4. Hierarchical Agglomerative Clustering

4.1. Introduction

Hierarchical clustering algorithms create sets of clusters. They differ in how the sets are created. A tree data structure called dendrogram is used to illustrate the hierarchical clustering technique and the sets of different clusters. The space complexity of hierarchical algorithms is $O(n^2)$ as this is the space required for the adjacency matrix. The time complexity is $O(kn^2)$ because there is one iteration for each level in the dendrogram. HAC start with each individual item in its own cluster and iteratively merge clusters until all items belong to one cluster. The basic process of HAC is the merging of clusters based on their proximity [143].

The main advantages of the HAC approach are as follows [144]:

1. Simple computation and easy implementation.

2. Less restricted assumptions and more flexibility: HAC could use simple qualitative connectivity information of a network or quantitative data through Received Signal Strength (RSS) or GPS. In addition, other factors could easily be incorporated into the algorithm. For instance, different weights could be assigned to different nodes or connections for specific scenarios.

3. Less resource for clusters establishment: Using the HAC approach, nodes can finish the CHs election and announcement, cluster establishment, and scheduling at the same time. It can greatly reduce resource dissipation.

4. Without the need of periodic re-clustering or network updating: The HAC approach generates a logical CH backup chain during the cluster generation process. It makes clusters easily adaptive to network changes without extra information exchanges or the need of periodic announcement, such as CH.

Different hierarchical clustering techniques are classified as shown in the Figure 4.1. Hierarchical algorithms are categorized as agglomerative or divisive [145]. Agglomerative implies that the clusters are created in a bottom-up fashion, while divisive algorithms work in...
4. Hierarchical Agglomerative Clustering

Figure 4.1.: Hierarchical clustering
4.2. Hierarchical Quantitative and Qualitative Clustering

The current chapter proposes a HAC algorithm to have minimum energy consumption in a WSN. Section 4.2 explains the Hierarchical Quantitative and Qualitative Clustering. Section 4.3 deals with HAC process and the section 4.4 talks about the simulation and results for qualitative and quantitative data.

4.2. Hierarchical Quantitative and Qualitative Clustering

The motivation of the research is to provide efficient clustering without requiring the global knowledge of network by using the bottom-up called the Hierarchical agglomerative clustering (HAC)[144]. With the bottom-up approach, sensing nodes build clusters before they select CHs. In this manner, the bottom-up approach can be a better way to implement self-organization, scalability and flexibility. If the distance between the sensing nodes is calculated using their location then it’s quantitative HAC. If the received signal strength is used to calculate the distance between the nodes then it’s qualitative HAC. This chapter compares the various agglomerative clustering techniques applied in a WSN. The simulations are done in MATLAB and the comparisons are made between the different protocols using dendrograms[14].

4.3. HAC Process

The process of HAC comprises three common key steps: obtaining the input data set, computation of the resemblance coefficients, and executing the clustering method.

4.3.1. Input Data Set

An input data set for HAC is a component-attribute data matrix. Components are the sensing nodes that are to be clustered based on their similarities. Attributes are the properties of the components. The attributes are the x and y coordinates of the sensing nodes that give the exact position of the node. The components or the attributes can be easily added or removed from the data set for different applications. Obviously, the more factors are considered, the more restricted assumptions and computations are needed. The type of input data set can be classified into quantitative data and qualitative data. The location information is used as the quantitative input data and the connectivity information is used as the qualitative data.
4. Hierarchical Agglomerative Clustering

4.3.2. Computation of Resemblance Coefficients

A resemblance coefficient for a given pair of components indicates the degree of similarity or dissimilarity between these two components, depending on the way in which the data is represented. It could be quantitative or qualitative. For quantitative data, the current research work uses Euclidean distance as the similarity measure. Less the Euclidean distance and more similar the two sensing nodes and thus belong to the same cluster.

If \( S_1 = (x_1, y_1) \) and \( S_2 = (x_2, y_2) \) are the two nodes, then the distance \( \text{Dis} \) between \( S_1 \) and \( S_2 \) is given by Eqn. 4.1.

\[
\text{Dis}(S_1, S_2) = D_{1,2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}
\]  (4.1)

The distance between each pair of nodes is calculated and the adjacency matrix is obtained as given in Table 4.2. For quantitative data [146], following ways are used to calculate the resemblance coefficients Eqns.4.2-4.5:

- Jaccard:
  \[
  SIM(a, b) = \frac{M_{1-1}}{M_{1-1} + M_{1-0} + M_{0-1}}
  \]  (4.2)

- Sorenson:
  \[
  SIM(a, b) = \frac{2M_{1-1}}{2M_{1-1} + M_{1-0} + M_{0-1}}
  \]  (4.3)

- Simple Matching:
  \[
  SIM(a, b) = \frac{M_{1-1} + M_{1-0}}{M_{1-1} + M_{1-0} + M_{0-1} + M_{0-0}}
  \]  (4.4)

- Dice:
  \[
  SIM(a, b) = \frac{M_{1-1}}{M_{1-1} + 0.5M_{1-0} + 0.5M_{0-1}}
  \]  (4.5)

where \( M \) is the attribute which comes from the entries in Connectivity matrix based on the direct connection of the node,

- \( M_{1-1} \): represents the total number of attributes where \( a \) and \( b \) both have a value of 1,
- \( M_{0-0} \): represents the total number of attributes where \( a \) and \( b \) both have a value of 0,
- \( M_{0-1} \): represents the total number of attributes where the attribute of \( a \) is 0 and the attribute of \( b \) is 1,
- \( M_{1-0} \) : represents the total number of attributes where the attribute of \( a \) is 1 and the attribute of \( b \) is 0.

Dissimilarity coefficient

\[
DSIM(a, b) = 1 - SIM(a, b)
\]  (4.6)
4.3. HAC Process

In the simulation, the clustering of the qualitative and quantitative data is done using single link, complete link and average link methods as explained in the next sections.

4.3.3. Execution of the HAC Method

Execution of HAC involves various steps and each step merges two clusters together and updates the Resemblance Matrix. Updating the Resemblance Matrix is an important step and various methods could be adopted. With the same data set, we may get different clustering results by using different HAC algorithms. The type of algorithm depends on how the distance between the motes in two clusters is calculated. This is not an easy task as there are many interpretations for the distance between clusters. There are three main types of HAC [8]:

1. Single link: It’s also called as nearest neighbor method. Two clusters are merged if the minimum distance between any two points is less than or equal to the threshold distance being considered. It considers the smallest distance between a node in one cluster and a node in the other. Thus

\[
Dis(K_i, K_j) = \min(D_{il}, D_{jm}) \quad \forall S_{il} \in K_i, S_{jm} \not\in K_j
\]

and

\[
S_{jm} \in K_j, S_{il} \not\in K_i
\] (4.7)

where \(K_i\) and \(K_j\) are two different clusters in the sensor network \(N\), \(S_{pq}\) is the \(q^{th}\) sensor node of \(p^{th}\) cluster.

2. Complete link: Although it’s similar to the single link algorithm, the difference lies in the fact that it considers the largest distance between a node in one cluster and a node in the other. Thus

\[
Dis(K_i, K_j) = \max(D_{il}, D_{jm}) \quad \forall S_{il} \in K_i, S_{jm} \not\in K_j
\]

and

\[
S_{jm} \in K_j, S_{il} \not\in K_i
\] (4.8)

3. Average link: The average link technique merges two clusters if the average distance between any two points in the two target clusters is below the distance threshold. It takes the average distance between a node in one cluster and a node in the other. Thus

\[
Dis(K_i, K_j) = \text{mean}(D_{il}, D_{jm}) \quad \forall S_{il} \in K_i, S_{jm} \not\in K_j
\]
4. Hierarchical Agglomerative Clustering

and

\[ S_{jm} \in K_j \notin K_i \quad (4.9) \]

4.3.4. Single Complete and Average Link Clustering Algorithms

The steps involved in single link clustering algorithms are as follows:

1. Find the minimum distance in the adjacency matrix.
2. Cluster the two nodes with the minimum distance.
3. The distance between the clustered nodes is calculated with the rest of the unclustered nodes.
4. Let’s say \( i \) and \( j \) nodes are clustered since \( D_{(i,j)} \) was minimum in the adjacency matrix.
5. Now the adjacency matrix is updated as
   a) \( D_{(i,k)} \) and \( D_{(j,k)} \) is replaced with \( \min(D_{(i,k)}, D_{(j,k)}) \forall k \in N \)
   b) The dimension of the adjacency matrix is also reduced by one.
6. The steps 1-6 are repeated until the adjacency matrix is left with 2*2 elements.

The complete link algorithm works in the same way except the 5\(^{th}\) step is replaced with \( \max(D_{(i,k)}, D_{(j,k)}) \forall k \in N \). The dimension of the adjacency matrix is also reduced by one. In the average link algorithm the distance is the average of the two distances and replaced with \( \text{mean}(D_{(i,k)}, D_{(j,k)}) \forall k \in N \).

4.4. Simulation and Results

Some of the assumptions made during simulation are as follows:

1. The nodes in the network are quasi-stationary, i.e. location of neighbour need not be determined every time clustering is performed.[10]
2. Propagation channel is symmetric.
3. Nodes are left unattended after deployment.
4. All nodes have similar capabilities, processing, communication and initial energy.
5. The coordinates of the nodes deployed is known.
6. The nodes are deployed at fixed locations.
4.4. Simulation and Results

Ten sensing nodes were randomly deployed in a room. The sensing nodes are clustered based on hierarchical agglomerative qualitative and quantitative clustering. As mentioned earlier for quantitative data, Euclidian method is used to calculate the resemblance coefficients and for qualitative data, Sorensen method is used to calculate the resemblance coefficients. The plot of the deployed nodes is depicted in the Figure 4.2.

### 4.4.1. Quantitative Data

Firstly for quantitative data, the adjacency matrix is calculated, based on Euclidian distances. The x and y coordinates of the sensors are depicted in Table 4.1. The adjacency matrix is as shown in the Table 4.2.

<table>
<thead>
<tr>
<th>Node ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X data</td>
<td>5</td>
<td>9.5</td>
<td>2.3</td>
<td>6.1</td>
<td>4.9</td>
<td>8.9</td>
<td>7.6</td>
<td>4.6</td>
<td>8.2</td>
<td>4.4</td>
</tr>
<tr>
<td>Y data</td>
<td>5</td>
<td>6.2</td>
<td>7.9</td>
<td>9.2</td>
<td>7.4</td>
<td>1.8</td>
<td>4.1</td>
<td>9.4</td>
<td>4.7</td>
<td>8.9</td>
</tr>
</tbody>
</table>
4. Hierarchical Agglomerative Clustering

Table 4.2.: Adjacency matrix for quantitative data

<table>
<thead>
<tr>
<th>Node ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4.66</td>
<td>3.96</td>
<td>4.34</td>
<td>2.40</td>
<td>5.04</td>
<td>2.75</td>
<td>4.42</td>
<td>3.21</td>
<td>3.95</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0</td>
<td>7.40</td>
<td>4.53</td>
<td>4.75</td>
<td>4.44</td>
<td>2.83</td>
<td>5.86</td>
<td>1.98</td>
<td>5.77</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>4.01</td>
<td>2.64</td>
<td>8.99</td>
<td>6.52</td>
<td>2.75</td>
<td>6.71</td>
<td>2.33</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>2.16</td>
<td>7.90</td>
<td>5.31</td>
<td>1.50</td>
<td>4.97</td>
<td>1.73</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>6.89</td>
<td>4.26</td>
<td>2.02</td>
<td>4.26</td>
<td>1.58</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>2.64</td>
<td>8.73</td>
<td>2.98</td>
<td>8.40</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>6.09</td>
<td>0.85</td>
<td>5.77</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>5.92</td>
<td>0.54</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>5.66</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Based on the above distance matrix, clustering is done using the three approaches of single link, complete link and average link. Correspondingly the Figures 4.3, 4.4 and 4.5 represent the dendrograms formed in single link, complete link and average link respectively.

Biggest disadvantage in the single link HAC is that the clustering creates clusters with long chain as shown in the Figure 4.3.

4.4.2. Qualitative Data

In the case of qualitative data, the x and y coordinates are not required to be calculated. The current research work uses the one-hop network connectivity data as the qualitative input data, where the “1” value represents a one-hop connection and the “0” value represents
4.4. Simulation and Results

**Figure 4.4.:** Complete link HAC for quantitative data

**Figure 4.5.:** Average link HAC for quantitative data
no direct connection. The received signal strength (RSS) is used to find out the network connectivity. If two nodes are within the transmission range then it’s a one hop connection. The Table 4.3 is the connectivity matrix obtained by the one hop connectivity of the nodes is shown in the Figure 4.6.

The adjacency matrix obtained by Sorenson coefficients is shown in the Table 4.4.

Based on the above distance matrix, clustering is done using the three approaches of single link, complete link and average link. Correspondingly the Figures 4.7, 4.8 and 4.9 represent the dendrograms formed in single link, complete link and average link respectively for the quantitative data.

Figure 4.6.: One hop connectivity diagram
Table 4.3.: Connectivity matrix

<table>
<thead>
<tr>
<th>Node ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.4.: Adjacency matrix by Sorenson dissimilarity coefficients

<table>
<thead>
<tr>
<th>Node ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.56</td>
<td>0.33</td>
<td>0.45</td>
<td>0.14</td>
<td>0.56</td>
<td>0.33</td>
<td>0.45</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0.64</td>
<td>0.33</td>
<td>0.33</td>
<td>1</td>
<td>0.33</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.09</td>
<td>0.14</td>
<td>1</td>
<td>0.67</td>
<td>0.09</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.23</td>
<td>1</td>
<td>0.82</td>
<td>0</td>
<td>0.82</td>
<td>0.09</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.64</td>
<td>0.43</td>
<td>0.23</td>
<td>0.43</td>
<td>0.14</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.33</td>
<td>1</td>
<td>0.33</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.82</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.82</td>
<td>0.09</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.7.: Single link HAC for qualitative data
4. Hierarchical Agglomerative Clustering

Figure 4.8.: Complete link HAC for qualitative data

Figure 4.9.: Average link HAC for qualitative data

4.5. Results and Observations

A wireless sensor network consists of a large number of sensing nodes randomly deployed in a region. Clustering not only reduces energy consumption but also achieves better network performance. Hierarchical agglomerative clustering uses bottom up approach and hence the complete network information is not required. The dendrograms for the various HAC proto-
4.6. Summary

cols for qualitative and quantitative data were shown in the previous section. If the position of the nodes is not known due to the issues like cost, time, energy conservation then the Sorensen method is used to find out the similarity matrix and then perform the clustering of the nodes. The dendrograms for single and complete link protocol clearly shows that single link has a chain effect, making the clustering inefficient. The dendrograms of quantitative single link, complete link, and average link are almost similar to the dendrograms of qualitative single link, complete link and average link respectively. So, it shows that there is no need to find out the location of the nodes in a randomly deployed WSN. Finding out the position of the nodes not only make the clustering complex but also consumes a lot of energy. When the number of nodes increase or behave asymptotically, then it becomes computationally complex to calculate the adjacency matrix. Hence in such cases, Qualitative approach is used.

4.6. Summary

Clustering not only reduces energy consumption but also achieves better network performance. Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The dendrograms of quantitative single link, complete link, and average link are almost similar to the dendrograms of qualitative single link, complete link and average link respectively. So, it demonstrates that there is no need to find out the location of the nodes in a randomly deployed WSN. Finding out the position of the nodes not only make the clustering complex but also consumes a lot of energy.