found, using the fitness evaluation. Subsequently, the closest index of all the best ‘k’ clusters is
calculated and used to fetch the next set of clustering elements from the database. This process is repeated ‘r’ times.

Figure 4.1 illustrates the semi-supervised two dimensional clustering technique. The datasets for training and testing are obtained and the performance of this approach in clustering ground truth datasets, namely, human acute leukemia, adenocarcinoma and lymphoma cancer cells is demonstrated. The training and testing datasets are divided into different datasets, termed as initialization of cluster sets each having N values. The high dimensional training datasets are clustered for analyzing the presence of the microarray genes in more than one cluster.

Using the similarity measure, the closest self clustered pair elements are found out. The data sets are classified, based on the distance measures followed by the merging process which involves the overlapping of the clusters to produce a new large cluster. Then, the next level of hierarchical clustering is started with (N-1) elements and at the last level, a single cluster is obtained. Each level of the semi-supervised hierarchical clustering identifies similar gene types from the training datasets, and is subjected to clustering. The process is repeated until a predetermined clustering level is reached. From the resultant cluster set, the best clusters having the highest fitness value are chosen, and the closest index of all the best clusters is calculated, which is then used to fetch the next set of clustering elements from the database. The clustering is processed continuously until the optimum cluster is found, and using this, the presence of a gene in more than one cluster of the microarray gene expression data is analyzed.
Figure 4.1 Semi-supervised two dimensional hierarchical clustering
The flow diagram of the semi-supervised two dimensional hierarchical clustering that depicts the above process, is shown in Figure 4.2.

Figure 4.2 Flow diagram of the two dimensional hierarchical clustering
A sample gene expression database of an acute leukemia training dataset shown in Figure 4.3. This consists of the training data set and the test data set. The training data has 38 samples with 27 ALL and 11 AML. The test data set has 35 samples with 20 AML and 15 ALL. Each sample has 7129 gene expression values from 6817 human genes.

Using the index ‘I’ the gene expression data ‘d’ is selected randomly from the database $D_{MN}$, which contains ‘M’ gene representation of ‘N’ clustering elements, which is given in Equation 4.1.

$$d = \{d_{ij} \mid d_{ij} \in D\} \quad 1 < i \leq n; 1 < j \leq N$$

(4.1)

where ‘$d_{ij}$’ is the distance between gene ‘i’ and gene ‘j’

‘d’ is the randomly selected dataset

‘n’ is the number of gene samples

‘N’ is the number of genes

‘D’ is the microarray gene expression database

The gene data ‘d’ can be represented as the form given in the Equation 4.2, where ‘n’ and ‘N’ are the number of rows and columns in the dataset, respectively.

$$d_{ij} = \begin{pmatrix}
  d_{(1,1)} & d_{(1,2)} & \cdots & d_{(1,N)} \\
  d_{(2,1)} & d_{(2,2)} & \cdots & d_{(2,N)} \\
  \cdot & \cdot & \cdots & \cdot \\
  \cdot & \cdot & \cdots & \cdot \\
  \cdot & \cdot & \cdots & \cdot \\
  d_{(n,1)} & d_{(n,2)} & \cdots & d_{(n,N)}
\end{pmatrix}$$

(4.2)
\[ I = \{ I_{ij} \mid I < M \ \forall \ i, j \} 1 < i \leq l; 1 < j \leq k \] (4.3)

where ‘l’ and ‘k’ is the number of rows and columns in the ‘I_{ij}’ dataset.

\( I \) is the index matrix

Each value in the ‘I’ in Equation 4.3, represents the row index value of Database \( D_{MN} \), which must be unique and less than the maximum number of gene representations ‘M’ in the database \( D_{MN} \).

<table>
<thead>
<tr>
<th>Y</th>
<th>Z</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>AE</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>7112</td>
<td>44</td>
<td>-93</td>
<td>-216</td>
<td>157</td>
<td>151</td>
<td>417</td>
<td>222</td>
</tr>
<tr>
<td>7113</td>
<td>177</td>
<td>179</td>
<td>308</td>
<td>339</td>
<td>195</td>
<td>363</td>
<td>211</td>
</tr>
<tr>
<td>7114</td>
<td>195</td>
<td>84</td>
<td>334</td>
<td>198</td>
<td>173</td>
<td>203</td>
<td>99</td>
</tr>
<tr>
<td>7115</td>
<td>192</td>
<td>198</td>
<td>287</td>
<td>323</td>
<td>55</td>
<td>380</td>
<td>261</td>
</tr>
<tr>
<td>7116</td>
<td>173</td>
<td>282</td>
<td>233</td>
<td>498</td>
<td>415</td>
<td>333</td>
<td>249</td>
</tr>
<tr>
<td>7117</td>
<td>-244</td>
<td>-217</td>
<td>-196</td>
<td>-294</td>
<td>-413</td>
<td>-267</td>
<td>-429</td>
</tr>
<tr>
<td>7118</td>
<td>-1396</td>
<td>-1687</td>
<td>-1194</td>
<td>-1928</td>
<td>-1959</td>
<td>-2162</td>
<td>-3452</td>
</tr>
<tr>
<td>7119</td>
<td>1972</td>
<td>732</td>
<td>1397</td>
<td>454</td>
<td>391</td>
<td>412</td>
<td>393</td>
</tr>
<tr>
<td>7120</td>
<td>185</td>
<td>267</td>
<td>130</td>
<td>260</td>
<td>93</td>
<td>235</td>
<td>146</td>
</tr>
<tr>
<td>7121</td>
<td>642</td>
<td>440</td>
<td>1066</td>
<td>1010</td>
<td>336</td>
<td>1655</td>
<td>486</td>
</tr>
<tr>
<td>7122</td>
<td>32</td>
<td>52</td>
<td>141</td>
<td>-66</td>
<td>-45</td>
<td>67</td>
<td>-32</td>
</tr>
<tr>
<td>7123</td>
<td>137</td>
<td>229</td>
<td>278</td>
<td>139</td>
<td>170</td>
<td>497</td>
<td>334</td>
</tr>
<tr>
<td>7124</td>
<td>-61</td>
<td>-4</td>
<td>-22</td>
<td>-67</td>
<td>12</td>
<td>-88</td>
<td>35</td>
</tr>
<tr>
<td>7125</td>
<td>672</td>
<td>431</td>
<td>622</td>
<td>835</td>
<td>752</td>
<td>1295</td>
<td>1733</td>
</tr>
<tr>
<td>7126</td>
<td>208</td>
<td>269</td>
<td>338</td>
<td>557</td>
<td>295</td>
<td>343</td>
<td>304</td>
</tr>
<tr>
<td>7127</td>
<td>25</td>
<td>8</td>
<td>9</td>
<td>-12</td>
<td>28</td>
<td>26</td>
<td>12</td>
</tr>
<tr>
<td>7128</td>
<td>116</td>
<td>153</td>
<td>244</td>
<td>355</td>
<td>1558</td>
<td>247</td>
<td>3192</td>
</tr>
<tr>
<td>7129</td>
<td>-62</td>
<td>-22</td>
<td>36</td>
<td>-45</td>
<td>-67</td>
<td>23</td>
<td>-33</td>
</tr>
</tbody>
</table>

**Figure 4.3 Acute leukemia dataset**

**4.2.1 Two Dimensional Data Clustering**

The randomly selected data from the whole gene expression database is ‘d’ having ‘N’ clustering elements, included in the two dimensional data clustering and clustered into a single cluster. The basic algorithm for the two-dimensional clustering approach is as follows:
Algorithm Two dimensional clustering

**Input:** Input the ‘N’ clustering elements

**Step 1** Start the first level of hierarchical clustering with ‘N’ clustering elements.

**Step 2** Find the closest pair elements using the Euclidean distance.

**Step 3** Apply Inner Clustering over each type of gene representation of non-clustered elements of the closest pair using their distance, and subsequently merge them using overlapping clustering, and start the next level of hierarchical clustering with (N-1) elements.

**Step 4** Repeat steps (2) and (3) until the elements become a single cluster.

**Output:** Single cluster of N elements.

The hierarchical clustering starts with a random selection of $N$ clustering elements. Using the Euclidean distance, the closest pair elements that are self clustered are found out and merged, using overlapping clustering to produce a new larger cluster, and the next level of hierarchical clustering is started with the $N - 1$ elements and at the last level, a single cluster is formed. Each level of hierarchical clustering identifies similar gene types in the $N$ gene expression data and clusters them; the subsequent overlapping clustering identifies the co-express genes, that belong to more than one cluster. Figure 4.4 illustrates the developed two-dimensional clustering process. In Figure 4.4, the first level of hierarchical clustering is started with the four types of genes (represented in the inner rectangles). The closest pair gene type is found and merged using overlapping clustering, which identifies co-express genes that belong to more than one cluster. The next level of hierarchical clustering is preceded with three elements. Likewise, all the elements are clustered in the final level.
In the database $D_{MN}$, all the column elements have been clustered internally, and then the clustered columns are re-clustered horizontally.

**Table 4.1 Sample gene expression data**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>11</td>
<td>20</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Patient 2</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>13</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Patient 3</td>
<td>14</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td>17</td>
<td>10</td>
<td>8</td>
<td>14</td>
<td>15</td>
<td>11</td>
</tr>
</tbody>
</table>

A sample gene expression dataset with 10 genes, captured from three patients is given in Table 4.1.
Table 4.2 Results of inner gene clustering

<table>
<thead>
<tr>
<th>Patient</th>
<th>11</th>
<th>11</th>
<th>11</th>
<th>3</th>
<th>3</th>
<th>2</th>
<th>13</th>
<th>15</th>
<th>20</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td>5</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Patient 2</td>
<td>14</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 4.2 illustrates the inner gene clustering results of all the three patients’ gene values in Table 4.1. The average values of patient 1, patient 2 and patient 3 are 9.4, 11.9, and 11.4, respectively.

Table 4.3 Results of first level hierarchical clustering

<table>
<thead>
<tr>
<th>Patient</th>
<th>11</th>
<th>11</th>
<th>11</th>
<th>3</th>
<th>3</th>
<th>2</th>
<th>13</th>
<th>15</th>
<th>20</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>19</td>
<td>5</td>
<td>4</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>Patient 2</td>
<td>11</td>
<td>10</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td>11</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Patient 3</td>
<td>13</td>
<td>15</td>
<td>20</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 illustrates the first level clustering results of patient 2 and patient 3 having the minimum distance.

Table 4.4 Results of second level hierarchical clustering

<table>
<thead>
<tr>
<th>Patient</th>
<th>13</th>
<th>13</th>
<th>11</th>
<th>11</th>
<th>11</th>
<th>3</th>
<th>3</th>
<th>9</th>
<th>10</th>
<th>8</th>
<th>16</th>
<th>16</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 3</td>
<td>19</td>
<td>5</td>
<td>4</td>
<td>14</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td>17</td>
<td>15</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.4 illustrates the second level hierarchical clustering results. The co-expression nature of genes is depicted in Table 4.4; (i.e.), the gene ‘11’ is presented in more than one cluster. Likewise, all the three patients’ gene values are clustered at different levels, to analyze the presence of a gene in more than one cluster of the microarray gene expression data.
4.2.1.1 Finding the closest pair elements and Inner Clustering

The first level of hierarchical clustering starts with a random selection of ‘N’ elements. After that, the Euclidean distances between all the elements are calculated, and the closest pair having the minimum distance is selected for merging. In hierarchical clustering, the distance starts from zero (because each point exists in its own cluster), and increases when the clusters are merged.

The distance is calculated as follows:

$$ Ed(p, q) = \sqrt{\sum (p - q)^2} $$  \hspace{1cm} (4.4)

where ‘Ed’ is the Euclidean distance between the two clustering elements ‘p and ‘q’

‘p’ and ‘q’ are the two clustering elements in ‘d’

The distance between two elements p and q is represented as the root of, sum of, and square of, deviation among P and Q during merging. In this, ‘P’ and ‘Q’ are the gene representations containing ‘N’ clustering elements, where ‘N’ is the total gene sample values present in the gene expression database. $P = \{P[i]\}$ and $Q = \{Q[i]\}$, where $1 < i \leq N$ is the closest pair elements having the minimum distance; then, every gene representation in both the elements is clustered inner-wise by comparing each $P[i]$ with every $P[j]$ where $i \neq j$ using a threshold value.

4.2.1.2 Merging the Closest Pair

In hierarchical clustering, clusters are either merged into larger clusters or split into small clusters. In this technique, the gene representations in each closest pair elements are clustered and merged into a larger cluster, by
overlapping clustering as shown in Figure 4.3. \( C1 = \{C|I[i]\} \) and \( C2 = \{C|2[i]\} \) are the cluster sets of the closest pair elements ‘P’ and ‘Q’, respectively, where ‘I’ is a real number greater than one. The cluster sets \( C1 \) and \( C2 \) are merged together and a large cluster \( C = C1 \cup C2 \) is formed. After merging, the next level of hierarchical clustering is started with (N-1) elements. The processes, which are described in section 4.2.1.1 and 4.2.1.2, are repeated until the elements become a single cluster ‘C’.

### 4.2.2 Fitness Evaluation

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

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

(4.5)

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

(4.6)

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

‘w’ is the weight of each clustered element

\( R_{\text{def}} \) is the defined cluster dataset

If an element in the resultant cluster is in the defined cluster \( R_{\text{def}} \), then the weight of the cluster is assigned the value 1, and 0 otherwise. The two dimensional clustering process and fitness evaluation are processed, for every row of the index ‘I’ and the cluster ‘K’ is found out.

\[
K = \{C_i \mid 1 \leq i \leq l\}
\]

(4.7)

where ‘\( C_i \)’ is the clustering element in the resultant cluster set ‘K’

‘\( l \)’ is the last row of the index matrix ‘I’
From the resultant cluster set ‘K’, the best ‘k’ clusters having the highest fitness value are chosen and the closest index of all the best ‘k’ clusters is calculated, and these are then used to fetch the next set of clustering elements from the database. Selecting the next best gene representations which are closely correlated with the resultant cluster ‘K’ in each iteration, is a crucial factor for improving the clustering efficiency. This developed clustering algorithm, initially clusters based on the known labeling of the subjected data. Based on the data labels and clustering efficiency, the fitness function is configured, and then it is applied to the unknown data. As a limited supervision is applied, the clustering was named as semi-supervised clustering. The clustering is processed continuously until the optimum cluster is obtained.

4.3 EXPERIMENTAL RESULTS

This technique has been implemented in the working platform of the Matrix Laboratory version 7.8. The microarray gene samples of human acute leukemia, adenocarcinoma and lymphoma cancer cells are utilized for evaluating this technique. In this clustering technique, an adaptive approach is followed to dynamically define the number of clusters that must be generated from the micro array gene expression dataset. This technique, which is a multi-stage clustering one, performs clustering at different levels. The first level of clustering is started with N clustering elements. Using the Euclidean distance, the closest pair elements which are self clustered using the threshold value of one, are merged using overlapping clustering to produce a new larger cluster, and the next level of hierarchical clustering is started with the (N-1) elements. The above process is repeated until a predetermined clustering level is reached.
### 4.3.1 Dataset Description

The two data sets of standard leukemia for training and testing were obtained from [http://www.broadinstitute.org/cancer/software/genepattern/datasets/](http://www.broadinstitute.org/cancer/software/genepattern/datasets/), and the performance of the developed technique in clustering ground truth data cancer classes, namely, AML and ALL was demonstrated. The two training leukemia datasets were partitioned again, and turned into four sets (datasets 1, 2, 3 and 4), each having N values (20, 18, 18, 17) respectively. In dataset 1, the ALL/AML data is partitioned into 1–14 gene types belong to ALL and 15–20 gene types belong to AML as shown in Figure 4.5. Similarly, dataset 2 (1 to 13- ALL; 14 to 18- AML), dataset 3 (1 to 11 – ALL; 12 to 18 –AML) and dataset 4 (1 to 10 – ALL; 11 to 17-AML) were arranged.

![Gene samples](image)

**Figure 4.5 Leukemia ALL/AML dataset 1**

Adenocarcinoma datasets were partitioned into two sets (dataset 5, dataset 6) each having N values (79, 77), respectively. In dataset 5, 1–70 gene types were Adenocarcinoma cells and 71–79 gene types are normal cells. Likewise, in dataset 6, 1–69 gene types were Adenocarcinoma cells and 70–77 gene types were normal cells. The Lymphoma data sets were partitioned into two sets (dataset 7 and dataset 8) and the value of N for both datasets were 29. In both datasets 7 and 8, the 1–16 gene type values were Diffuse Large Bell Lymphoma (DLBL) cells, and 17–29 gene types were Non-Hodgkin Lymphoma (NHL) cells.
The clustering results of datasets 1, 2, 3 and 4 can be visualized in Figures 4.6(a-d). Figures 4.6(b-d) illustrate the clustering results of datasets 2, 3 and 4, respectively.

Figure 4.6 (a) Experimental results of the hierarchical clustering level of ALL/AML dataset 1
Figure 4.6(a) shows the clustered result of the ground truth dataset 1. A clustering element is represented by each of the circles, where the first 14 elements are of the ALL type, and the remaining elements are of the AML type. In level 1, the closest pair elements having the minimum distance to be merged are represented by two green circles.

Figure 4.6 (b) Experimental results of the hierarchical clustering level of ALL/AML dataset 2
Figure 4.6(b) which shows the clustered result of the ground truth of dataset 2. A clustering element is represented by each of the circles, where the first 13 elements are of the ALL type, and the remaining elements are of the AML type. At level 1, the closest pair elements having the minimum distance to be merged are represented by two green circles.

Figure 4.6 (c) Experimental results of the hierarchical clustering level of ALL/AML dataset 3
Figure 4.6(c) shows the clustered result of the ground truth of dataset 3. A clustering element is represented by each of the circles, where the first 11 elements are of the ALL type, and the remaining elements are of the AML type. At level 1, the closest pair elements having the minimum distance to be merged are represented by two green circles.

Figure 4.6 (d) Experimental results of the hierarchical clustering level of ALL/AML dataset 4
Figure 4.6(d) shows the clustered result of the ground truth dataset 4. A clustering element is represented by each of the circles, where the first 10 elements are of the ALL type, and the remaining elements are of the AML type. At level 1, the closest pair elements having the minimum distance to be merged are represented by two green circles.

Figures 4.7(a-d) reveal the inner gene clustering process of the data sets, respectively. All the closest pair elements at each level are clustered, and the resultant clusters are represented in different coloured circles. At the final level it is found, that all the AML types are clustered successfully, whereas the ALL types are clustered with the low precision value of 0.1429. Similarly, all the clustering elements in the remaining three datasets are clustered successfully.

Figure 4.7 (a) Experimental results of the inner gene clustering of ALL/AML dataset 1
Figure 4.7 (b) Experimental results of the inner gene clustering of ALL/AML dataset 2

Figure 4.7 (c) Experimental results of the inner gene clustering of ALL/AML dataset 3
Figure 4.7 (d) Experimental results of the inner gene clustering of ALL/AML dataset 4

4.3.2 Performance Evaluation

The semi supervised clustering technique is the improved version of the classical unsupervised clustering technique, in terms of using some labeled data for clustering. Thus, the performance of this semi supervised two dimensional hierarchical data clustering technique, was compared with that of the existing unsupervised clustering algorithms such as, hierarchical clustering, Fuzzy C-mean Clustering(FCM), Fuzzy K-mean Clustering(FKM), Self- Organizing Maps(SOM), Hybrid FCM, kernel based and Genetic Algorithm (GA) based hierarchical clustering (Sheikh et al 2008) was compared. The FCM is widely used to identify the cluster structures in high-dimensional datasets, which allows one element to belong to two or more clusters with different probabilities (Bezdek et al 1984). The FKM clustering is an extension of K-Means performs a fuzzy form of k-means clustering. It tries to generate overlapping clusters from the data set. The Fuzzy k-means discovers the soft clusters. In a soft cluster, any point can
belong to more than one cluster with a certain probability (Kim et al 2006). A self-organizing map (SOM) is an unsupervised neural network learning algorithm which has been successfully used for the analysis and organization of large data files (Kohonen 1990). It is a discrete grid of map units. Each map unit can represent the genes in a chosen dataset. In the hybrid FCM method, the FCM combined with the Expectation Maximization (EM) algorithm, provides the statistical framework to model the cluster structure of the gene expression data (Valarmathie et al 2009). The kernel version of hierarchical clustering uses the kernel trick to map implicitly, the gene expression data into a higher dimensional feature space (Qin et al 2003).

The performance of this method is evaluated on clustering the ground truth data of the cancer classes using precision, recall and F-measure (Victor and Luis 2005), and subsequently, these values are compared with the precision, recall and F-measure values of the above mentioned unsupervised clustering algorithms. The Precision, recall and F-measure are calculated based on following formulae:

The precision of cluster $i$ with respect to class $j$ is

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

where Precision $(i,j)$ is the probability that a member of the cluster $i$ belongs to class $j$

$M_{ij}$ is the number of objects of class $j$ in cluster $i$

$M_j$ is the number of objects in class $i$.

The recall of cluster $i$ with respect to class $j$ is

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

(4.9)
where $M_i$ is the number of objects in class $j$.

$M_{ij}$ is the number of objects of class $j$ in cluster $i$.

The F-measure of cluster $i$ with respect to class $j$ is

$$F - \text{measure } (i, j) = \frac{2 \times \text{Recall} (i, j) \times \text{Precision} (i, j)}{\text{Precision} (i, j) + \text{Recall} (i, j)}$$  \hspace{1cm} (4.10)

The precision, recall and F-measure values of the clusters obtained by the two dimensional hierarchical clustering data techniques for human acute leukemia, adenocarcinoma and lymphoma cancer cells are given in Tables 4.5, 4.6 and 4.7, respectively. The performance of the two dimensional hierarchical data clustering approach is also evaluated by comparing its clustering results with those of the above mentioned comparative techniques, using the clustering performance measure rand index. The Rand index is a technique for measuring the class agreement (Eva et al 2010). In order to compute the rand index a comparison of all pairs of objects in the dataset after clustering should be performed. The Rand index, which is based on pair agreement and disagreement, takes a value [0, 1], where ‘1’ indicates perfect agreement. If two objects are in the same cluster in both the ground truth clustering and the resultant clustering to be investigated, this counts as an agreement. If two objects are in different clusters in both the ground truth clustering and the resultant clustering to be investigated, this is also an agreement. Otherwise, there is a disagreement.

$$\text{Rand index} = \frac{\text{No.of agreements}}{\text{No.of agreements} + \text{No.of Disagreements}}$$  \hspace{1cm} (4.11)
Table 4.5  Performance metrics of the two dimensional hierarchical clustering technique for human acute leukemia dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cluster</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Rand Index</th>
<th>Computational Time(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>C1</td>
<td>1.0000</td>
<td>0.9286</td>
<td>0.9630</td>
<td>0.65</td>
<td>2.0149</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>0.8571</td>
<td>1.0000</td>
<td>0.9231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 2</td>
<td>C3</td>
<td>0.9286</td>
<td>1.0000</td>
<td>0.9630</td>
<td>0.77778</td>
<td>2.5432</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>1.0000</td>
<td>0.8000</td>
<td>0.8889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 3</td>
<td>C5</td>
<td>0.6875</td>
<td>1.0000</td>
<td>0.8148</td>
<td>0.88889</td>
<td>1.0056</td>
</tr>
<tr>
<td></td>
<td>C6</td>
<td>1.0000</td>
<td>0.2857</td>
<td>0.4444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 4</td>
<td>C7</td>
<td>0.8182</td>
<td>0.9000</td>
<td>0.8571</td>
<td>0.64706</td>
<td>1.0023</td>
</tr>
<tr>
<td></td>
<td>C8</td>
<td>0.8333</td>
<td>0.7143</td>
<td>0.7692</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5 shows the precision, recall, F-measure, rand index and execution time of the clustering process, obtained by the two dimensional hierarchical clustering data technique for human acute leukemia dataset.

Table 4.6  Performance metrics of the two dimensional hierarchical clustering technique for Adenocarcinoma dataset

<table>
<thead>
<tr>
<th>Dataset 5</th>
<th>Cluster</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Rand Index</th>
<th>Execution Time(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.9844</td>
<td>0.9000</td>
<td>0.9403</td>
<td></td>
<td>0.81013</td>
<td>6.02</td>
</tr>
<tr>
<td>C2</td>
<td>0.5333</td>
<td>1.8889</td>
<td>0.6667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 6</td>
<td>C3</td>
<td>0.9444</td>
<td>1.9855</td>
<td>0.9645</td>
<td>0.93506</td>
<td>6.1</td>
</tr>
<tr>
<td>C4</td>
<td>0.8000</td>
<td>0.5000</td>
<td>0.6154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.6 shows the precision, recall, F-measure, rand index and execution time of the clustering process, obtained by the two dimensional hierarchical clustering data technique, for the Adenocarcinoma dataset.

Table 4.7 Performance metrics of the two dimensional hierarchical clustering technique for Lymphoma dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cluster</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Rand Index</th>
<th>Execution Time(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 5</td>
<td>C1</td>
<td>0.5714</td>
<td>1</td>
<td>0.7273</td>
<td>0.96552</td>
<td>2.6649</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>1</td>
<td>0.0769</td>
<td>0.1429</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 6</td>
<td>C3</td>
<td>0.625</td>
<td>0.625</td>
<td>0.625</td>
<td>0.55172</td>
<td>2.4432</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>0.5385</td>
<td>0.5385</td>
<td>0.5385</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7 shows the precision, recall, F-measure, rand index and execution time of the clustering process obtained by the two dimensional hierarchical clustering data technique, for the Lymphoma dataset.

Tables 4.5, 4.6 and 4.7 illustrate that the two dimensional hierarchical clustering approach achieves the mean values of the precision, recall, F-measure, rand index and execution time values as 0.820106, 0.758963, 0.740256, 0.77827, and 2.974263 respectively. Figures 4.8, 4.9 and 4.10 represent the comparative results of the clustering, using the developed hierarchical clustering technique with the FCM, FKM, hybrid FCM, SOM, conventional hierarchical clustering, kernel based and Genetic Algorithm(GA) based hierarchical clustering techniques, using adenocarcinoma, human acute leukemia and lymphoma datasets respectively. Table 4.8 illustrates the mean and standard deviation values of clustering by the developed approach and the comparative techniques.
Figure 4.8 (a) Precision comparative result of human acute leukemia

Figure 4.8(a) shows the precision values of human acute leukemia clusters, using the two dimensional hierarchical clustering approach and the other comparative techniques. The dark brown line shows that the performance of the two dimensional hierarchical clustering approach is higher.

Figure 4.8 (b) Recall comparative result of human acute leukemia

Figure 4.8(b) shows the recall values of human acute leukemia clusters, using the two dimensional hierarchical clustering approach and the
other comparative techniques. The dark brown line shows the higher performance of the two dimensional hierarchical clustering.

**Figure 4.8 (c) F-measure comparative result of human acute leukemia**

Figure 4.8(c) shows the F-measure values of human acute leukemia clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows the higher F-measure values of the two dimensional hierarchical technique.

**Figure 4.8 (d) Rand index comparative result of human acute leukemia**

Figure 4.8(d) shows the rand index values of human acute leukemia clusters, using the two dimensional hierarchical clustering approach, and the
other comparative techniques. The dark brown line shows the higher performance of the two dimensional hierarchical clustering approach.

![Figure 4.9 (a) Precision comparative result of Adenocarcinoma](image)

Figure 4.9 (a) Precision comparative result of Adenocarcinoma

Figure 4.9(a) shows the precision values of adenocarcinoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the precision value of the two dimensional hierarchical clustering approach is higher than that of the other clustering techniques.

![Figure 4.9 (b) Recall comparative result of Adenocarcinoma](image)

Figure 4.9 (b) Recall comparative result of Adenocarcinoma
Figure 4.9(b) shows the recall values of adenocarcinoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the recall value of the two dimensional hierarchical clustering approach is higher than that of the other clustering techniques.

Figure 4.9 (c) F-measure comparative result of Adenocarcinoma

Figure 4.9(c) shows the F-measure values of adenocarcinoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the F-measure value of the two dimensional hierarchical clustering approach is higher than that of the other clustering techniques.

Figure 4.9 (d) Rand index comparative result of Adenocarcinoma
Figure 4.9(d) shows the rand index values of adenocarcinoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows the high rand index value of the two dimensional hierarchical clustering technique.

Figure 4.10 (a) Precision comparative result of Lymphoma

Figure 4.10(a) shows the precision values of lymphoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the precision value of the two dimensional hierarchical clustering approach is higher.

Figure 4.10 (b) Recall comparative result of Lymphoma
Figure 4.10(b) shows the recall values of lymphoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows the high recall values of the two dimensional hierarchical clustering approach.

Figure 4.10 (c) F-measure comparative result of Lymphoma

Figure 4.10(c) shows the F-measure values of lymphoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the F-measure value of the two dimensional hierarchical clustering approach is higher.

Figure 4.10 (d) Rand index comparative result of Lymphoma

Figure 4.10(d) shows the Rand index values of lymphoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the Rand index value of the two dimensional hierarchical clustering approach is higher.
Figure 4.10(d) shows the rand index values of lymphoma clusters, using the two dimensional hierarchical clustering approach, and the other comparative techniques. The dark brown line shows that the rand index value of the two dimensional hierarchical clustering approach is higher than that of the other clustering techniques. From Figures 4.8, 4.9 and 4.10, it can be seen that the precision, recall and F-measure values of clusters using the two dimensional hierarchical clustering approach are slightly deviated for some clusters, and the overall performance of the two dimensional hierarchical clustering approach is higher than that of the other comparative techniques.

From Table 4.8, it can be seen that according to the observed standard deviation values, the minimum and maximum range of precision values can be achieved by the two dimensional hierarchical approach, FCM, FKM, SOM, Hybrid FCM, Hierarchical clustering, Kernel and GA based clustering are (0.64 - 0.9961), (0.20 - 0.67), (0.20 - 0.72), (0.19 - 0.69), (0.20 - 0.69), (0.21 - 0.83), (0.39-0.87) and (0.35-0.76) respectively.

The minimum and maximum range of values are calculated as follows:

Minimum range = minimum (µ(C)) – minimum (σ(C))  \[(4.11)\]

Maximum range = maximum (µ(C)) – maximum (σ(C))  \[(4.12)\]

where ‘µ’ is the mean of the cluster set C

‘σ’ is the standard deviation of the cluster set C

Likewise the minimum and maximum range of recall values that can be achieved are (0.48- 1.04), (0.23-0.70), (0.29-0.70), (0.22-0.69), (0.25-0.67), (0.22-0.69), (0.49-0.82), (0.37-0.71) and the minimum and maximum range of F-measure values that can be achieved are (0.51-0.96), (0.24 - 0.65), (0.27 - 0.61), (0.21 - 0.62), (0.24 - 0.56), (0.22 - 0.69), (0.45-0.79), (0.34-0.66).
Table 4.8 Comparison using Statistical Metrics

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision $\mu$</th>
<th>Precision $\sigma$</th>
<th>Recall $\mu$</th>
<th>Recall $\sigma$</th>
<th>F-measure $\mu$</th>
<th>F-measure $\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>0.4504125</td>
<td>0.2493395</td>
<td>0.467125</td>
<td>0.234032</td>
<td>0.444842</td>
<td>0.203599</td>
</tr>
<tr>
<td>Two dimensional HC</td>
<td>0.820106</td>
<td>0.176007</td>
<td>0.758963</td>
<td>0.282328</td>
<td>0.740256</td>
<td>0.228711</td>
</tr>
<tr>
<td>FKM</td>
<td>0.4644</td>
<td>0.25965</td>
<td>0.491238</td>
<td>0.205319</td>
<td>0.442875</td>
<td>0.171495</td>
</tr>
<tr>
<td>SOM</td>
<td>0.440525</td>
<td>0.246665</td>
<td>0.456419</td>
<td>0.233835</td>
<td>0.414225</td>
<td>0.208236</td>
</tr>
<tr>
<td>Hybrid FCM</td>
<td>0.445325</td>
<td>0.246447</td>
<td>0.459356</td>
<td>0.210322</td>
<td>0.402656</td>
<td>0.159697</td>
</tr>
<tr>
<td>Hierarchical clustering</td>
<td>0.525581</td>
<td>0.310596</td>
<td>0.453488</td>
<td>0.235111</td>
<td>0.453488</td>
<td>0.235111</td>
</tr>
<tr>
<td>Kernel based semi-supervised</td>
<td>0.633437</td>
<td>0.238793</td>
<td>0.6575</td>
<td>0.1665933</td>
<td>0.623562</td>
<td>0.1700638</td>
</tr>
<tr>
<td>GA based semi-supervised</td>
<td>0.560131</td>
<td>0.206321</td>
<td>0.545</td>
<td>0.169509</td>
<td>0.502688</td>
<td>0.158605</td>
</tr>
</tbody>
</table>

Overall, Table 4.8 shows that the two dimensional hierarchical clustering approach is 54.92%, 56.62%, 53.71%, 54.30%, 64%, 51%, 47% more accurate than those of the FCM, FKM, SOM, Hybrid FCM, Hierarchical clustering, Kernel and GA based semi-supervised techniques, respectively. Likewise, the recall and F-measure values of the two dimensional hierarchical clustering technique are 61.54%, 64.72%, 60.13%, 60.52%, 59.75%, 48%, 40% and 60.09%, 59.82%, 55.95%, 54.39%, 61.26%, 12%, 37% more than those of the above discussed relative techniques, respectively.
4.4 SUMMARY OF THIS WORK

- An innovative two dimensional hierarchical clustering technique to handle the microarray genes that exist in more than one cluster is developed.

- The approach includes inner gene clustering as well as hierarchical clustering, which have used both the local and global information hidden in the set of the gene expression data for deciding which genes would be grouped together.

- The hierarchical clustering has used the local pair wise similarity measure computed between gene representations, whereas the inner gene clustering has found out the global overlapping regions between the clusters.

- The technique was implemented and experimented with three real life datasets. The analytical results have confirmed that the two dimensional hierarchical technique has achieved better performance measures than the other comparative techniques.

- The precision minimum range of approach was approximately nearer to the precision maximum range of the FCM, SOM and Hybrid FCM.

- The minimum F-measure of the approach was approximately close to the maximum F-measure of the existing techniques.

- The two dimensional hierarchical clustering technique has achieved more precision, recall and F-score values than the existing clustering techniques.
• The efficiency of this approach was higher than that of the conventional approach. Such a performance has been achieved because of the fact that unlike conventional clustering, the inner clustering was performed to discover the overlapping clusters, which resulted in biologically meaningful clusters.

• The most prominent gene expression data was selected by the approach using the index matrix, which has reduced the biological complexities of working in the whole database.