科研合作数据分析流程

一 所用数据：

cnv：

1. 使用cnvkit软件batch模块计算每个样本的panel区间的log2（每一组所有的血细胞样本作为baseline），之后可以使用call计算拷贝数变化的绝对值；
2. 对于一个基因有多个panel区间的取其中位值，然后生成如下表格（所用数据log2或者cn的中位值，包含两组的数据）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gene | cf\_S13 | cf\_S82 | cf\_S106 | cf\_S27 |
| MTOR | 0.094197 | 0.169443 | 0.030789 | 0.121609 |
| AGTRAP | 0.760645 | 0.053456 | 0.19567 | 0.118796 |
| MTHFR | 0.170298 | 0.182852 | -0.00495 | 0.048328 |
| CLCN6 | 0.20477 | -0.07085 | 0.111363 | -0.13455 |
| CASP9 | -0.79295 | -0.1698 | -0.29705 | -0.12274 |
| ZBTB17 | -1.1495 | -0.22018 | -0.17059 | 0.107315 |
| EPHA2 | 0.17365 | 0.109873 | 0.225221 | 0.042558 |
| . | -0.08847 | -0.00201 | 0.032762 | 0.009509 |
| CDA | -0.67094 | 0.036059 | 0.107614 | 0.311844 |
| SH2D5 | -0.62872 | 0.230529 | 0.138261 | 0.189029 |
| GALE | -0.73085 | -0.16658 | -0.22406 | 0.204103 |
| ARID1A | -0.16826 | 0.057032 | -0.04213 | -0.13112 |
| NUDC | -0.34676 | 0.074749 | -0.08328 | 0.304277 |
| OPRD1 | -0.21694 | 0.00342 | -0.05236 | -0.22504 |
| HDAC1 | 0.617212 | 0.221218 | 0.155792 | 0.353039 |
| AGO1 | 0.083123 | 0.212579 | -0.29339 | -0.04122 |
| CSF3R | 0.024845 | -0.1977 | 0.12904 | 0.066575 |

... ...  
.... ... ... ... ...

snp&indel

1. 使用varscan软件somatic模块计算每个样本的somatic突变，之后利用annovar注释，然后过滤尽量获得每个样本真实且治病的somatic突变；
2. 统计每个基因的突变个数或者平均突变频率，获得如下表格（包含两组样本）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gene | P2004 | P2046 | P2122 | P2156 |
| MTOR | 0 | 5.503333 | 0.63 | 0.57 |
| AGTRAP | 0 | 0 | 0 | 0 |
| MTHFR | 0 | 0 | 0.4 | 0.543333 |
| CASP9 | 0.65 | 0 | 0 | 0 |
| ZBTB17 | 0 | 0 | 0 | 0.54 |
| EPHA2 | 0.65 | 0.62 | 0 | 0 |
| CDA | 0.45 | 0 | 0 | 0.61 |
| SH2D5 | 0.92 | 0 | 0 | 0 |
| GALE | 0 | 0.42 | 0 | 0 |
| ARID1A | 0.513333 | 0.47 | 0.8975 | 0.62 |
| NUDC | 0 | 0 | 0 | 0 |
| HDAC1 | 0 | 0 | 0 | 0 |
| AGO1 | 0.49 | 0.56 | 0 | 0 |
| CSF3R | 0.89 | 3.12 | 0.5525 | 1.9 |
| MPL | 0 | 0 | 0 | 0 |
| MUTYH | 0 | 0 | 0 | 0 |
| FAAH | 0 | 0 | 0 | 0.43 |
| CMPK1 | 0.58 | 0 | 0 | 0 |
| CDKN2C | 0 | 0.63 | 0 | 0.68 |
| JAK1 | 0 | 0 | 0 | 0 |

综上：目前公司收集到的样本不多，一般用于训练和预测的比例为1:1，总共样本为30左右。数据整理及清洗之后均获得标准的格式化数据， 对于以后可能会将多组学的数据（甲基化、RNA-seq、cnv、snp-indel、位点突变等）添加到这一个格式化数据文件中来，打算使用标准化（x` = (x-mean)/S）、区间缩放法(x`=(x-min)/(max-min))等方法进行组合，不知道有没有哪些推荐的组合方法。

SNP的wild type 0/0跟missing data区分，cnv的no gain no loss no call 的严格区分？

没有区分,这个的必要性

somatic的fp过滤问题？

伴随检测怎么进行？

panel的问题是否也需要谈论？

二 数据分析：

特征基因选取：

1. wilcoxon秩和检验：如果这两个独立样本来自正态分布和具有相同方差时，我们可以采用t检验比较均值。但当这两个条件都不能确定时，我们常替换t检验法为Wilcoxon秩和检验来决定两个独立样本是否来自相同的或相等的总体，对计算结果排序选取特征基因（pvalue小于0.05或者0.01）；
2. pearsonr相关系数：

95c7b4484dc46f28390c4de96c83b915_r.png

皮尔逊相关系数衡量线性相关性，即两组数据是否和分组有线性相关，对结果排序选取特征基因（pvalue小于0.05或者0.01，或者相关系数）；之后会加入互信息系数，

6af9a077b49f587a5d149f5dc51073ba_b.png

其不仅能够度量线性相关并且能够很好地度量各种相关性，另外对于相关性计算，响应变量（分组信息）最好是连续型，目前都是离散型。

1. chi2检验：卡方校验的方法比较适合0/1型特征和稀疏矩阵的特征选择，目前snp\_indel的基因突变个数矩阵一般比较稀疏，所以不知道有没有推荐的稀疏矩阵选取特征基因的方法。
2. 方差选择法: 计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征
3. logistic回归：

http://img.blog.csdn.net/20140528194036875

这里http://img.blog.csdn.net/20140528194524250称为**Logistic函数。**其中http://img.blog.csdn.net/20140528201109953

对训练集进行拟合选取特征基因系数非零或者其绝对值大于一定标准的基因为特征基因（之后会根据变量系数的显著性选择，使用L1或者L2正则化防止拟合欠佳问题，调整正则化强度参数C值已增加或者减少特征数量，这均依据之后的分类效果进行调整）。

1. lasso回归：L1正则线性回归，使用aic评判标准进行拟合回归选取达到一定标准的权值系数。
2. RandomizedLasso：随机lasso，依旧分类效果调整alpha调整正则化参数alpha（随机lasso可以自动选择最优alpha，所以可以不用设置），防止拟合过度，提高泛化能力，是一种稳定选择特征基因的方法。
3. 递归特征消除法：递归消除特征法使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练，可以使用logistic或者其他线性回归模型进行递归。
4. 综合L1和L2正则化与logistic回归的特征选择法：

[L1惩罚项降维的原理在于保留多个对目标值具有同等相关性的特征中的一个](http://www.zhihu.com/question/28641663/answer/41653367)，所以没选到的特征不代表不重要。故，可结合L2惩罚项来优化。具体操作为：若一个特征在L1中的权值为1，选择在L2中权值差别不大且在L1中权值为0的特征构成同类集合，将这一集合中的特征平分L1中的权值

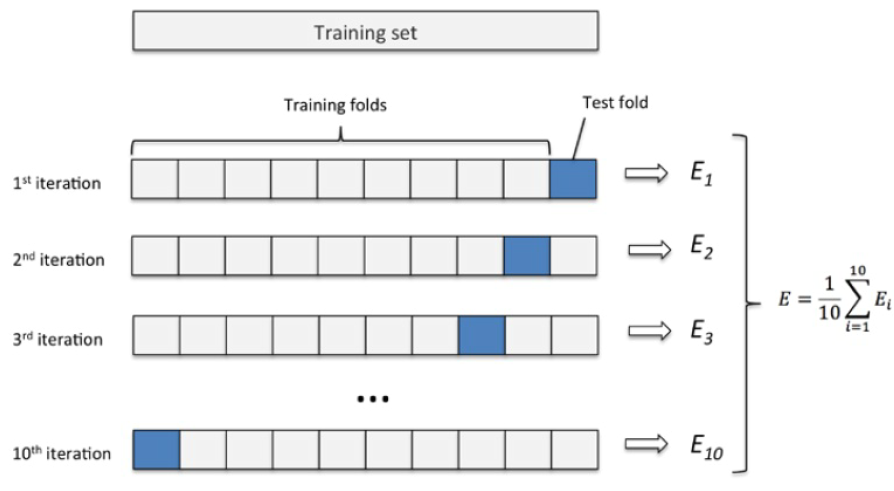
综上：1-4为单因素分析方法，5-9为多因素回归分析方法，另外可以使用组合方式进行特征选取，如先使用logistic回归去除对分类变量没有关系的基因，然后在使用wilcoxon检验去除掉两组样本数据不是显著差异的基因，不知是否值得考虑。

以上特征基因选取完毕之后即可进行模型训练，所用训练数据为特征基因的相关数据。

模型训练及模型评估

与之前的特征选择那里一样均会出现模型训练拟合欠佳问题，这里依据K折交叉验证评估模型性能不断反馈模型性能。

1. LinearSVC: 线性支持向量机
2. SVC：支持向量机
3. logistic：这一个方法参数也可与特征选择那里完全一致



使用以上训练集利用K折交叉验证评估模型性能，对于获取验证过程中的真阳性率及假阳性率，绘制ROC曲线，并计算AUC值

以上还需要进行调参，计划选用sklearn的自动调参，此处参数包括正则方法（L1、L2），正则强弱参数C

线性logistic和线性SVM通常效果近似。逻辑回归目标是最大化训练集的条件似然,使得她更易受奇异值影响。SVM只关心那些离决策界最近的点(即，支持向量)。另一方面，逻辑斯蒂回归的优点是易于实现，特别是并行化实现。此外，面对流式数据，逻辑斯蒂回归的易于更新的特点也很明显。

分类预测：

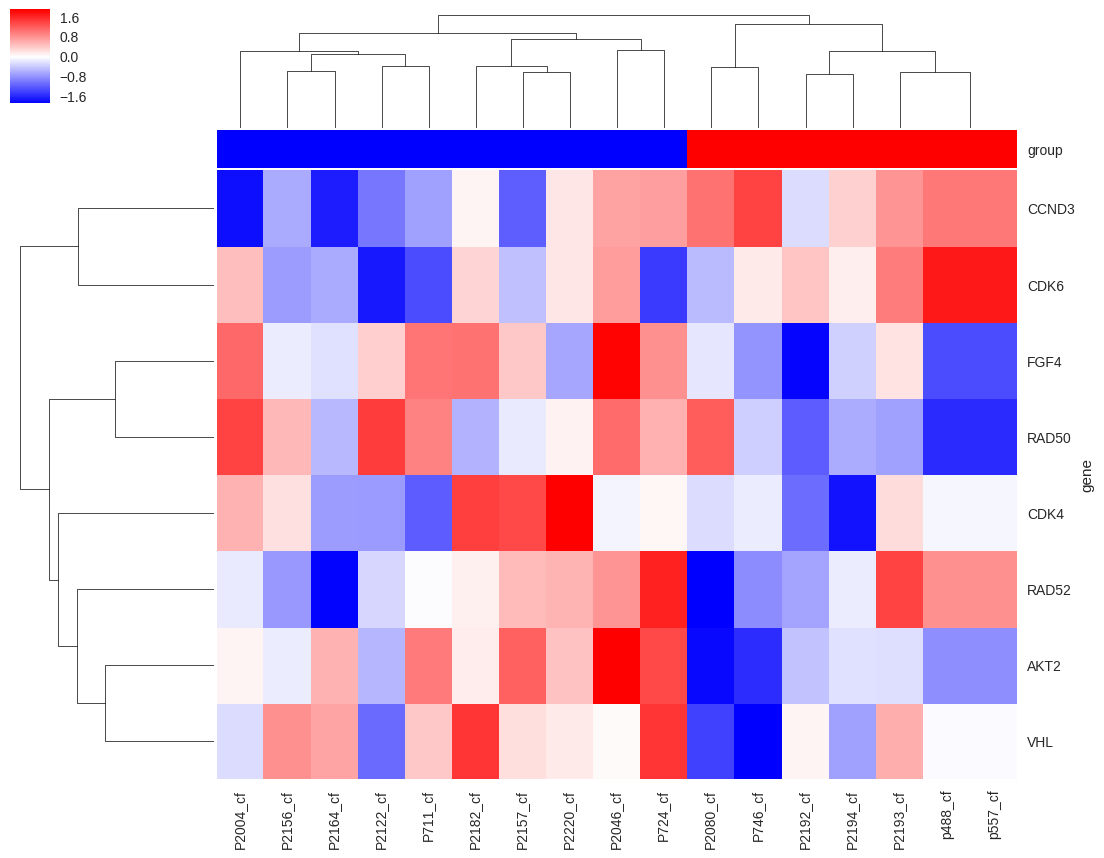
利用以上选取的特征基因及所选模型及参数对未知样本进行预测

1. 前期的测序及基本的变异检测要求除样本不一致外其他所用试剂及方法均一致。
2. 给出未知样本的预测准确率

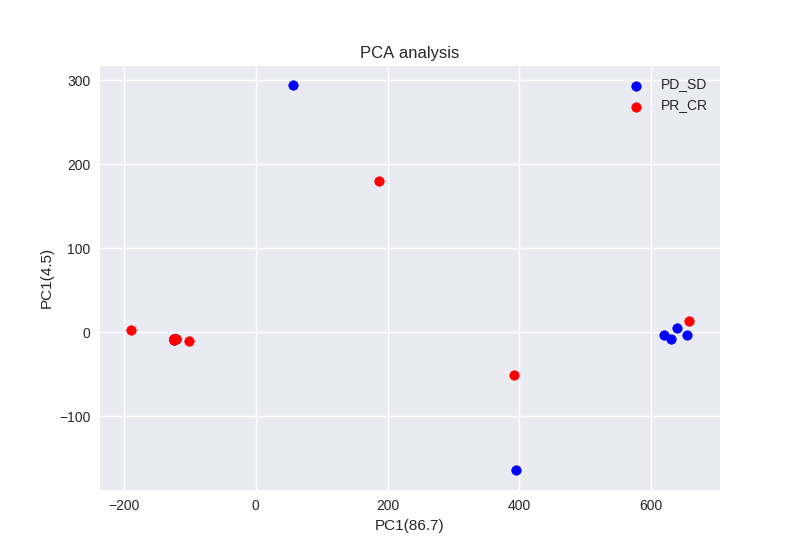
以上均使用基于python的sklearn编写的genecast包进行分析

结果展示：

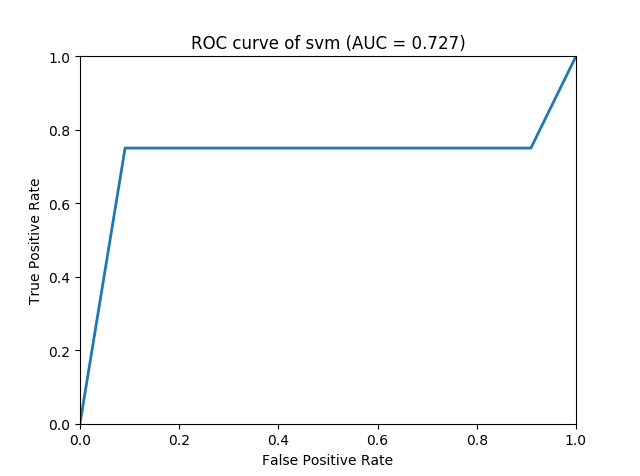
1. 对cnv数据的特征基因进行聚类：



1. 利用特征基因进行PCA分析对cnv数据



1. ROC曲线评估模型性能



1. 对未知样本的准确率83.56%

一个完整的例子-河南肿瘤医院的食管癌项目