CHAPTER 6

ANALYSIS AND DISCUSSION

6.1 INTRODUCTION

A preparatory study was performed to discover whether there was a likelihood of the existence of the knowledge sought in the data being investigated. The knowledge sought being the enhancement of characterisation of primary brain tumours in Quantitative MRI data. Precursory procedures used to make these preliminary investigations were the inspection of the sorted values of the variables, their means, standard deviations and scatterplots of the variables. The analysis obtained using these summarization methods are described in section 6.2. Then the more penetrating data-mining techniques were used to discover knowledge from the clinical *in-vivo* MR imaging data. These tools are described in chapter 4 and are Cluster Analysis, Discriminant Analysis, Decision Trees and Inductive Logic Programming. The analyses using these methods are described in sections 6.3 - 6.6.

6.2 SUMMARISATION METHODS

6.2.1 Introduction

The basic summarisation method used first was an inspection of the sorted values of the variables, their means and standard deviations. Then scatterplots of pairs of variables were used to make a visual scrutiny
of the data sets. These methods may not be conclusive in themselves, but
give an indication as to which of the variables are likely to be significant
and reveal the presence of outliers, if any. Since the data matrix is very
large, only the significant results of the summarisation methods are
reported here.

6.2.2 Appraisal of the variables

**Contrast values:** $C_w(\text{grase})$ values were determined for 88 cases, of which
16 are astrocytomas, 24 are meningiomas and 26 are pituitary adenomas.
The $C_w(\text{grase})$ values for astrocytomas are generally higher than those for
meningiomas with some overlapping of the ranges of values. The range of
values for pituitary adenomas is wider and overlaps with the ranges for
the aforementioned tumours. However, in the $T_2$-weighted FLAIR scans no
such observation could be made. $C_w(\text{flair})$ values were determined for 62
cases, of which 11 are astrocytomas, 18 are meningiomas and 19 are
pituitary adenomas.

$C_w(\text{se})$ values were determined for 44 cases, of which 12 are
astrocytomas, 19 are meningiomas and 2 are pituitary adenomas.
Astrocytomas are seen to be hypo-intense with respect to white matter
when compared to meningiomas. There is still quite a bit of overlapping.
However, in the $C_w(\text{se}')$ values, there is good demarcation between the
astrocytomas and meningiomas. Almost without exception the
astrocytomas tend to be hypo-intense, whereas the meningiomas tend to be
hyper-intense with respect to white matter. $C_w(\text{se}')$ values were determined
for 43 cases, of which 11 are astrocytomas, 19 are meningiomas and 1 is a
pituitary adenoma.

$C_w(\text{ffe})$ values were determined for 54 cases, of which 5 are
astrocytomas, 8 are meningiomas and 26 are pituitary adenomas.
Cw(ffe) values for astrocytomas, meningiomas and pituitary adenomas appear to be distributed over almost the entire range of Cw(ffe) values, not including the outliers. However the meningiomas tend to have higher contrast values post Gd-DTPA (Cw(ffe')) compared to pituitary adenomas, though with a little overlap in the ranges. Though a few pituitary adenomas have a positive contrast, no meningiomas have a negative contrast. Cw(ffe') values were determined for 56 cases, of which 7 are astrocytomas, 7 are meningiomas and 27 are pituitary adenomas.

**Intensity values:** The range of I(grase) values for pituitary adenomas and meningiomas overlapped. The astrocytomas lay at the higher end of the scale of intensity values, but there was no clear demarcation. The observations for the I(flair) values are similar.

There are no significant demarcations between the different tumours under study in the I(se) values. The I(se') values of meningiomas generally have higher values than those of astrocytomas, with some overlap.

The ranges of the I(ffe) values for meningiomas and pituitary adenomas overlap. The I(ffe) values of meningiomas tend to lie in the upper half of the range of intensity values, but those of pituitary adenomas tend to be distributed through the entire range.

**Standard deviation values:** Astrocytomas tend to have high SD(grase) values and meningiomas tend to have low SD(grase) values. The SD(grase) values for pituitary adenomas are very variable and are spread through the entire range of SD(grase) values. There is a similar tendency in SD(flair) values, although the tendency is less and some astrocytomas have a very low value.
Meningiomas generally have lower SD\(_t\)(se) and SD\(_t\)(se') values and astrocytomas generally have higher values, with a little overlap.

A wider range of SD\(_t\)(ffe) values is observed for pituitary adenomas than for meningiomas and the values for meningiomas tend to be lower. The SD\(_t\)(ffe') values do not show such behaviour.

### 6.2.3 Scatterplots

The findings from the scatterplots concur with reports in the literature regarding the behaviour of primary brain tumours in T\(_r\)-weighted, T\(_r\)-weighted and gadolinium contrast (Gd-DTPA) enhanced T\(_r\)-weighted images. This is indicative of the validity of the data collected in that it conforms to the known reported behaviour of normal and diseased brain parenchyma in MR images.

The scatterplot of I(grase) versus I(se) for the Data Set RP-GS is shown in figure 6.1. The well-known fact that usually I\(_t\) > I\(_w\) and I\(_t\) > I\(_g\) in T\(_r\)-weighted images can be visualised in this scatterplot. This is the reason T\(_r\)-weighted images are routinely used for detecting neoplasms.

The scatterplot of I(se) versus I(se') for the Data Set RP-GS is shown in figure 6.2. Two groups can be discerned - the normal tissues and the tumour tissues. In general, the intensity of tumours tends to be less than that of normal tissue before Gd-DTPA is administered and greater after administration. The intensity of tumour tissue is generally increased after administration of Gd-DTPA. All tumours do not enhance equally.

The scatterplot of I(grase) versus I(ffe) for the Data Set RP-GF is shown in figure 6.3. The cluster of the tumour tissues and the cluster of
Figure 6.1 Scatterplot of $I_{\text{grase}}$ vs $I_{\text{se}}$ for Data Set RP-GS
Figure 6.2 Scatterplot of $I(\text{se})$ vs $I(\text{se}')$ for Data Set RP-GS
Figure 6.3  Scatterplot of I(grade) vs I(ffe) for Data Set RP-GF
the normal tissues overlap to a greater extent in this scatterplot than in the corresponding scatterplot for the T₁-weighted SE scan (Figure 6.1). However, the populations of tumours in the two data sets follow different distributions for clinical reasons, and so it would be incorrect to infer from these scatterplots that T₁-weighted SE gives different information from T₁-weighted FFE. Grey matter and white matter are separated better in figure 6.3 than in figure 6.1.

The scatterplot of I(ffe) versus I(ffe') for the Data Set RP-GF is shown in figure 6.4. It is similar to Figure 6.2. An outlying white matter value is observed. This does not conform to the likely range of values for white matter signal intensities and therefore this sample was not taken into consideration in analysing the remaining scatterplots and in the cluster analysis. The sample is a case of histologically verified craniopharyngioma and therefore does not affect the analyses performed using the other three methods, since those methods were used for classification of meningiomas and pituitary adenomas.

The scatterplot of I(grase) versus I(flair) for the Data Set P-GIS + P-GIF is shown in figure 6.5. The T₂-weighted GRASE scan clearly shows a greater separation of the clusters of tumour, white and grey matter than the T₂-weighted FLAIR scan. The values of tumour and normal tissue overlap to a great extent in the T₂-weighted FLAIR scan. The outlier in this scatterplot is another craniopharyngioma. The intensities of craniopharyngiomas in this study conform to the reports in the literature (Osborn 1994) of widely ranging values of I in craniopharyngiomas.

The scatterplot shown in figure 6.6 plots the intensities (I(grase) vs I(se)) and in figure 6.7 plots the contrast with white matter (Cₙ(grase) vs Cₙ(se)) for the Data Set RP-GS and for the same pair of scans, T₂-weighted GRASE and T₁-weighted SE. The advantage of analysing C
Figure 6.4 Scatterplot of I(ffe) vs I(ffe') for Data Set RP-GF
Figure 6.5 Scatterplot of $I_{\text{grase}}$ vs $I_{\text{flair}}$ for Data Set P-GIS + P-GIF
Figure 6.6 Scatterplot of $I_t(\text{grase})$ vs $I_t(\text{se})$ for Data Set RP-GS
Figure 6.7 Scatterplot of $C_w(\text{grase})$ vs $C_w(\text{se})$ for Data Set RP-GS
rather than I is immediately apparent, since the separation of the meningioma cluster and the astrocytoma cluster is greater by far in the scatterplot of contrasts than in the scatterplot of intensities. Figure 6.7 also shows that when the contrast values in one T1-weighted scan are used in conjunction with the contrast values in the T2-weighted scan, rather than each in isolation, the qualitative information about the probable diagnostic outcome is increased. This agrees with the findings in the literature (Osborn 1994).

The scatterplot shown in figure 6.8 plots the intensities (I1(se) vs I1(se')) and in figure 6.9 plots the contrast with white matter (Cw(se) vs Cw(se')) for the Data Set RP-GS and for the same pair of scans, T1-weighted SE before and after Gd-DTPA administration. Figures 6.8 and 6.9 show the known fact that meningiomas enhance differently from astrocytomas after Gd-DTPA administration. There is a small range of overlapping values in the scatterplot of intensities that actually disappears completely in the scatterplot of contrast values. Hence, if the selected four methods of analysis are to have any degree of utility in data mining from quantitative MRI data, they should be successful in the simpler application of characterisation of astrocytomas and meningiomas.

The scatterplot shown in figure 6.10 plots the intensities (I1(grase) vs I1(ffe)) and in figure 6.11 plots the contrast with white matter (Cw(grase) vs Cw(ffe)) for the Data Set RP-GF and for the same pair of scans, T2-weighted GRASE and T1-weighted FFE. Both these scatterplots show overlapping clusters of meningiomas and pituitary adenomas. This reflects the difficulties in discriminating between meningiomas and pituitary adenomas based on known methods of qualitative analysis. The pituitary adenomas show a greater range of contrast values both in T1-weighted as well as T2-weighted images. The outlier is a pituitary macroadenoma which in the MR image showed a very heterogeneous range of
Figure 6.8 Scatterplot of $I(se)$ vs $I(se')$ for Data Set RP-GS
Figure 6.9 Scatterplot of $C_w(se)$ vs $C_w(se')$ for Data Set RP-GS
Figure 6.10 Scatterplot of $I_{tg} (grase)$ vs $I_{tg} (ffe)$ for Data Set RP-GF
Figure 6.11 Scatterplot of $C_w(\text{grase})$ vs $C_w(\text{ffe})$ for Data Set RP-GF
intensities. The other outlier is a pituitary adenoma which in qualitative assessment of the scanned image showed that there was bleeding into the tumour region. Scatterplots highlight such outliers.

The scatterplot shown in figure 6.12 plots the intensities \((I_{\text{ffe}})\) vs \((I'_{\text{ffe}})\) and in figure 6.13 plots the contrast with white matter \((C_{\text{ffe}})\) vs \((C'_{\text{ffe}})\) for the Data Set RP-GF and for the same pair of scans, \(T_1\)-weighted FFE before and after Gd-DTPA administration. Both these scatterplots also show overlapping clusters of meningiomas and pituitary adenomas.

The scatterplot shown in figure 6.14 plots the intensities \((I_{\text{grase}})\) vs \((I_{\text{flair}})\) and in figure 6.15 plots the contrast with white matter \((C_{\text{grase}})\) vs \((C_{\text{flair}})\) for the Data Set P-GIS + P-GIF and for the same pair of scans, \(T_2\)-weighted GRASE and \(T_2\)-weighted FLAIR. The ranges of \(I\) values of pituitary adenomas and meningiomas in \(T_2\)-weighted FLAIR scans are almost identical, whereas in the \(T_2\)-weighted GRASE scans, the range for meningiomas is slightly more than that for pituitary adenomas. The scatterplot of contrasts for the same pair of scans reveals the same information. Very little information is visualisable in these two scatterplots of two \(T_2\)-weighted images for the discrimination between pituitary adenomas and meningiomas and there is not much difference in the information visualisable in these two scans.

### 6.2.4 Conclusions

The summarisation methods used were appraisal of the sorted values of the variables and visualisation using scatterplots. These indicated that obvious and separate clusters for meningiomas and pituitary adenomas are not revealed in the quantitative data. This finding is supported by the reports in the literature (Just et al., 1988).
Figure 6.12 Scatterplot of $I_{(ffe)}$ vs $I_{(ffe')}$ for Data Set RP-GF
Figure 6.13 Scatterplot of $C_w(ffe)$ vs $C_w(ffe')$ for Data Set RP-GF
Figure 6.14 Scatterplot of $I_{(grase)}$ vs $I_{(flair)}$ for Data Set P-GIS + P-GIF
Figure 6.15 Scatterplot of $C_w$ (grase) vs $C_w$ (flair) for Data Set P-GIS + P-GIF
These summarisation methods also indicated that meningiomas and astrocytomas may be differentiated more easily but there could be occasions where one could be mistaken for the other. This agrees with the known behaviour of meningiomas and astrocytomas in MR images. (Osborn 1994).

The scatterplots gave a better indication as to which variables were likely to be of significance in classification. The scatterplot that offered the maximum discrimination between meningiomas and astrocytomas was \( C_w(\text{se}') \) in conjunction with \( C_w(\text{se}) \) and the scatterplot that offered the maximum discrimination between meningiomas and pituitary adenomas was \( C_w(\text{ffe}') \) in conjunction with \( C_w(\text{ffe}) \). These scatterplots show that pairs of variables offer more discriminative information than the individual variables taken on their own. This conforms to the knowledge used in accepted radiological practice that sets of scans with different weightings of relaxation times, used in conjunction are useful for differential diagnosis.

6.3 CLUSTER ANALYSIS

6.3.1 Introduction

The Cluster Analysis method has been described in section 5.3. K-Means analysis was performed on the data set \( N \) to cluster the normal grey and white matter in brain tissues. Successful characterisation of normal tissues has been reported (Siromoney 1993, Siromoney et al., 1993a). Cluster analysis was performed on the data set \( P\text{-GIS} \) to test for clustering of meningiomas and astrocytomas. It was then used on the data set \( P\text{-GIF} \) to test for clustering of meningiomas and pituitary adenomas. Section 6.3.2 describes the clustering of meningiomas and astrocytomas and section 6.3.3 that of meningiomas and pituitary adenomas.
6.3.2 Clustering of astrocytomas and meningiomas

K-Means cluster analysis was performed on Data Set P-GIS containing 30 cases. The variables used were \( I_t(\text{grase}) \), \( \text{SD}_t(\text{grase}) \), \( C_w(\text{grase}) \), \( C_g(\text{grase}) \), \( C_c(\text{grase}) \), \( D_w(\text{grase}) \), \( D_g(\text{grase}) \), \( D_c(\text{grase}) \), and similarly in the T\(_2\)-weighted FLAIR, T\(_1\)-weighted SE (before and after administering gadolinium contrast agent) scans, and \( C_{gad}, D_{gad} \). Thus 34 variables in all were used.

Partitioning was performed for \( K=2, 3, 4 \), etc., clusters. The number of clusters for which the partitioning was most meaningful was three. When two clusters were formed, one astrocytoma and a cystic lesion separated into one cluster and the remaining 28 cases into the other cluster. When three clusters were formed, the major reallocation took place, with the astrocytoma and cystic lesion still remaining together in one cluster (cluster no. 3); 6 astrocytomas, 1 meningioma and 1 schwannoma in another cluster (cluster no. 1), and the remaining cases including one astrocytoma in another cluster (cluster no.2). Further division into a greater number of clusters produced redistributions that were not meaningful. Removal of outliers did not change the performance of clustering. The final cluster centres for \( K=3 \) clusters are as shown in the table 6.1. The distances in the 34-dimensional space between the final cluster centres were determined. They are summarised as shown in table 6.2.

The number of cases in each cluster is shown in table 6.3. There are 8 samples in cluster 1, 20 in cluster 2 and two in cluster 3. The case listing of cluster membership is as shown in the table 6.4.
Table 6.1 Final cluster centres for K=3 clusters for the Data Set P-GIS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>FLCONC</th>
<th>FLCONG</th>
<th>FLCONW</th>
<th>FLDIFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8304.1250</td>
<td>396.0000</td>
<td>522.8750</td>
<td>1303.3750</td>
</tr>
<tr>
<td>2</td>
<td>6635.2000</td>
<td>175.7000</td>
<td>352.7000</td>
<td>994.0000</td>
</tr>
<tr>
<td>3</td>
<td>2314.5000</td>
<td>-257.0000</td>
<td>-259.0000</td>
<td>549.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>FLDIFG</th>
<th>FLDIFW</th>
<th>FLSDT</th>
<th>FLSIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>351.0000</td>
<td>425.5000</td>
<td>137.7500</td>
<td>1465.0000</td>
</tr>
<tr>
<td>2</td>
<td>156.7500</td>
<td>292.2500</td>
<td>85.5000</td>
<td>1144.8500</td>
</tr>
<tr>
<td>3</td>
<td>-237.5000</td>
<td>-182.5000</td>
<td>91.5000</td>
<td>736.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>GADCON</th>
<th>GADDIF</th>
<th>GRCONC</th>
<th>GRCONG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>206.6250</td>
<td>114.7500</td>
<td>-93.1250</td>
<td>688.1250</td>
</tr>
<tr>
<td>2</td>
<td>657.2000</td>
<td>404.5500</td>
<td>-363.7500</td>
<td>265.9000</td>
</tr>
<tr>
<td>3</td>
<td>69.0000</td>
<td>43.0000</td>
<td>49.5000</td>
<td>902.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>GRCONW</th>
<th>GRDIFC</th>
<th>GRDtFG</th>
<th>GRDIFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1377.8750</td>
<td>-179.6250</td>
<td>698.1250</td>
<td>1000.2500</td>
</tr>
<tr>
<td>2</td>
<td>568.6500</td>
<td>-780.0000</td>
<td>219.2500</td>
<td>418.2500</td>
</tr>
<tr>
<td>3</td>
<td>1841.0000</td>
<td>96.5000</td>
<td>950.0000</td>
<td>1290.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>GRSIT</th>
<th>POCONC</th>
<th>POCONG</th>
<th>POCONW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1731.0000</td>
<td>2082.3750</td>
<td>11.8750</td>
<td>-171.1250</td>
</tr>
<tr>
<td>2</td>
<td>1160.1000</td>
<td>3739.3000</td>
<td>479.3000</td>
<td>214.1500</td>
</tr>
<tr>
<td>3</td>
<td>2004.0000</td>
<td>109.0000</td>
<td>-425.0000</td>
<td>-511.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>PODIFC</th>
<th>PODIFG</th>
<th>PODIFW</th>
<th>PRDIFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>502.1250</td>
<td>1.3750</td>
<td>-173.2500</td>
<td>352.3750</td>
</tr>
<tr>
<td>2</td>
<td>816.7000</td>
<td>334.1000</td>
<td>181.2500</td>
<td>396.0000</td>
</tr>
<tr>
<td>3</td>
<td>62.0000</td>
<td>-362.0000</td>
<td>-459.5000</td>
<td>5.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>PRDIFG</th>
<th>PRDIFW</th>
<th>PRSDT</th>
<th>PRSIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-117.5000</td>
<td>-333.5000</td>
<td>68.3750</td>
<td>642.6250</td>
</tr>
<tr>
<td>2</td>
<td>-71.2500</td>
<td>-241.8000</td>
<td>40.2500</td>
<td>632.5500</td>
</tr>
<tr>
<td>3</td>
<td>-307.5000</td>
<td>-476.5000</td>
<td>54.5000</td>
<td>368.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>PRCONW</th>
<th>GRSDT</th>
<th>POSDT</th>
<th>POSIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-336.0000</td>
<td>199.7500</td>
<td>125.7500</td>
<td>757.3750</td>
</tr>
<tr>
<td>2</td>
<td>-270.9500</td>
<td>103.3000</td>
<td>76.5500</td>
<td>1037.1000</td>
</tr>
<tr>
<td>3</td>
<td>-558.5000</td>
<td>99.5000</td>
<td>98.0000</td>
<td>411.5000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster</th>
<th>PRCONC</th>
<th>PRCONG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1205.8750</td>
<td>-139.7500</td>
</tr>
<tr>
<td>2</td>
<td>1717.5500</td>
<td>-98.6000</td>
</tr>
<tr>
<td>3</td>
<td>-6.5000</td>
<td>-457.0000</td>
</tr>
</tbody>
</table>
Table 6.2 Distances between the final cluster centres in the Data Set P-GIS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3063.6118</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6778.3650</td>
<td>6731.9729</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 6.3 Number of cases in each cluster in Data Set P-GIS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.0</td>
</tr>
<tr>
<td>2</td>
<td>20.0</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Table 6.4 Case listing of cluster membership in Data Set P-GIS

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Case ID</th>
<th>Disease code</th>
<th>Cluster</th>
<th>Distance from cluster centroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>a</td>
<td>1</td>
<td>2320.661</td>
</tr>
<tr>
<td>2</td>
<td>103</td>
<td>te</td>
<td>2</td>
<td>2575.906</td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>x</td>
<td>3</td>
<td>2112.721</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
<td>t</td>
<td>2</td>
<td>1145.811</td>
</tr>
<tr>
<td>5</td>
<td>114</td>
<td>m</td>
<td>2</td>
<td>1936.953</td>
</tr>
<tr>
<td>6</td>
<td>115</td>
<td>m</td>
<td>2</td>
<td>1268.388</td>
</tr>
<tr>
<td>7</td>
<td>117</td>
<td>m</td>
<td>2</td>
<td>1449.360</td>
</tr>
<tr>
<td>8</td>
<td>119</td>
<td>a</td>
<td>1</td>
<td>1658.337</td>
</tr>
<tr>
<td>9</td>
<td>120</td>
<td>m</td>
<td>2</td>
<td>1877.217</td>
</tr>
<tr>
<td>10</td>
<td>121</td>
<td>a</td>
<td>2</td>
<td>1426.554</td>
</tr>
<tr>
<td>11</td>
<td>122</td>
<td>p</td>
<td>2</td>
<td>938.985</td>
</tr>
<tr>
<td>12</td>
<td>124</td>
<td>m</td>
<td>2</td>
<td>1644.313</td>
</tr>
<tr>
<td>13</td>
<td>125</td>
<td>a</td>
<td>3</td>
<td>2275.238</td>
</tr>
<tr>
<td>14</td>
<td>126</td>
<td>m</td>
<td>2</td>
<td>1335.770</td>
</tr>
<tr>
<td>15</td>
<td>128</td>
<td>a</td>
<td>1</td>
<td>1563.713</td>
</tr>
<tr>
<td>16</td>
<td>129</td>
<td>a</td>
<td>1</td>
<td>3007.160</td>
</tr>
<tr>
<td>17</td>
<td>130</td>
<td>m</td>
<td>2</td>
<td>1764.282</td>
</tr>
<tr>
<td>18</td>
<td>131</td>
<td>s</td>
<td>2</td>
<td>1293.810</td>
</tr>
<tr>
<td>19</td>
<td>132</td>
<td>s</td>
<td>1</td>
<td>2167.767</td>
</tr>
<tr>
<td>20</td>
<td>134</td>
<td>s</td>
<td>2</td>
<td>1314.278</td>
</tr>
<tr>
<td>21</td>
<td>137</td>
<td>m</td>
<td>2</td>
<td>2183.537</td>
</tr>
<tr>
<td>22</td>
<td>141</td>
<td>m</td>
<td>2</td>
<td>1151.291</td>
</tr>
<tr>
<td>23</td>
<td>147</td>
<td>a</td>
<td>1</td>
<td>2105.679</td>
</tr>
<tr>
<td>24</td>
<td>148</td>
<td>s</td>
<td>2</td>
<td>2173.664</td>
</tr>
<tr>
<td>25</td>
<td>151</td>
<td>a</td>
<td>1</td>
<td>2202.115</td>
</tr>
<tr>
<td>26</td>
<td>153</td>
<td>m</td>
<td>1</td>
<td>2798.872</td>
</tr>
<tr>
<td>27</td>
<td>154</td>
<td>nl</td>
<td>2</td>
<td>802.048</td>
</tr>
<tr>
<td>28</td>
<td>156</td>
<td>s</td>
<td>2</td>
<td>1180.143</td>
</tr>
<tr>
<td>29</td>
<td>159</td>
<td>m</td>
<td>2</td>
<td>1549.046</td>
</tr>
<tr>
<td>30</td>
<td>165</td>
<td>m</td>
<td>2</td>
<td>1334.660</td>
</tr>
</tbody>
</table>
The clusters can be summarised as shown below.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Tumour Code and No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>a: 6; m: 1; s: 1</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>a: 1; m: 11; p: 2; s: 4; te: 1; nl: 1</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>a: 1; x: 1</td>
</tr>
</tbody>
</table>

Six out of 8 astrocytomas were grouped in a separate cluster. Eleven out of twelve meningiomas appeared in the second cluster. The third cluster contained an astrocytoma but no meningiomas.

The single astrocytoma in cluster 3 is found to be a sample from a post-operative patient. Post-surgical tissue changes affect the appearance and intensities in the MRI. So it is meaningful that this case did not become clustered with the other astrocytomas (or even meningiomas). Thus K-Means cluster analysis succeeded in clustering astrocytomas and meningiomas in this data set.

### 6.3.3 Clustering of pituitary adenomas and meningiomas

Cluster Analysis was performed on the Data Set P-GIF to seek for clustering of pituitary adenomas and meningiomas. This data set contained 25 samples, of which 15 were pituitary adenomas and 4 were meningiomas. There were also 3 craniopharyngiomas and one each of astrocytoma, neurocytoma and primitive neuroectodermal tumour. Partitioning for 2, 3, etc., ... upto 10 clusters were tried, but no meaningful partition was observed. The results of these 10 attempts at clustering are shown in table 6.5.
Table 6.5  Summary of results of partitioning Data Set P-GIF into K=2,3,...,10 clusters

<table>
<thead>
<tr>
<th>For No. of clusters → formed</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>Cluster Affiliation of Sample ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumour Code ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>a2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>nc</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>c</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>pe</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
From the table, the following can be observed:

For K = 2 clusters: All the tumours are in one cluster, except a solitary craniopharyngioma that is in the second cluster.

For K = 3 clusters: The same craniopharyngioma now shifts into the first cluster, a solitary pituitary adenoma lies in the third cluster and all the rest of the samples are clustered together in the second group.

For K = 4 clusters: Now the first cluster contains the solitary craniopharyngioma, the second another craniopharyngioma, the third the pituitary adenoma and all the rest of the samples are in the fourth cluster.

For K = 5 clusters: The craniopharyngioma continues to occupy the first cluster alone. The pituitary adenoma shifts to the third cluster again and alone again. The third cluster contains a single meningioma, the fourth the other craniopharyngioma and the fifth all the remaining samples.

For K = 6 - 10 clusters: the larger groups now begin to disintegrate without any recognisable pattern. Pituitary adenomas and meningiomas are found in the same cluster and also in all the large clusters. Some of them appear as solitary samples in a cluster.

To sum up: partitioning for 2, 3, ..., 10 clusters were tried, but no meaningful partition was observed. For partitioning into 2, 3, ... 6 clusters most of the samples remained in one large cluster and a few less common tumours like craniopharyngiomas occurred in the other cluster. For 7 -10 groups the pituitary adenomas and meningiomas were still found together in several of the groups. In other words no meaningful clusters were obtained.
Thus from the table 6.5 it is seen that the 15 pituitary adenomas and 4 meningiomas did not become clustered and so were not classifiable by K-Means cluster analysis. If clustering exists, an indication of meaningful partition would normally be obtained when the data set is partitioned into two groups. So statistically valid results were not observable for distinguishing between meningiomas and pituitary adenomas by use of K-means Cluster Analysis of the signal intensities, contrast and other calculated properties of the image. The findings in literature (Just et al., 1988) indicate that $T_1$ and $T_2$ values of meningiomas and pituitary adenomas also overlap.

K-Means analysis was then performed on the Data Set RP-GF to seek for clustering of pituitary adenomas and meningiomas. In this analysis also partitioning was attempted for $K = 2, 3, ..., 10$ clusters. These results are summarised in the table 6.6.

The salient features of this analysis are very similar to the analysis for the data set P-GIF given above. (Compare with the previous table 6.5). Again the pituitary adenomas and meningiomas are found in the same cluster or in several clusters. Once again the removal of possible outliers did not improve the clustering.

Thus the results did not appear to be statistically significant. Clustering was also attempted by removing the outliers prior to the analysis. However, natural clusters were not formed inspite of this.

6.3.4 Conclusions

K-Means clustering was successful in distinguishing between astrocytomas and meningiomas. Six out of seven astrocytomas were clustered correctly and 11 out of 12 meningiomas were clustered correctly.
Table 6.6 Results of partitioning Data Set RP-GF into K=2,3,...,10 clusters

<table>
<thead>
<tr>
<th>K = →</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>Tumour Code ↓</td>
<td>m</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>s</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>pp</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>hh</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>nc</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>pe</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
K-Means clustering is an unsupervised learning tool. Tissue characterisation between astrocytomas and meningiomas is enhanced by the use of K-Means Cluster Analysis for mining clinical images.

Separate clusters of meningiomas and pituitary adenomas were not observed. K-Means clustering did not enhance tissue characterisation between pituitary adenomas and meningiomas in this data set, as the clusters contained mixed categories of tumour samples. Division of the entities into $K = 2, 3, 4, \ldots$ up to 10 clusters did not produce significant clustering either before or after the removal of outliers. To summarise, in this data set the data mining tool K-Means Cluster Analysis was unable to mine the knowledge sought (discrimination between meningiomas and pituitary adenomas). This is due to the absence of natural clusters in the values of the variables presented for analysis in this unsupervised learning environment.

6.4 DISCRIMINANT ANALYSIS

6.4.1 Introduction

The data was partitioned in a supervised learning environment using Discriminant Analysis. Fisher's method was used for separating tumour samples into distinct classes. Discriminant analysis using Fisher's method was performed on the data set P-GIS to test for discrimination between meningiomas and astrocytomas; and on the data sets P-GIF and RP-GF to test for discrimination between meningiomas and pituitary adenomas.

6.4.2 Classification of astrocytomas and meningiomas

Discriminant analysis was used on the astrocytomas and meningiomas in the Data Set P-GIS. T-test of equality of means for the
individual variables was performed and it was found that the following 29 variables were significant: $D_c(\text{flair})$, $SD_t(\text{flair})$, $I(\text{flair})$, $C_{\mu \text{flair}}$, $D_{\mu \text{flair}}$, $C_c(\text{grase})$, $C_g(\text{grase})$, $C_w(\text{grase})$, $D_c(\text{grase})$, $D_g(\text{grase})$, $SD_t(\text{grase})$, $I(\text{grase})$, $C_c(\text{se'})$, $C_g(\text{se'})$, $C_w(\text{se'})$, $D_c(\text{se'})$, $D_g(\text{se'})$, $SD_t(\text{se'})$, $I(\text{se'})$, $C_c(\text{se})$, $C_g(\text{se})$, $C_w(\text{se})$, $D_c(\text{se})$, $D_g(\text{se})$, $SD_t(\text{se})$, $I(\text{se})$.

The prior probability for each group (astrocytomas and meningiomas) was taken to be 0.5. Out of the above 29 variables, the following 12 variables failed the tolerance test: $C_c(\text{se'})$, $D_c(\text{se'})$, $D_g(\text{se'})$, $SD_t(\text{se'})$, $I(\text{se'})$, $C_c(\text{se})$, $C_g(\text{se})$, $D_c(\text{se})$, $D_g(\text{se})$, $D_w(\text{se})$, $SD_t(\text{se})$, $I(\text{se})$. Fisher's classification coefficients were determined for those that passed the tolerance test. The coefficients of Fisher's linear discriminant functions for these variables that passed the tolerance test are shown in table 6.7.

All the astrocytomas were correctly classified as astrocytomas by Fisher's classification function. All meningiomas were correctly classified as meningiomas. No astrocytoma was classified as a meningioma and no meningioma was classified as an astrocytoma. The results of the classification are summarised in table 6.8.

When the discriminant analysis program was allowed to select the statistically significant variables stepwise, by minimising Wilks' Lambda, from the variables found to be significant by the t-test, the following variables were found to be significant: $C_g(\text{se'})$, $D_g(\text{grase})$. The significant variables selected by the program are given in table 6.9. Fisher's classification coefficients for each of these were determined and are shown in table 6.10.

The groups formed by using these variables in Discriminant Analysis correctly grouped all the astrocytomas and meningiomas. The
Table 6.7 Fisher's coefficients for classification of astrocytomas and meningiomas in Data Set P-GIS using 17 variables.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>FISHER'S COEFFICIENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Program Code</td>
</tr>
<tr>
<td>$D_{c}(flair)$</td>
<td>FLDIFC</td>
</tr>
<tr>
<td>$SD_{c}(flair)$</td>
<td>FLSDT</td>
</tr>
<tr>
<td>$I_{c}(flair)$</td>
<td>FLSIT</td>
</tr>
<tr>
<td>$C_{gad}$</td>
<td>GADCON</td>
</tr>
<tr>
<td>$D_{gad}$</td>
<td>GADDIF</td>
</tr>
<tr>
<td>$C_{c}(grase)$</td>
<td>GRCONC</td>
</tr>
<tr>
<td>$C_{g}(grase)$</td>
<td>GRCONG</td>
</tr>
<tr>
<td>$C_{w}(grase)$</td>
<td>GRCONW</td>
</tr>
<tr>
<td>$D_{c}(grase)$</td>
<td>GRDIFC</td>
</tr>
<tr>
<td>$D_{g}(grase)$</td>
<td>GRDIFG</td>
</tr>
<tr>
<td>$D_{w}(grase)$</td>
<td>GRDIFW</td>
</tr>
<tr>
<td>$SD_{c}(grase)$</td>
<td>GRSDT</td>
</tr>
<tr>
<td>$I_{c}(grase)$</td>
<td>GRSIT</td>
</tr>
<tr>
<td>$C_{c}(se')$</td>
<td>POCONG</td>
</tr>
<tr>
<td>$C_{w}(se')$</td>
<td>POCONW</td>
</tr>
<tr>
<td>$D_{c}(se')$</td>
<td>PODIFG</td>
</tr>
<tr>
<td>$C_{w}(se)$</td>
<td>PRCONW</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.8 Results of classification of astrocytomas and meningiomas in Data Set P-GIS using 17 variables.

<table>
<thead>
<tr>
<th>Actual Group</th>
<th>No. Of Cases</th>
<th>Predicted Group Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Group 2</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0%</td>
</tr>
<tr>
<td>Ungrouped</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40.0%</td>
</tr>
</tbody>
</table>

Table 6.9 Results of the stepwise selection of variables for classification of astrocytomas and meningiomas in Data Set P-GIS.

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
<th>Vars Entered</th>
<th>Wilks’ Lambda</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cg(se')</td>
<td>POCONG</td>
<td>0.17826</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>Dw(grase)</td>
<td>GRDIFW</td>
<td>0.14154</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Table 6.10  Fisher's coefficients for classification of astrocytomas and meningiomas in Data Set P-GIS using the significant variables selected stepwise.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Fisher's Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Program Code</td>
</tr>
<tr>
<td>Dw(grase)</td>
<td>GRDIFW</td>
</tr>
<tr>
<td>Cg(se')</td>
<td>POCONG</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.11 Results of classification of astrocytomas and meningiomas in Data Set P-GIS using the significant variables selected stepwise.

<table>
<thead>
<tr>
<th>Actual Group</th>
<th>No. Of Cases</th>
<th>Predicted Group Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Group 1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Group 2</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0%</td>
</tr>
<tr>
<td>Ungrouped</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.0%</td>
</tr>
</tbody>
</table>
results of the classification were again 100% and are summarised in table 6.11.

6.4.3 Classification of pituitary adenomas and meningiomas

**Data set P-GIF:** Discriminant analysis was used on the Data Set P-GIF. This set contained 25 cases. All pituitary adenomas (15 cases) were labelled as group 1, meningiomas (4 cases) as group 2 and the rest of the tumours (6 cases) as group 3. Discriminant analysis was run using only group 1 (pituitary adenomas) and group 2 (meningiomas) and achieved 100% discrimination.

T-test of equality of means for the individual variables was performed and it was found that the following variables were significant: $C_w(\text{grase})$, $D_t(\text{grase})$, $D_v(\text{grase})$, $I_1(\text{grase})$, $C_s(\text{ffe'})$, $C_s(\text{ffe'})$, $C_s(\text{ffe'})$, $D_s(\text{ffe'})$, $D_s(\text{ffe'})$, $I_1(\text{ffe'})$, SD$_t$(ffe).

The prior probability for each group (pituitary adenomas and meningiomas) was taken to be 0.5. Out of the above 12 variables, one variable ($I_1(\text{ffe'}))$ failed the tolerance test. Fisher's classification coefficients were determined for those that passed the tolerance test. The coefficients of Fisher's linear discriminant functions for these variables that passed the tolerance test are shown in table 6.12. The groups formed by Fisher's classification function correctly classified all the pituitary adenomas and meningiomas. The results of the classification are summarised in table 6.13.

When the discriminant analysis program was allowed to select the statistically significant variables stepwise, by minimising Wilks' Lambda, from the variables found to be significant by the T-test, the following variables were found to be significant: $C_s(\text{ffe'})$, $C_s(\text{grase})$, SD$_t$(ffe). This is shown in table 6.14. Fisher's classification coefficients for each of
Table 6.12 Fisher's classification coefficients for classification of pituitary adenomas and meningiomas in Data Set P-GIF using 11 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Fisher's Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Program Code</td>
</tr>
<tr>
<td>Cw(grase)</td>
<td>GRCONW</td>
</tr>
<tr>
<td>Dg(grase)</td>
<td>GRDIFG</td>
</tr>
<tr>
<td>Dw(grase)</td>
<td>GRDIFW</td>
</tr>
<tr>
<td>It(grase)</td>
<td>GRSIT</td>
</tr>
<tr>
<td>Cc(ffe')</td>
<td>POCONC</td>
</tr>
<tr>
<td>Cg(ffe')</td>
<td>POCONG</td>
</tr>
<tr>
<td>Cw(ffe')</td>
<td>POCONW</td>
</tr>
<tr>
<td>Dc(ffe')</td>
<td>PODIFC</td>
</tr>
<tr>
<td>Dg(ffe')</td>
<td>PODIFG</td>
</tr>
<tr>
<td>Dw(ffe')</td>
<td>PODIFW</td>
</tr>
<tr>
<td>SDt(ffe)</td>
<td>PRSDT</td>
</tr>
<tr>
<td>(Constant)</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.13 Results of classification of pituitary adenomas and meningiomas in Data Set P-GIF using 11 variables.

<table>
<thead>
<tr>
<th>Actual Group</th>
<th>No. of Cases</th>
<th>Predicted Group Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Group 1</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Group 2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Ungrouped cases</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>100.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 6.14 Results of the stepwise selection of variables for classification of pituitary adenomas and meningiomas in Data Set P-GIF.

<table>
<thead>
<tr>
<th>Step</th>
<th>Action</th>
<th>Vars In</th>
<th>Wilks' Lambda</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Removed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C_ff_else</td>
<td>POCONW</td>
<td>1</td>
<td>0.50146</td>
</tr>
<tr>
<td>2</td>
<td>C_grapse</td>
<td>GRCONW</td>
<td>2</td>
<td>0.22179</td>
</tr>
<tr>
<td>3</td>
<td>SD_ff_else</td>
<td>PRSDT</td>
<td>3</td>
<td>0.16018</td>
</tr>
</tbody>
</table>
these were determined and are shown in table 6.15. The groups formed by Fisher's classification function correctly classified all the pituitary adenomas and meningiomas. The results of the classification are summarised in table 6.16.

**Data set RP-GF:** Discriminant analysis was used on the Data Set RP-GF. There were 40 cases. T-test of equality of means for the individual variables was performed and it was found that the following variables were significant: Cw(ffe'), C_g(ffe'), D_w(ffe'), D_g(ffe'), SD_{ffe}.

Discriminant analysis was then done using these five significant variables. The probability of a sample belonging to either group was taken to be equal. Fisher's classification coefficients for each of these were determined and are shown in table 6.17. 89% accuracy of discrimination was achieved. Out of the 22 pituitaries and 6 meningiomas, two pituitary adenomas were misclassified as meningiomas, and one meningioma was misclassified as pituitary adenoma. The results of the classification are summarised in table 6.18.

Ten-fold cross validation was then performed, i.e., one-tenth of the data set was removed at a time and set as unknown cases. Discriminant analysis was performed with the remaining nine-tenths and the resultant Fisher's classification function was used on the unknown 1/10th. The overall percentage of accuracy in predicting the unknowns was 75%. However the accuracy of prediction of the knowns also ranged from 77 - 92%. The results of the ten-fold cross-validation on the classification are shown in table 6.19.
Table 6.15 Fisher's coefficients for classification of pituitary adenomas and meningiomas in Data Set P-GIF using the significant variables selected stepwise.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Fisher's Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Program Code</td>
</tr>
<tr>
<td>$C_w(\text{grase})$</td>
<td>GRCONW</td>
</tr>
<tr>
<td>$C_w(ffe')$</td>
<td>POCONW</td>
</tr>
<tr>
<td>$SD(ffe)$</td>
<td>PRSDT</td>
</tr>
<tr>
<td>(Constant)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.16 Results of classification of pituitary adenomas and meningiomas in Data Set P-GIF using the significant variables selected stepwise.

<table>
<thead>
<tr>
<th>Actual Group</th>
<th>No. of Cases</th>
<th>Predicted Group Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Group 1</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Group 2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0%</td>
</tr>
<tr>
<td>Ungrouped cases</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>83.3%</td>
</tr>
</tbody>
</table>
Table 6.17 Fisher’s classification coefficients for classification of pituitary adenomas and meningiomas in Data Set RP-GF using 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Program Code</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_i(ffe')$</td>
<td>POCONG</td>
<td>-0.0081982</td>
<td>-0.0006639</td>
</tr>
<tr>
<td>$C_s(ffe')$</td>
<td>POCONW</td>
<td>0.0188624</td>
<td>0.0333181</td>
</tr>
<tr>
<td>$D_i(ffe')$</td>
<td>PODIFG</td>
<td>0.0469598</td>
<td>0.0378113</td>
</tr>
<tr>
<td>$D_s(ffe')$</td>
<td>PODIFW</td>
<td>-0.0476584</td>
<td>-0.0557864</td>
</tr>
<tr>
<td>$SD_i(ffe)$</td>
<td>PRSDT</td>
<td>0.1590309</td>
<td>0.1248413</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
<td>-9.0238791</td>
<td>-8.7261873</td>
</tr>
</tbody>
</table>

Table 6.18 Results of classification of pituitary adenomas and meningiomas in Data Set RP-GF using 5 variables.

<table>
<thead>
<tr>
<th>Actual Group</th>
<th>No. of Cases</th>
<th>Predicted Group Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Group 1</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90.9%</td>
</tr>
<tr>
<td>Group 2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16.7%</td>
</tr>
<tr>
<td>Ungrouped cases</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75.0%</td>
</tr>
</tbody>
</table>
Table 6.19 Results of the ten-fold cross-validation on the classification of pituitary adenomas and meningiomas in Data Set RP-GF using 5 variables.

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Training Set</th>
<th>Test Set</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of examples</td>
<td>No. of errors</td>
<td>% error</td>
<td>No. of examples</td>
<td>No. of errors</td>
<td>% error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>4</td>
<td>16.0%</td>
<td>3</td>
<td>1</td>
<td>33.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>4</td>
<td>16.0%</td>
<td>3</td>
<td>2</td>
<td>66.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>2</td>
<td>8.0%</td>
<td>3</td>
<td>1</td>
<td>33.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>3</td>
<td>12.0%</td>
<td>3</td>
<td>0</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>4</td>
<td>16.0%</td>
<td>3</td>
<td>1</td>
<td>33.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>6</td>
<td>24.0%</td>
<td>3</td>
<td>0</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>5</td>
<td>20.0%</td>
<td>3</td>
<td>0</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>3</td>
<td>12.0%</td>
<td>3</td>
<td>1</td>
<td>33.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>6</td>
<td>23.1%</td>
<td>2</td>
<td>0</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>2</td>
<td>7.7%</td>
<td>2</td>
<td>1</td>
<td>50.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>25.2</td>
<td>3.9</td>
<td>15.5%</td>
<td>2.8</td>
<td>2.1</td>
<td>25.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.4.4 Conclusions

Discriminant analysis achieved 100% classification of astrocytomas and meningiomas. Similarly in the classification of pituitary adenomas versus meningiomas, using the data set P-GIF, 100% discrimination was achieved between pituitary adenomas and meningiomas by this statistical data mining tool. The environment was one of supervised learning.

Discriminant analysis was then done on the data set RP-GF. The accuracy of discrimination achieved was 89%. Ten-fold cross validation was then performed. The overall percentage of accuracy in predicting the unknowns was 75%. However the accuracy of prediction of the knowns also ranged from 77 - 92%.

A few variables from the FLAIR scan were found to be statistically significant on the basis of the t-test for discrimination between astrocytomas and meningiomas. However, for discrimination between pituitary adenomas and meningiomas, they were found to be insignificant.

This supervised learning tool shows promise of usefulness in enhancement of tissue characterisation by quantitative MRI in situations where classification manually or using unsupervised techniques is relatively unsuccessful, as in the case of discrimination between meningiomas and pituitary adenomas in the pituitary region of the brain.

6.5 DECISION TREES

6.5.1 Introduction

Decision trees were constructed for the two classes astrocytomas and meningiomas and also for the two classes pituitary adenomas and
meningiomas. In this supervised learning setting, the well-known algorithm for generating decision trees called C4.5 was used. The following sections examine the features of the trees generated from the Data Sets P-GIS, RP-GS, P-GIF and RP-GF.

6.5.2 Decision trees for astrocytomas and meningiomas

Data set P-GIS: The decision tree for astrocytomas and meningiomas was constructed using the Data Set P-GIS. This data set contains 30 cases. Of these, 8 are astrocytomas, 12 are meningiomas, and 10 are other primary brain tumours. The number of attributes used was 46.

The decision tree generated is given below:

\[
\begin{align*}
\text{podifw} &\leq -52 : \text{a (8.0)} \\
\text{podifw} &> -52 : \text{m (12.0)} 
\end{align*}
\]

The decision tree is shown in figure 6.16.

The single node of the decision tree used the attribute \(D_w(se')\). If \(D_w(se') \leq -52\) then the sample is an astrocytoma. Eight samples were classified as astrocytomas at this node. If \(D_w(se') > -52\), then the sample is a meningioma. Twelve samples were classified as meningiomas.

The decision tree is also expressed as ripple-down rules. These rules output by the program are

Rule 1:

\[
\text{podifw} \leq -52 \\
\rightarrow \text{class a}
\]

Rule 2:

\[
\text{podifw} > -52 \\
\rightarrow \text{class m}
\]
Figure 6.16 Decision tree for classifying astrocytomas and meningiomas in Data Sets P-GIS and RP-GS

Figure 6.17 Decision tree for classifying pituitary adenomas and meningiomas in Data Set P-GIF and RP-GF
Default class: m

Rule one states that if $D_w(se')$ is less than or equal to -52 the sample is separated into the astrocytoma group. Rule 2 states that if $D_w(se')$ is greater than -52 then the sample is separated into the meningioma group. The default class is meningioma.

The 12 meningiomas were classified correctly as meningiomas and the 8 astrocytomas were classified correctly as astrocytomas.

Ten-fold cross-validation was performed on a subset of the Data Set P-GIS containing 20 cases of 8 astrocytomas and 12 meningiomas. The number of attributes used was 46. The results are shown in table 6.20. One misclassification occurred during the cross validation. This gives an average percentage error of 5% in the test sets, e.g., a success rate of 95%.

**Data set RP-GS:** A decision tree was then constructed for a subset of the Data Set RP-GS comprising 26 cases of which 16 are meningiomas and 10 are astrocytomas. The $T_2$ weighted GRASE and the $T_1$ weighted Spin-echo data alone are used. The number of attributes used was 35.

The decision tree generated is given below:

podifw <= -52 : a (10.0)
podifw > -52 : m (16.0)

The decision tree is shown in figure 6.16.
Table 6.20 Results of the ten-fold cross-validation on the decision tree classification of astrocytomas and meningiomas in Data Set P-GIS

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Training set</th>
<th>Test set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of examples</td>
<td>No. of errors</td>
<td>%error</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Average</td>
<td>18</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
The first node of the decision tree used the attribute $D_w(se')$. If $D_w(se') \leq -52$ then the sample is an astrocytoma. If $D_w(se') > -52$ then the sample is a meningioma. The number of nodes in this tree is one.

The decision tree is also expressed as ripple-down rules. These rules output by the program are shown below:

Rule 1:

\[ podifw \leq -52 \rightarrow \text{class a} \]

Rule 2:

\[ podifw > -52 \rightarrow \text{class m} \]

Default class: m

Rule one states that if $D_w(se')$ is less than or equal to -52 then the sample is separated into the astrocytoma group. Rule 2 states that if $D_w(se')$ is greater than -52 the sample is separated into the meningioma group; and the default class is meningioma.

The 16 meningiomas were classified correctly as meningiomas, and the 10 astrocytomas were classified correctly as astrocytomas.

Ten-fold cross-validation was performed on the subset of the Data Set RP-GS containing 26 cases of 10 astrocytomas and 16 meningiomas. The number of attributes used was 35. The results are shown in table 6.21. One misclassification occurred during the cross validation. This gives an average percentage error of 3.3% in the test sets, i.e., a success rate in the test sets is 96.7%. This data set contained 6
Table 6.21 Results of the ten-fold cross-validation on the decision tree classification of astrocytomas and meningiomas in Data Set RP-GS

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Training set</th>
<th>Test set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of examples</td>
<td>No. of errors</td>
<td>%error</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Average</td>
<td>23.4</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
samples more than the previous data set. The error rate dropped from 5% in the previous data set to 3.3% in this data set.

6.5.3 Decision trees for pituitary adenomas and meningiomas

Data set P-GIF: The decision tree for pituitary adenomas and meningiomas was constructed using the Data Set P-GIF. This data set contains 25 cases. Of these, 15 are pituitary adenomas, 4 are meningiomas, and 6 are other primary brain tumours. The number of attributes used was 46.

The decision tree generated is given below:

\[
\begin{align*}
\text{poconw} & \leq 166 : \text{p (15.0)} \\
\text{poconw} & > 166 : \text{m (4.0)} 
\end{align*}
\]

The first node of the decision tree used the attribute \(C_w(ffe')\). If \(C_w(ffe') \leq 166\) then the sample is a pituitary adenoma. Fifteen samples were classified as pituitary adenomas at this node. If \(C_w(ffe') > 166\), then the sample is a meningioma. Four samples were classified as meningioma at this node.

The Ripple-down rules output by the program are shown below.

Rule 1:
\[
\begin{align*}
\text{poconw} & \leq 166 \\
\rightarrow & \text{ class p}
\end{align*}
\]

Rule 2:
\[
\begin{align*}
\text{poconw} & > 166 \\
\rightarrow & \text{ class m}
\end{align*}
\]
Default class: p
Rule one states that if \( C_w(ffe') \leq 166 \), then the sample is classified as a pituitary adenoma. Rule 2 states that if \( C_w(ffe') > 166 \), the sample is classified as a meningioma. The 4 meningiomas were correctly classified as meningiomas and the 15 pituitary adenomas were correctly classified as pituitary adenomas.

Ten-fold cross-validation was performed on these 19 cases and the result of this test is shown in table 6.22. Two misclassifications occurred during the cross validation. This gives an average percentage error of 10.0% in the test sets, i.e., a success rate in the test sets is 90%. In the training set the error is 0%.

Data set RP-GS: A decision tree was then constructed for a subset of the Data Set RP-GS comprising 28 cases of which 6 are meningiomas and 22 are pituitary adenomas. The \( T_2 \) weighted GRASE and the \( T_1 \) weighted Spin-echo data alone are used. The number of attributes used was 35.

The decision tree generated is given below:

\[
\begin{align*}
\text{poconw} \leq 166 & : p \ (22.0/1.0) \\
\text{poconw} > 166 & : m \ (6.0/1.0)
\end{align*}
\]

The decision tree is shown in figure 6.17.

The first node of the decision tree used the attribute \( C_w(ffe') \). If \( C_w(ffe') \leq 166 \) then the sample is an pituitary adenoma. All 22 pituitary adenoma samples were classified as pituitary adenomas at this node. If \( C_w(ffe') > 166 \), then the sample is a meningioma. All 6 meningioma samples were classified as meningioma at this node.
Table 6.22 Results of the ten-fold cross-validation on the decision tree classification of pituitary adenomas and meningiomas in Data Set P-GIF

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of examples</td>
<td>No. of errors</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>17.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>
The decision tree was expressed as ripple-down rules. These rules output by the program are:

Rule 1:
\[ \text{poconw} \leq 166 \]
\[ \rightarrow \text{class p} \]

Rule 2:
\[ \text{poconw} > 166 \]
\[ \rightarrow \text{class m} \]

Default class: p

Rule one states that if \( C'_w(ffe') \) is less than or equal to 166, then the sample is classified as a pituitary adenoma. Rule 2 states that if \( C'_w(ffe') \) is greater than 166, the sample is classified as a meningioma.

One pituitary adenoma was misclassified as a meningioma whereas the remaining 21 pituitary adenomas were correctly classified. One meningioma was misclassified as a pituitary adenoma whereas the remaining 5 meningiomas were correctly classified. These results are shown in table 6.23.

Ten-fold cross-validation was performed on these 28 cases. The results of this cross validation test are shown in table 6.24. Five misclassifications occurred during the cross validation. This gives an average percentage error of 16.6% in the test sets, i.e., a success rate in the test sets is 83.4%. In the training set the error is 5.9%.

6.5.4 Conclusions

Decision trees were constructed for the classification of astrocytomas and meningiomas in the two data sets P-GIS and RP-GS. In
Table 6.23 Results of the decision tree classification of pituitary adenomas and meningiomas in Data Set RP-GF

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Classified as</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>21</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Table 6.24 Results of the ten-fold cross-validation on the decision tree classification of pituitary adenomas and meningiomas in Data Set RP-GF

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of examples</td>
<td>No. of errors</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>Average</td>
<td>25.2</td>
<td>1.5</td>
</tr>
</tbody>
</table>
the first instance, the decision trees showed an average accuracy of prediction of 95% on the test sets in a ten-fold cross-validation test. In the second instance, when the number of samples (by using the data set RP-GS) was increased by 10, the accuracy increased to 96.7% in the test sets. In the training sets the accuracy of classification was 100% in both instances.

Decision trees were also constructed using the same algorithm for the classification of pituitary adenomas and meningiomas. In this experiment, the two data sets P-GIF and RP-GF were mined. In the first instance, the accuracy of prediction on the test set was 90%. In the second instance the accuracy dropped to 83.4%. In the training set too the accuracy dropped from 100% to 94.1%. This behaviour is similar to the behaviour observed in the discriminant analysis.

The accuracy of prediction using decision trees ranges from 83% to 100% in these two classification problems and this indicates that decision trees could be useful to enhance the diagnostic usefulness of MR images in discriminating between astrocytomas and meningiomas and between pituitary adenomas and meningiomas. The accuracy of classification of meningiomas and pituitary adenomas is of the same order of magnitude using both discriminant analysis and decision trees.

These decision trees are small and are thus useful for gaining some insights into the basis of decision making in this classification problem. It is thus shown that this supervised learning technique is able to discover knowledge from MRI data.
6.6 INDUCTIVE LOGIC PROGRAMMING

6.6.1 Introduction

In the first of these studies, the Inductive Logic Programming tool Progol was used for mining the data set N to classify normal grey and white matter (Siromoney et al., 1996). In the next study, the data set P-GIS was mined using Progol for classification of meningiomas and astrocytomas. Section 6.6.3 describes this experiment to classify meningiomas and astrocytomas.

6.6.2 Characterisation of normal tissues using ILP

The Data Set N was used in the study of normal tissues using Progol. CProgol version 4.1 dated 2.4.95 was used on a 80486, 33 MHz, 4 MB RAM computer running Linux. Progol was given the signal intensities $I_g$ of the 80 positive examples (grey matter) and the signal intensities $I_w$ of the 83 negative examples (white matter) as background knowledge. It used the mode declarations (that indicated to Progol that the algebraic comparisons '< or '=' could be useful), examined the examples and induced a logic program with five rules, that could correctly determine to which of these two types of normal brain tissue a given sample belongs (Siromoney et al., 1996).

Progol found five generalised rules for classification of these samples as positive or negative, that is as grey matter or white matter, on the basis of signal intensities. Five positive samples were not classifiable. No negative samples were covered.
The rules formulated by Progol are as follows:

\[ \text{grey}(A,B,C) : \ A < 7823. \]
\[ \text{grey}(A,B,C) : 5730 < B, \ A < 8509. \]
\[ \text{grey}(A,B,C) : 6614 < B, \ A < 8649. \]
\[ \text{grey}(A,B,C) : 7959 < B, \ 13618 < C. \]
\[ \text{grey}(A,B,C) : 13453 < C, \ A < 9749. \]

where A is the signal intensity from the T₁ weighted image, B the signal intensity from the T₂ weighted image, and C the signal intensity from the proton density weighted image.

In summary, in this data set, a sample is grey matter if the signal intensity from T₁ weighted image is less than or equal to 782.3; or if the signal intensity from T₂ weighted image is greater than or equal to 573.0 AND the signal intensity from T₁ weighted image is less than or equal to 850.9; etc.

The supervised learning examples were input to Progol, which then used the mode declarations, examined the examples and generated five rules that successfully characterised these normal tissues. Thus it is seen that Progol can distinguish between normal grey and white matter using their signal intensities.

6.6.3 Characterisation of astrocytomas and meningiomas using ILP

The Data Set P-GIS was used in the study of astrocytomas and meningiomas using Progol. Progol version 4.4 dated 25.08.98 was used on a Sun Sparcstation 20. In the first run Progol was asked to use the signal intensity values I themselves. The next run used the relationships
between the signal intensities, for example between $I_t$ and $I_w$, $I_t$ and $I_g$, etc. Then leave one out testing was performed.

In the first run Progol was asked to use the signal intensity values themselves. The $I_t(grase)$, $I_w(grase)$, $I_g(grase)$, $I_c(grase)$, $I_t(flair)$, $I_w(flair)$, $I_g(flair)$, $I_c(flair)$, $I_t(se)$, $I_w(se)$, $I_g(se)$, $I_c(se)$, $I_t(se')$, $I_w(se')$, $I_g(se')$ and $I_c(se')$ values were used. The Progol input file is given below.

```
:- modeh(l,al234(+sernum))? 
:- modeb(*,+int =< #int)? 
:- modeb(*,#int =< +int)? 
:- set(inflate,100000)? 
:- set(nodes,10000)? 
:- set(c,10)? 
:- set(i,5)? 
:- set(verbose,0)? 
:- modeb(10,sigintgt2(+sernum,-int,-int,-int,-int))? 
:- modeb(10,sigintlt2(+sernum,-int,-int,-int,-int))? 
:- modeb(10,sigintstl(+sernum,-int,-int,-int,-int))? 
:- modeb(10,sigintsgl(+sernum,-int,-int,-int,-int))? 
:- consult(am_p_gls)? 
```

The file am_p_gls.pl (used in the consult statement above) contains clauses in the following form.

```
sigintgt2(s111,1480,767,946,2054).
sigintstl(s111,780,1014,810,247).
sigintsgl(s111,911,1001,699,226).
s1234(s111).
sernum(s111).
sigintgt2(s114,1018,778,931,1962).
```
Progol generated the following hypothesis.

\[ \text{a1234}(A) : \text{sigintgt2}(A, B, C, D, E), 1194 =< B, B =< 1224. } \]
\[ \text{a1234}(A) : \text{sigintgt2}(A, B, C, D, E), 1607 =< B, C =< 708. } \]
\[ \text{a1234}(A) : \text{sigintgt2}(A, B, C, D, E), 1380 =< B, 1863 =< E, C =< 769. } \]

The clauses mean that the tumour is an astrocytoma if for the GRASE T2 weighted scans, the mean tumour signal intensity is \( \geq 1194 \) AND the mean tumour signal intensity \( \leq 1224 \), OR the mean tumour signal intensity \( \geq 1607 \) AND the white matter signal intensity \( \leq 708 \), OR the mean tumour signal intensity \( \geq 1380 \) AND the white matter signal intensity \( \leq 769 \) AND the cerebrospinal fluid signal intensity \( \geq 1863 \).

Progol took more than 16 minutes on a Sun Sparcstation 20 and could not complete all possible paths in its search since the resources available were exceeded. In other words, by using the signal intensity of any particular region, Progol could not determine any useful hypothesis to distinguish patients with astrocytoma from those with meningioma. Thus this run did not yield significant results.

The next run used the relationship between the signal intensities. The input file follows:

\[- \text{modeh}(1, \text{a1234}(+\text{sernum}))? \]
\[- \text{modeb}(*, +\text{int} =< +\text{int})? \]
\[- \text{set}(\text{inflate}, 100000)? \]
set(nodes, 10000)?
:- set(verbose,0)?
:- modeb(*,sigintgt2(+sernum,-int,-int,-int,-int))?  
:- modeb(*,sigintlt2(+sernum,-int,-int,-int,-int))?  
:- modeb(*,sigintstl(+sernum,-int,-int,-int,-int))?  
:- modeb(*,sigintsgl(+sernum,-int,-int,-int,-int))?  
:- consult(am_p_gls)?

The modeb declaration with '+int =< +int' tells Progol that any two mean signal intensities (tumour, white matter, grey matter, cerebrospinal fluid) can be compared.

Progol quickly generated a meaningful hypothesis when it was asked to make an algebraic comparison between the signal intensities. The hypothesis generated was

\[ a_{1234}(A) :\text{sigintsgl}(A,B,C,D,E), B =< C. \]

This clause means that the tumour is an astrocytoma if the mean signal intensity of the tumour is less than that of white matter in the spin echo post gadolinium scan. Progol took around 8 seconds to determine this.

Leave-one-out testing was then done. In leave one out testing, one example at a time is removed from the data set before the analysis and is later used as an unknown test case. The hypothesis is induced from the remaining examples in the data set. The induced hypothesis is then used on the unknown test case to predict the tumour type. This is repeated for each example in the data set. The overall percentage of accuracy in predicting the tumour type of the unknown test cases is then calculated.
Progol took 7:33.2 minutes to do the entire leave-one-out testing and achieved 100% results. The results of the leave one out testing are given in Table 6.25. The hypothesis clause is the same in each run and is given below.

\[ a1234(A) \leftarrow \text{sigintsg1}(A,B,C,D,E), B = C. \]

It is significant that the induced hypothesis was found to be identical in each of the runs of the leave-one-out testing. Since the induced hypothesis did not require any information from the \( T_1 \)-weighted FLAIR scan, leave-one-out testing was also performed on the Data Set RP-GS, the identical hypothesis was induced in all the runs and achieved 100% results.

### 6.6.4 Conclusions

The ILP tool Progol was used on medical imaging data to quantitatively discriminate between meningiomas and astrocytomas. Progol was especially suited for this type of knowledge discovery from medical images where the relationships between data from different regions in the medical image were used. Analysis of the signal intensity of one region alone (the tumour tissue) did not discover any useful knowledge. However, the application of Progol to the relationships between the signal intensity of different regions (tumour, white matter, grey matter and cerebrospinal fluid) gave a very clear result. Thus the knowledge discovery process in images of meningiomas and astrocytomas is facilitated by the application of ILP. ILP was also shown to be successful in mining useful knowledge for discriminating between grey and white matter.
Table 6.25 Results of the leave-one-out testing of the classification by Progol of astrocytomas and meningiomas in the Data Set P-GIS.

<table>
<thead>
<tr>
<th></th>
<th>Actual Positive examples (Astrocytomas)</th>
<th>Actual Negative examples (Meningiomas)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Positive examples (Astrocytomas)</td>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Predicted Negative examples (Meningiomas)</td>
<td>0</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Totals</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>