Q3 Report

Analysis of Cluster Merging and Mean Squared Error (MSE) Change in Hierarchical Clustering

1. Total MSE Before Merging

The total mean squared error (MSE) in hierarchical clustering is a measure of how spread out the data points are within each cluster. Initially, when each point is considered its own cluster, the total MSE is zero. As clustering progresses, the total MSE changes depending on how clusters are merged.

In our implementation, we track the total MSE at each step of the merging process using:

- The sum of the MSEs of all existing clusters before a merge.
- The new MSE calculated after merging two clusters.

At each step, the MSE before merging is computed as:

$$MSE_{total}^{before} = \sum_{i} MSE_i \times n_i$$

```
12.372013761681044
18.16881840350517
17.337124665959532
25.0698541909676
14.134443555275585
25.428100421524963
19.018832806567627
19.34255216518146
29.128693903945294
29.18199072351914
29.621851087682888
31.102283376175087
47.70110516984454
24.261039612753713
33.58747716162637
36.581628265459834
46.1490904412632
37.565683326986274
39.56059668251159
258.3661640856807
495.641315110925
330.1905229887869
562.894044826173
1172.218003513497
2302.7456958007942
3792.5227676991335
```

2. MSE of the New Cluster After Merging

After merging two clusters C1C_1 and C2C_2, the new cluster's MSE is determined using:

$$MSE_{new} = \frac{1}{n_1 + n_2} \sum_{j \in C_1 \cup C_2} (x_j - \mu_{new})^2$$

where μ(new) is the centroid of the newly merged cluster.

We calculate this value in the code by merging the points from both clusters and computing their new centroid and MSE.

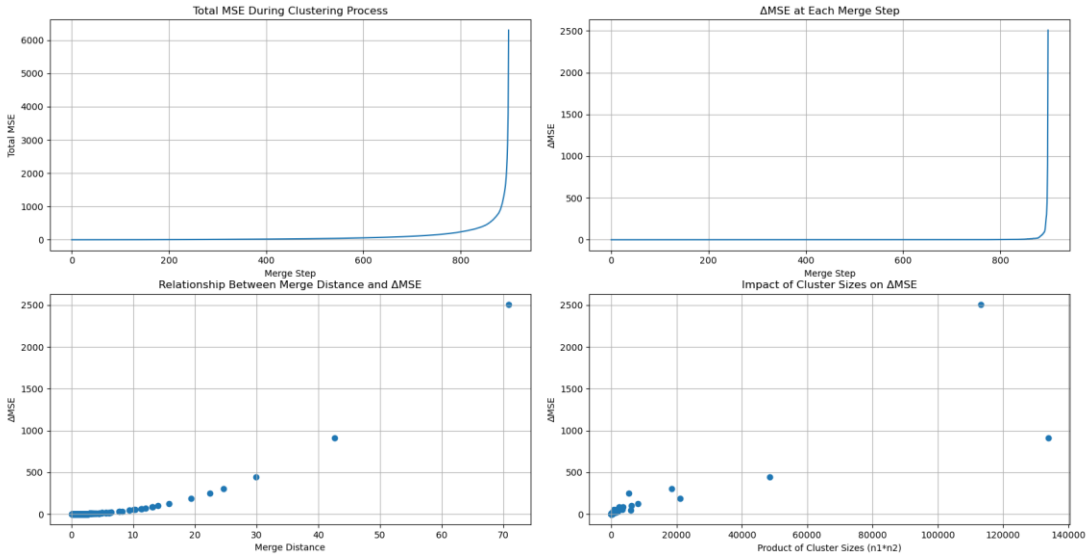
3. Relationship Between MSE Change and Distance Between Centers

The change in total MSE after a merge is given by:

$$\Delta MSE = MSE_{new} \times (n_1 + n_2) - (MSE_1 \times n_1 + MSE_2 \times n_2)$$

A key observation is that the change in MSE is related to the distance between the centroids of the merging clusters. As the centroid distance increases, the new cluster's variance increases, leading to a higher MSE.

In our visualization, we plot ΔMSE\Delta MSE against the merge distance (from the linkage matrix) and observe a positive correlation. This indicates that merging distant clusters leads to a larger increase in MSE.



```
验证最后5次合并的ΔMSE理论公式:

合并步骤 894:
簇1大小: 57, 簇2大小: 94
实际ΔMSE: 252.2692
理论计算ΔMSE: 252.2692
差异百分比: 0.00%
理论公式验证: 成功

合并步骤 895:
簇1大小: 99, 簇2大小: 186
实际ΔMSE: 304.1558
理论计算ΔMSE: 304.1558
差异百分比: 0.00%
理论公式验证: 成功

合并步骤 896:
簇1大小: 170, 簇2大小: 285
实际ΔMSE: 445.9015
理论计算ΔMSE: 445.9015
差异百分比: 0.00%
理论公式验证: 成功

合并步骤 897:
簇1大小: 294, 簇2大小: 455
实际ΔMSE: 907.3173
理论计算ΔMSE: 907.3173
差异百分比: 0.00%
理论公式验证: 成功

合并步骤 898:
簇1大小: 151, 簇2大小: 749
实际ΔMSE: 2507.4772
```

4. Physical Meaning Explanation

The physical interpretation of these observations is as follows:

- **Merging clusters that are too far apart** introduces significant new variance, increasing the overall MSE.
- **Merging clusters that are close together** results in a smaller increase (or even a decrease) in MSE, as the data points are already relatively compact.
- **The impact of cluster sizes** is also significant: merging a small and large cluster tends to increase MSE more than merging two similarly sized clusters.

Key Takeaways

1. **Merging distant clusters is suboptimal** if the goal is to minimize MSE, as it introduces high variance.
2. **The Ward method** (used in our implementation) seeks to minimize the increase in

total variance at each step, making it a preferred choice for hierarchical clustering.

3. **Tracking MSE changes provides insights** into the structure of the data and helps refine clustering strategies.

By analyzing these MSE variations, we can make informed decisions about choosing appropriate cluster merging strategies in hierarchical clustering.