Appendix A5

Appendix for Chapter 5

This chapter appendix contains additional information for the cluster analysis dissection study in Chapter 5. Section A5.1 provides a list of the clustering methods that were examined, and shows the results for the clusters that were formed for the data on eddie. Section A5.2 shows the clusters that resulted on the workstations when the two-stage density method was used, and Section A5.3 shows the clusters that were formed for marvin using Ward’s method.

A5.1 Clustering Methods

This section presents an overview of the different clustering methods that were examined using the SAS CLUSTER procedure. Tables and graphs to show the results of different clustering methods for the data on eddie are presented.

A5.1.1 Methods Examined

The ten different clustering methods that were used with the SAS CLUSTER procedure are listed below. More information about each method may be found in the SAS/STAT user’s guide [SAS90b].

1. Single Linkage Method
2. Centroid Method
3. Average Linkage Method
4. Complete Linkage Method
5. McQuitty’s Similarity Analysis
6. Gower’s Median Method
7. Lance Williams Flexible Beta Method
8. Ward's Minimum Variance Method
9. EML Maximum Likelihood Method
10. Two-stage Density Linkage Method ($k^{th}$ Nearest Neighbour)

Table A5.1 shows the frequencies of the clusters that result when each of these clustering methods is used to cluster the preliminary clusters on eddie. The number of clusters ($N_c$) specified for each method was 10. The number of the methods provided in the table corresponds to the numbered list of methods above.

<table>
<thead>
<tr>
<th>$N_c$</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>25540</td>
<td>25511</td>
<td>25480</td>
<td>25478</td>
<td>25428</td>
<td>25380</td>
<td>25185</td>
<td>24855</td>
<td>24851</td>
<td>13346</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>19</td>
<td>31</td>
<td>22</td>
<td>79</td>
<td>127</td>
<td>175</td>
<td>269</td>
<td>269</td>
<td>5040</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6</td>
<td>19</td>
<td>22</td>
<td>19</td>
<td>148</td>
<td>235</td>
<td>239</td>
<td>2433</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>127</td>
<td>127</td>
<td>1388</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>30</td>
<td>30</td>
<td>1065</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>14</td>
<td>14</td>
<td>751</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>534</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>457</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>294</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>249</td>
</tr>
</tbody>
</table>

Table A5.1: Cluster Frequency for Each Clustering Method on Eddie

The number of entries placed in each cluster differed for each of the ten clustering methods examined. Method 1, the single-linkage method, placed almost all entries into one very large cluster, whereas method 10, the two-stage density method, placed a relatively similar number of entries into each cluster. The remaining methods produced one cluster that had a very high frequency (ranging from 99.82% of all observations for the Centroid method to 97.24% for the EML method), and relatively smaller frequencies for each of the other clusters.

A5.1.2 Choosing a Clustering Method

To determine which method would be used in our study, scatter plot graphs containing cluster membership information were examined. Scatter plots for three methods that produced significantly different clustering results are shown in Figure A5.1.
Figure A5.1: Clusters Produced using 3 Clustering Methods on Eddie
The single-linkage method is an example of a clustering method that is very good at identifying outliers, however, this method was not able to distinguish between the commands that had low resource usage. Ward's method produced one very large cluster near the origin and several smaller clusters of combined outlier commands. The clusters produced by the two-stage density method had relatively uniform frequency. Of the ten methods studied, only the two-stage density method split up the commands that are close to the origin in the scatter plot.

Of the methods presented in Figure A5.1, Ward's method and the two-stage density method are most suitable for a dissection study. Methods such as the single-linkage method that create a lot of low-frequency clusters are not very practical for a dissection study because they do not provide information about the relative similarity of commands.

The two-stage density method was chosen for our dissection study because its bias towards equal-sized clusters enables it to subdivide the entries near the origin. The high resource usage cluster (cluster 1) produced by this initial application of the two-stage density method will be further subdivided as outlined in the next section.

A5.1.3 Reclustering the Large-Resource Usage Cluster

A large resource usage cluster (cluster 1 in the two-stage density graph in Figure A5.1) results when the two-stage density method is used. The $k^{th}$ nearest neighbour two-stage density linkage method is reapplied to this large resource usage cluster to produce separate clusters of CPU-intensive and I/O-intensive commands. Figure A5.2 shows the CPU-intensive (cluster 2) and I/O-intensive (cluster 1) clusters that result when the large resource usage cluster is split into two clusters.

The two clusters shown in Figure A5.2 replace cluster 1 produced by the initial use of the two-stage density method (Figure A5.1) to compose the final set of 11 clusters that are used for the dissection analysis in Section 5.3.
A5.2 Number of Clusters

Although the number of clusters is not of primary importance in a dissection analysis, a “good” value for the number of clusters should be chosen. The SAS CLUSTER procedure provides several statistics that can be examined to determine which \( N_c \) values are good choices for a particular method. These values are recommended for the dissection analysis, as they indicate the most natural clustering for the data set.

In this section, we show which \( N_c \) values are most suitable for clustering on eddie when Ward’s method and the two-stage density method are used. The number of clusters is determined differently for the \( k^{th} \) nearest neighbour two-stage density method than for the other clustering methods outlined in Section A5.1.1.

Ward’s Method

For Ward’s method, and for the other non-density based methods, statistics are automatically generated by the CLUSTER procedure to provide information about which numbers of clusters are most desirable. Since these statistics are not always reliable, a consensus between the various statistics should be sought.

The statistics provided by the CLUSTER procedure are the \( R^2 \) statistic, the cubic
clustering criterion (CCC), the pseudo $T^2$ statistic, and the pseudo F statistic. These statistics are explained in the SAS/STAT user’s guide. Desirable $N_c$ values are indicated by local peaks in the CCC and pseudo F statistic. Small values of pseudo $T^2$ that are followed by a larger pseudo $T^2$ value for the next number of clusters are indicators of desirable $N_c$ values. The slope of the $R^2$ curve between pairs of $N_c$ values generally decreases as the $N_c$ value increases, because the variance between cluster centroids decreases; thus, it is desirable to choose $R^2$ values that fall at the beginning of a section of the $R^2$ curve that has a smaller slope.

![Figure A5.3: Statistics to Determine $N_c$ for Ward’s Method](image)

Preferred $N_c$ values to use for dissection when Ward’s method is used are 3, 9, and 17 clusters, as indicated in Figure A5.3. Since the slope of the $R^2$ curve does not change much between 9 and 17 clusters, it may be desirable to chose 9 clusters. In general, choosing a larger number of clusters produces clusters that have lower variance; however, this larger number of clusters may induce more overhead in terms of representing and/or analyzing the clusters that are produced.

**Density Method**

To determine suitable $N_c$ values to be used with the $k^{th}$ nearest neighbour two-stage density method, the number of modal clusters that are produced by the method should be examined for several different $k$ values. The most desirable $N_c$ values are those that occur for a wide range of $k$ values.

In the graph on the left in Figure A5.4, the preferred modes are indicated by long
horizontal lines, as these modes span over a wide range of $k$ values [WS82]. The graph shown on the right in Figure A5.4 displays the same information, but in a more convenient form; the peaks in this graph indicate the preferred modes. There are peaks at 6, 8, 10, 13, and 16 clusters, indicating that these would be good choices for the number of clusters.

![Graph showing number of modes and range of k values](image)

**Figure A5.4: Number of Modal Clusters for Density Method**

### A5.3 Workstations

The clusters that were produced for the dissection study of the 65 client workstations are presented in this section. Section 5.4 of Chapter 5 provides a discussion of the variability within these clusters.

Several different combinations of clustering methods were considered for the data on the 65 workstations. The two-stage density method alone provided good separation of the low resource usage jobs, but produced a very large high-variance cluster containing all medium and high resource usage jobs. Reapplying the two-stage density method to this large resource usage cluster resulted in another large cluster that did not have distinct CPU and disk block usage.

A technique that gave better results, and thus the technique that was used for our dissection study, was the application of the EML method followed by the application of the two-stage density method.¹ The data for the workstations were first divided into 10 clusters

---

¹The FASTCLUS procedure was used prior to each of these clustering methods, as the number of jobs to be analyzed was too large for the CLUSTER procedure alone; 54231 jobs were clustered by the EML method and 50486 jobs by the two-stage density method.
using the EML method. These 10 clusters were then regrouped into a low resource usage, a high CPU usage, and a high disk block usage cluster, as shown in Figure A5.5. Since the low resource usage cluster contained the majority of the observations (93.09%), this cluster was then reclustered into 9 clusters using the two-stage density method.² Figure A5.6 shows the 9 clusters produced by the two-stage density method (clusters 1-9), combined with the high CPU usage (cluster 10) and the high disk block usage (cluster 11) clusters created by the EML method. As the scale in Figure A5.6 is too large to distinguish between low resource usage jobs, the graph legend is provided for this purpose.

<table>
<thead>
<tr>
<th>cluster</th>
<th>frequency</th>
<th>cpu</th>
<th>blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50468</td>
<td>0.94</td>
<td>2.71</td>
</tr>
<tr>
<td>2</td>
<td>351</td>
<td>145.55</td>
<td>56.96</td>
</tr>
<tr>
<td>3</td>
<td>3384</td>
<td>1.57</td>
<td>87.90</td>
</tr>
<tr>
<td>total</td>
<td>54361</td>
<td>1.38</td>
<td>8.41</td>
</tr>
</tbody>
</table>

Figure A5.5: 10 EML Clusters Regrouped into 3 Clusters

A5.4 Marvin

A single application of Ward’s method produced the most favourable clustering results for the data on marvin. The two-stage density method was unable to uncover the natural split of the data near the origin. Ward’s method, however, uncovered these clusters nicely.

The scatter plot in Figure A5.7 shows cluster membership information for all data on marvin, while Figure A5.8 displays a reduced range of the data near the origin. The reduced

²The number of clusters specified with the EML method (10) and with the two-stage density method (9) were chosen as outlined in section Section A5.2.
range plot clearly shows how Ward’s method placed the entries on either side of the sparse area at approximately 0.5 CPU seconds into separate clusters.

The 12 backup commands, 2 #aufs commands, and 3 makedbm commands on marvin were the only commands that occurred in the large resource usage clusters that had low frequencies (clusters 3-10). If these clusters were too sparse to meet the goals of the dissection analysis, it might be desirable to regroup these low-frequency clusters into one cluster. It may also be desirable to reapply a clustering method to split the low resource usage cluster (cluster 2), which contains the majority of the observations (93.64%). Since the data for marvin are not examined in detail in this thesis or included in the model, no further clustering techniques were examined.

With the exception of the 17 commands that were placed into the low-frequency clusters on marvin, the workload on marvin did not have very extreme or varied resource usage. Compared to the workload on eddie and on the workstations, the workload on marvin did not demonstrate as much variance; this is likely due to the deterministic nature of the system workload on marvin.
Figure A5.7: Ward's Method Clustering on Marvin - Full Range

Figure A5.8: Ward's Method Clustering on Marvin - Reduced Range