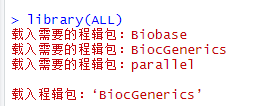
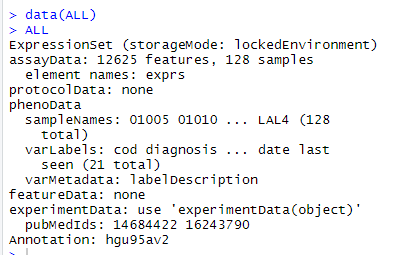
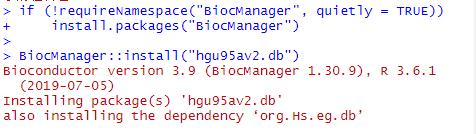
ALL数据集表型数据分析与可视化

基地一班 201730152009 刘利生

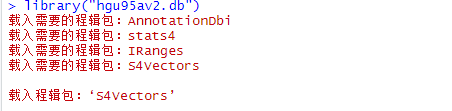
一、建立ExpressionSet object

#启用ALL数据包

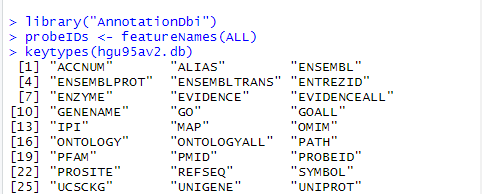
#查看ALL数据所用芯片类型

1. 将探针ID转化成基因symbol并获取注释

#本电脑不存在hgu95av2.db数据包，需从bioconductor下载。

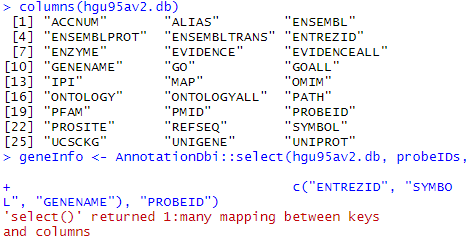


#启用芯片信息包

# 启用注释信息包

#将ALL数据包中的特征名称赋予给probeIDs向量

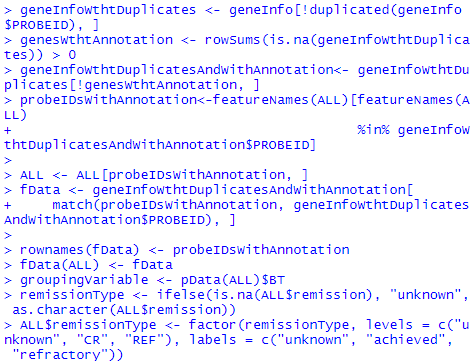
#查看数据包中有的哪些数据类型可作为select函数keytypes参数



#一般来说，column均可作为keytypes ，column输出可选取的列数，数据库中列数名为数据类型，故与keytypes输出结果相同（columns函数报错时手动输入该函数即可正常运行）

#使用select函数（数据库（hgu95av2.db），关键词(probeIDS)，columns(c(“ENTREZID”,”SYMBOL”,”GENENAME”)，keytpes(PROBEID)）将探针ID转化成基因symbol并获取注释，创建geneInfo数据框

（二）删除重复注释的基因

genesWthtAnnotation <- rowSums(is.na(geneInfoWthtDuplicates)) > 0

#使用is.na函数判断数据库是否为NA（空），rowsum汇总，将其赋予给genesWthtAnnotation，此时输出为逻辑性向量，等同于建立是否为空值索引

geneInfoWthtDulicaesAndWithAnnotation<geneInfoWthDuplicates[!genesWthtAnntation, ]

#利用之前的建立的索引创建新数据库，此时输出为无重复有注释的基因信息数据库

probeIDsWithAnnotation<featureNames(ALL)[featureNames(ALL)%in%geneInfoWthtDuplicatesAndWithAnnotation$PROBEID]

#选出既位于“PROBEID”eatureNames(ALL)gene又在WthtDuplicatesAndWithAnnotation中的数据

ALL <- ALL[probeIDsWithAnnotation, ]

#明确ALL的数据，到此完成ExpressionSet object的建立，用于之后的作图

fData <- geneInfoWthtDuplicatesAndWithAnnotation[

match(probeIDsWithAnnotation, geneInfoWthtDuplicatesAndWithAnnotation$PROBEID), ]

#建立fData矩阵

rownames(fData) <- probeIDsWithAnnotation

#将fData的行名称更改成probeIDsWithAnnotation

fData(ALL) <- fData #数据匹配

groupingVariable <- pData(ALL)$BT #将变量分组：B = B细胞，T = T细胞

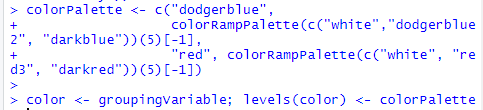
remissionType<-ifelse(is.na(ALL$remission),"unknown",as.character(ALL$remission))

#统计病情缓解类型，ifelse函数，如果ALL数据包中remission的数据为空，输入“unknown”，如不为空，正常输出

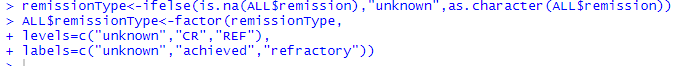
ALL$remissionType <- factor(remissionType, levels = c("unknown", "CR", "REF"), labels = c("unknown", "achieved", "refractory"))

#创建病情缓解类型用于之后作图，将三种情况 "unknown", "CR", "REF"更改对应"unknown", "achieved", "refractory"。

二、创建自定义调色板



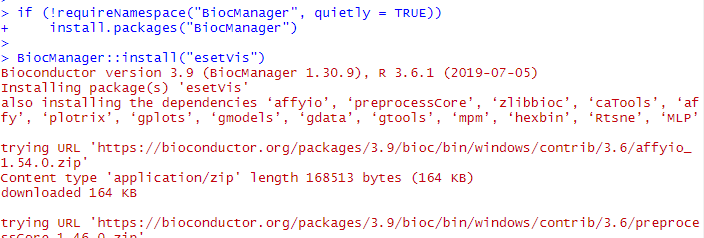
#自定义调色板，选择颜色、函数表面颜色用于分组。



#重新格式化缓解类型

三、作图

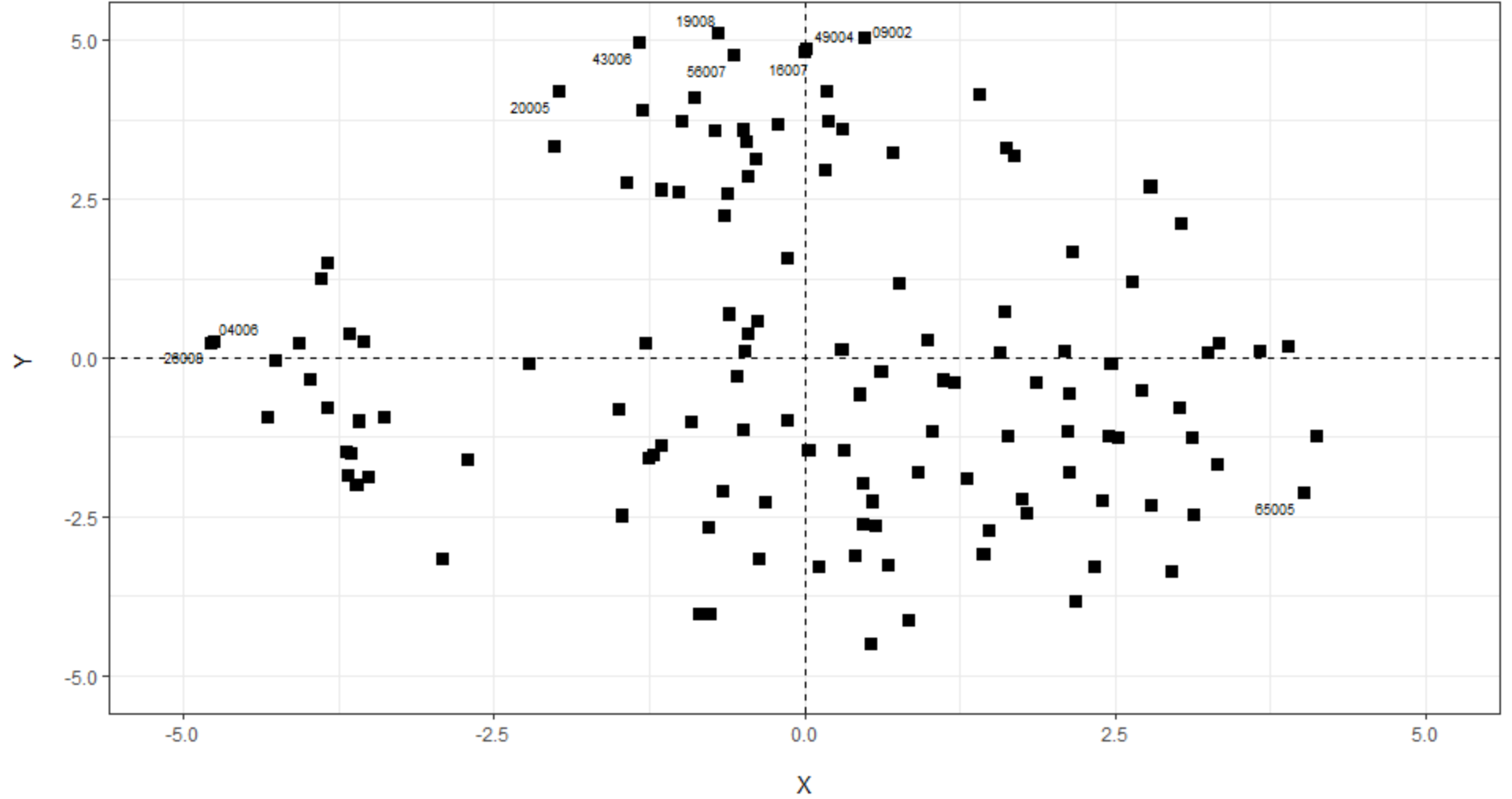
（一）进行t分布随机相邻嵌入esetTsne



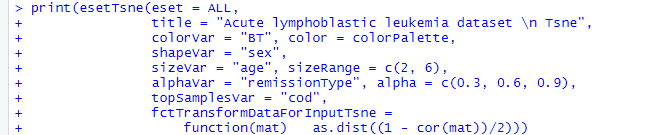
#从bioconductor下载esetVis数据包



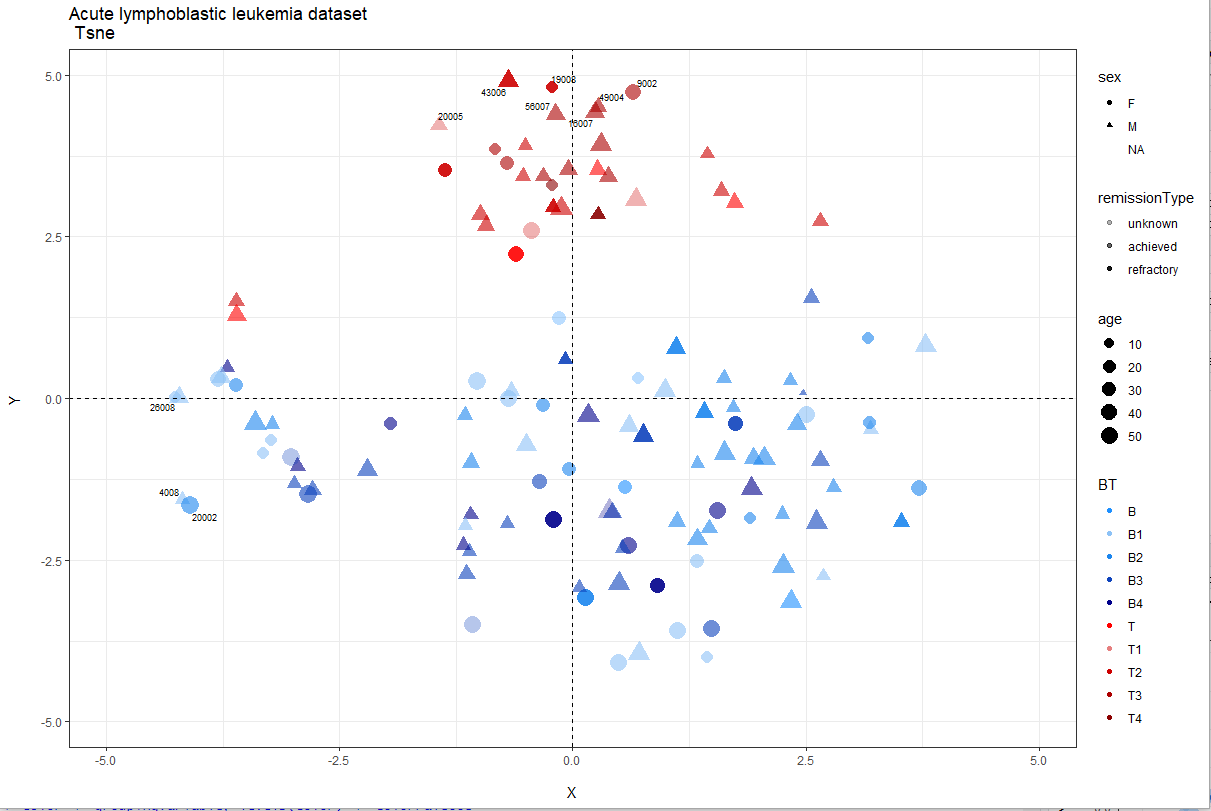
#调出esetvis数据包。使用默认参数将ALL中数据可视化



分析：将样品数据可视化，但不能从其中得出样品的性别、年龄、疾病类型、治疗效果等信息。

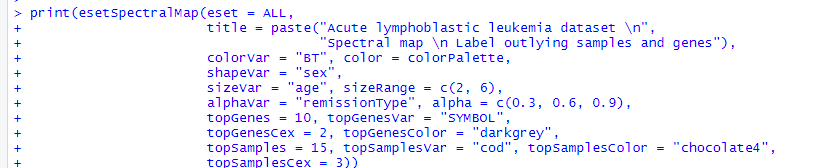


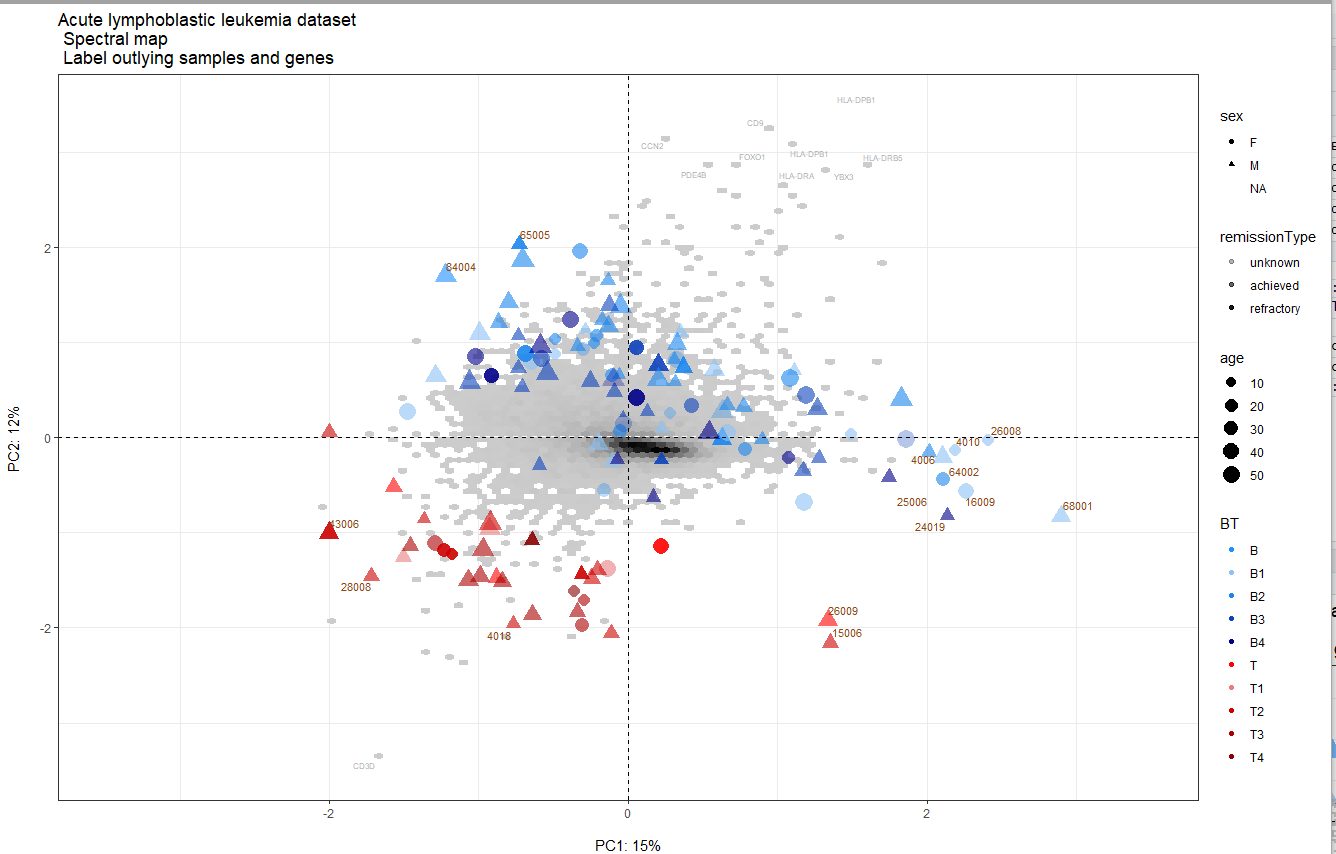
#esetTsne可将高维度数据降维，并可在散点图中进行可视化。



分析：图中圆形代表女性、三角形代表男性，透明表示缓解情况未知、半透明为病情症状减缓、不透明为难以治愈，图形面积大小表示年龄、面积越大年龄越大，颜色代表病症类型、蓝色为B淋巴细胞白血病、红色为T淋巴细胞白血病。相似的对象距离较近、不相似的点距离较远。例如图形下方为蓝色(B淋巴细胞白血病)，而上方多为红色（T淋巴细胞白血病）。

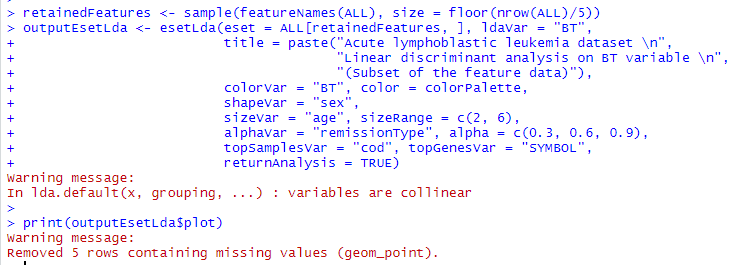
（二）制作esetSpectralMap，并标记最外围元素

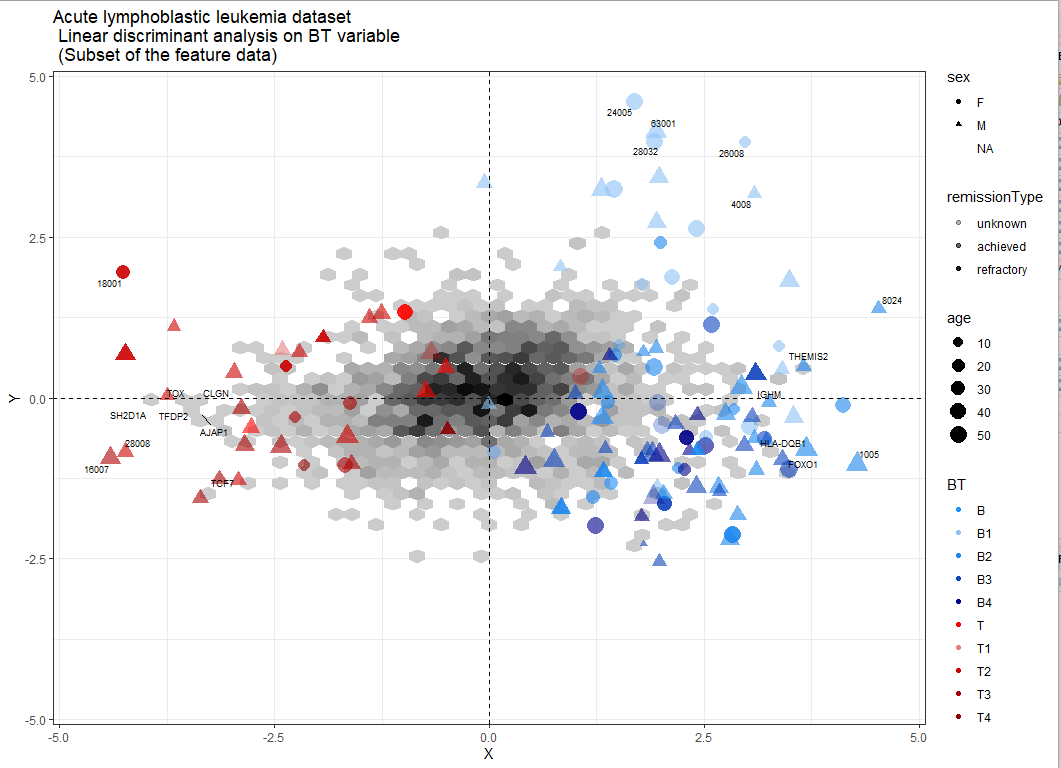


#将基因和样品放入同一图，并标记最离散的10个基因和15个样品

分析：图中包括样品与基因，前景图代表样品，右边为样品图形注释信息，灰色背景代表基因。根据样品与基因到起点的距离对元素进行排序，提取最离散的元素。图形中央为原点，并对最离散的10个基因（集中在图形右上角）和15个样品（分布于图形四周，右边偏多）在图上进行了标记。

（三）进行线性判别分析:esetLda



#提取随机特征、定义图上各数据的性质，其中对BT变量进行线性分析，最后提取并打印ggplot对象。

分析：对BT变量进行线性判别分析，根据数据自身的相似度进行聚类，原则是组内距离最小化和组间距离最大化。图中红色的标记代表T淋巴细胞白血病，蓝色标记代表B淋巴细胞白血病，红色区域与蓝色区域划分比较明显，在各区域内，最相似的各个样本之间有着最小的组内距离，在两个区域之间，差异最大的样本之间有着最大的组间距离。例如T淋巴细胞白血病位于左边、B淋巴细胞白血病位于图像右边，其中缓解情况未知的集中于右上角、症状缓解与难以治愈的位于右下角，使得各个样本之间达到了最佳的分离性。