肺癌、癌旁及正常组织的甲基化数据

94 probes，400 samples。7个正常组织，99个癌旁组织，294肺癌组织。

目标：利用甲基化信息对肺癌进行诊断

方法：classification model based on principal components

Step 1: 主成分分析

本例有400样本，94个变量，因此一般的主成分分析已经足够，可以利用SAS进行主成分分析，得到主成分。

Step 2: 主成分的挑选

从第一步得到的主成分中，利用logistic回归模型挑选三组显著差异的主成分。由于有三组数据：

① 癌旁与正常组织有差异的主成分

② 肺癌与癌旁组织有差异的主成分

③ 肺癌与癌旁、正常组织有差异的主成分

④ 比较2,3的结果

Step 3: 利用上一步得到的主成分构建相应的classification model。

问题：

1、缺失值 78个缺失

解决方法：以0代替

**第一部分：数据未经校正的分析结果:**

对原数据进行主成分分析后，利用logistic回归模型寻找与分组特异性有关的主成分。其结果如下：

1、癌旁与正常组织有差异的主成分：

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin2 | 0.006136 |
| Prin3 | 0.000867 |
| Prin9 | 0.001892 |
| Prin16 | 0.03186 |
| Prin17 | 0.002232 |
| Prin20 | 0.00354 |
| Prin21 | 0.003127 |
| Prin23 | 0.002739 |
| Prin30 | 0.033389 |
| Prin35 | 0.012164 |
| Prin39 | 0.007287 |
| Prin41 | 0.039608 |
| Prin42 | 0.03423 |
| Prin44 | 0.020568 |
| Prin45 | 0.025814 |
| Prin49 | 0.004844 |
| Prin53 | 0.013432 |
| Prin57 | 0.015478 |
| Prin58 | 0.001751 |
| Prin60 | 0.011403 |
| Prin67 | 0.002655 |
| Prin70 | 0.033742 |
| Prin76 | 0.020998 |
| Prin77 | 0.006641 |
| Prin85 | 0.03492 |
| Prin90 | 0.042629 |
| Prin94 | 0.001145 |

2、肺癌与癌旁组织有差异的主成分

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin1 | 5.2E-20 |
| Prin2 | 9.38E-09 |
| Prin3 | 3.36E-17 |
| Prin6 | 0.00045 |
| Prin8 | 0.000252 |
| Prin9 | 0.000562 |
| Prin11 | 0.005675 |
| Prin21 | 0.019565 |

3、肺癌与正常组织有差异的主成分

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin1 | 0.000414 |
| Prin2 | 0.006075 |
| Prin77 | 0.031034 |
| Prin94 | 0.002268 |

**以上结果显示，癌旁与正常组织有较大差异，但是由于正常组织只有7个样本，尚不能下差异显著性的结论。因此基于肺癌与正常组织建立的诊断模型有待商榷。**

**该结果亦表明癌旁数据与正常组织的数据不能轻易合并。**

**下面是基于肺癌与癌旁组织建立的肺癌诊断模型：**

**Leaving-one-out crossvalidation的结果显示the classification model based on principal components的结果是几种方法中最好的。**

The classification model based on principal components

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
| 1,2,3,6,8,9 | Tumor | 292 | 2 | 99.32% | 293 | 1 | 99.66% |
|  | Para | 1 | 98 | 98.99% | 1 | 98 | 98.99% |
|  |  | 99.66% | 98.00% | 99.24% | 99.66% | 98.99% | 99.49% |

Decision tree (C4.5)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 290 | 4 | 98.64% | 285 | 9 | 96.94% |
|  | Para | 1 | 98 | 98.99% | 7 | 92 | 92.93% |
|  |  | 99.66% | 96.08% | 98.73% | 97.60% | 91.09% | 95.93% |

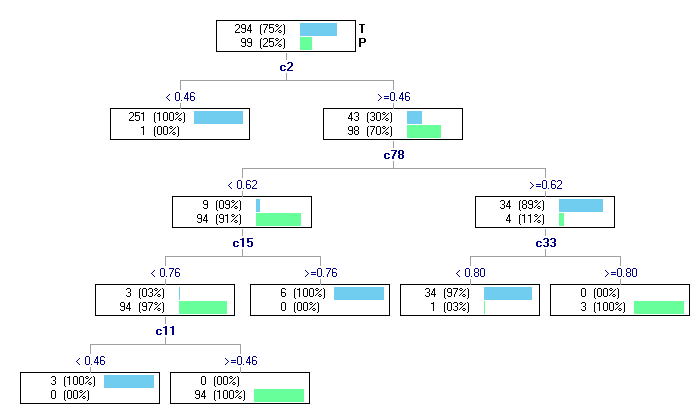


图 Decision Tree的结果

LDA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 293 | 1 | 99.66% | 289 | 5 | 98.30% |
|  | Para | 0 | 99 | 100.00% | 3 | 96 | 96.97% |
|  |  | 100.00% | 99.00% | 99.75% | 98.97% | 95.05% | 97.96% |

SVM

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 293 | 1 | 99.66% | 290 | 4 | 98.64% |
|  | Para | 0 | 99 | 100.00% | 2 | 97 | 97.98% |
|  |  | 100.00% | 99.00% | 99.75% | 99.32% | 96.04% | 98.47% |

**第二部分：经过校正的数据分析结果**

1、癌旁与正常组织有差异的主成分：

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin6 | 0.00996 |
| Prin9 | 0.041211 |
| Prin10 | 0.006885 |
| Prin12 | 0.025652 |
| Prin14 | 0.003171 |
| Prin15 | 0.022668 |
| Prin19 | 0.012008 |
| Prin21 | 0.000914 |
| Prin23 | 0.041448 |
| Prin32 | 0.00257 |
| Prin33 | 0.013074 |
| Prin39 | 0.000903 |
| Prin46 | 0.024676 |
| Prin54 | 0.031098 |
| Prin57 | 0.008729 |
| Prin61 | 0.021261 |
| Prin65 | 0.001107 |
| Prin67 | 0.02053 |
| Prin71 | 0.039475 |
| Prin82 | 0.011044 |
| Prin84 | 0.042855 |
| Prin88 | 0.034024 |
| Prin93 | 0.005682 |

2、肺癌与癌旁组织有差异的主成分

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin1 | 2.68E-23 |
| Prin2 | 1.52E-17 |
| Prin4 | 0.000708 |
| Prin6 | 4.8E-06 |
| Prin7 | 0.003407 |
| Prin9 | 0.016519 |
| Prin10 | 0.029661 |
| Prin14 | 0.001632 |

3、肺癌与正常组织有差异的主成分

|  |  |
| --- | --- |
| Variable | ProbChiSq |
| Prin1 | 0.008313 |
| Prin82 | 0.044478 |
| Prin93 | 0.014576 |

**以上结果显示，癌旁与正常组织有较大差异，癌旁数据与正常组织的数据不能轻易合并。由于正常组织只有7个样本，癌旁数据与正常组织以及肺癌数据与正常组织数据尚不能下差异显著性的结论。**

**下面是基于肺癌与癌旁组织建立的肺癌诊断模型：**

**Leaving-one-out crossvalidation的结果显示the classification model based on principal components的结果比较稳定，比Decision Tree和LDA的结果好，基于主成分1,2,6,7,14的结果比SVM的结果好，但是我们认为综合敏感性、特异性、模型简约性，the classification model based on principal components可以有两种模型可供选择。**

The classification model based on principal components

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
| 1,2,7,14 | Tumor | 292 | 2 | 99.32% | 292 | 2 | 99.32% |
|  | Para | 2 | 97 | 97.98% | 2 | 97 | 97.98% |
|  |  | 99.32% | 97.98% | 98.98% | 99.32% | 97.98% | 98.98% |
| 1,2,6,7,14 | Tumor | 292 | 2 | 99. 32% | 291 | 3 | 98.98% |
|  | Para | 1 | 98 | 98.99% | 0 | 99 | 100% |
|  |  | 99.66% | 98.00% | 99.24% | 100% | 97.06% | 99.24% |

Decision tree (C4.5)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 293 | 1 | 99.66% | 282 | 12 | 95.92% |
|  | Para | 2 | 97 | 97.98% | 8 | 91 | 91.92% |
|  |  | 99.32% | 98.98% | 99.24% | 97.24% | 88.35% | 94.91% |

LDA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 0 | 294 | 0.00% | 22 | 272 | 7.48% |
|  | Para | 0 | 99 | 100% | 1 | 98 | 98.99% |
|  |  | 0.00% | 25.19% | 25.19% | 95.65% | 26.49% | 30.53% |

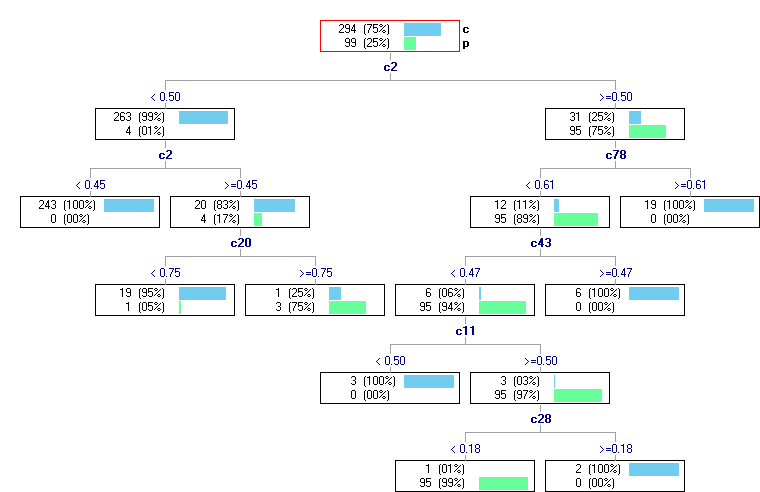


图 Decision Tree的结果

SVM

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Result | | | Leaving-one-out crossvalidation | | |
|  |  | Positive | Negative |  | Positive | Negative |  |
|  | Tumor | 293 | 1 | 99.66% | 291 | 3 | 98.98% |
|  | Para | 0 | 99 | 100% | 1 | 98 | 98.99% |
|  |  | 100% | 99.00% | 99.75% | 99.66% | 97.03% | 98.98% |