数据：共70，分别为33个肺腺瘤病人样本，37个对照组样本

随机分配60%为训练集，20%为验证集，20%为测试集。（具体分布情况可看表格data\_random.csv）

步骤：通过迭代的方法，逐步找出从1个特征到最大特征数量中PCA-LDA模型的准确率和样本距离最大的特征组合，随后使用验证集分别对这些组合进行测试，当特征数为5时，验证集的分类效果最佳，为85.71%。

迭代图如下：

图表

描述已自动生成

上述步骤得到最优biomarker组合，其PEAK-ID和保留时间如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| index | 保留时间 | 可能物质 | 名称 | CAS号 |
| 43 | 4.7088717073170745 | C4H8S | 1-Propene, 1-(methylthio)-, (Z)- | 52195-40-1 |
| 65(新峰) | 6.245888888888889 | C6H12O2 | Acetic acid, butyl ester | 123-86-4 |
| 144 | 12.290963571428572 | C9H18O | Nonanal | 124-19-6 |
| 149 | 12.65858595744681 | C10H14 | Benzene,1-ethyl-2,3-dimethyl- | 933-98-2 |
| 215 | 17.70627341463415 | C8H8N2O | 2H-Benzimidazol-2-one | 5400-75-9 |

在图谱中的位置如下：

图表, 直方图

描述已自动生成

取该biomarker组合，并用训练集训练PCA-LDA模型，并对测试集做出预测，预测结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Lung** | **Control** | **Total** |
| Train Set | 20 | 22 | 42 |
| Valid Set | 7 | 7 | 14 |
| **Test Set** | **5** | **9** | **14** |
| Positive | 5 | 1 | 6 |
| Negative | 0 | 8 | 8 |
| Specificity | **88.89%** | | |
| Sensitivity | **100%** | | |
| Positive predictive value | **83%** | | |
| Negative predictive value | **100%** | | |
| Total accuracy | **93%** | | |

PCA 两个主成分分布图：（训练集+测试集）

图表, 散点图

描述已自动生成