厦門大學

本 科 课 堂 论 文 （设 计）

（辅修专业）

单细胞测序问题探究

——基于高维非平衡小样本数据的二分类问题

姓　　名：王力为

学　　号：19020152203156

学　　院：王亚南经济研究院（辅修）

专　　业：金融学（数理）

年　　级：2016级

课　　程：数据挖掘

任课教师：方匡南　教授

二〇一八年六月二十一日

1. 背景介绍

本文数据挖掘的背景是单细胞RNA测序数据。单细胞RNA测序数据可以通过单细胞水平上基因表达的分子表征来洞察正常的细胞功能和各种疾病状态。单细胞RNA表达分析（scRNA-seq）是近年来兴起的技术，正在彻底改变整个生物体科学，可以在细胞水平上无偏地识别先前未鉴定过的分子异质性，提供了解剖复杂组织和特定细胞环境组成的手段。

基于单细胞RNA测序数据，我们能够从中得到什么？以该类型为典型的生物统计类数据本身有何特点？我们是否能成功分类出正常细胞和各种疾病状态？我们能否成功找出有显著影响的基因？本文基于以上出发点，对病变细胞的分类及显著影响基因进行数据分析，从数据角度为单细胞测序问题提供参考。

对本文需解决的主要问题和挑战进行描述：

- 有监督分类问题。分类首先是二分类，区分病变与不病变。在此基础上还潜在为多分类问题。本文所做的暂时是二分类。
- 最终目的不仅要成功分类（减轻人工压力），还要具有可解释，即找到显著影响分类的基因变量。在此基础上，还要确定分类阈值。
- 数据为非平衡+高维+零膨胀数据，并且小样本——这仍是一个业界难题。进行合适的分类及变量筛选的建模需要综合尝试。
- 即便从数据中得到了结果，即选出显著影响分类的基因变量，仍需要有较强大的生物学背景知识支持结论。这将是后续进一步要突破的地方。

2. 方法综述

基于单细胞RNA测序数据的一般特点是：高维、非平衡、零膨胀、小样本，还有潜在的噪声、及不符合IID样本数据的假设。对此类数据进行分类难度不小。其中，分类问题决定了这是一个有监督问题，小样本决定了不能使用太过于复杂的分类算法，比如神经网络。零膨胀可能会影响分类问题的估计，要进行一定的处理（本文处理方法为筛选删除。后续还将考虑其他方法）。非平衡和高维数据的问题则由如下叙述逐个解决。数据噪声及IID假设则将在后续继续深入的思路部分的文献方法中提到。

2.1 非平衡问题

非平衡数据的分类学习是机器学习和数据挖掘中的重要问题。在非平衡问题不解决前，所有树方法、集成方法、SVM方法都是不能用的。如果数据存在严重的不平衡，预测得出的结论往往也是有偏的，即分类结果会偏向于较多观测的

类。对于这种问题该如何处理呢？最简单粗暴的办法就是构造1:1的数据，要么将多的那一类砍掉一部分（即欠采样），要么将少的那一类进行Bootstrap抽样（即过采样）。但这样做会存在问题，对于第一种方法，砍掉的数据会导致某些隐含信息的丢失；而第二种方法中，有放回的抽样形成的简单复制，又会使模型产生过拟合。为了解决数据的非平衡问题，2002年Chawla提出了SMOTE算法，即合成少数过采样技术，它是基于随机过采样算法的一种改进方案。该技术是目前处理非平衡数据的常用手段，并受到学术界和工业界的一致认同。

SMOTE算法的基本思想就是对少数类别样本进行分析和模拟，并将人工模拟的新样本添加到数据集中，进而使原始数据中的类别不再严重失衡。

- 采样最邻近算法，计算出每个少数类样本的K个近邻；
- 从K个近邻中随机挑选N个样本进行随机线性插值；
- 构造新的少数类样本；
- 将新样本与原数据合成，产生新的训练集；

R中的主要参数有如下三个：perc.over:过采样时，生成少数类的样本个数;k:过采样中使用K近邻算法生成少数类样本时的K值，默认是5；perc.under:欠采样时，对应每个生成的少数类样本，选择原始数据多数类样本的个数。

目前非平衡解决的问题主要针对二分类，将算法扩展为能呈现多类分类形式，是未来需深入研究的问题。同时，如何区别少数类数据和噪声数据，也有待进一步研究。

注：其他类型做法如，杠杆、异常点检测方法等，泛化能力更好；比如说单类SVM检测。

2.2 高维数据变量筛选问题

基因数据维度动辄上万，相比起几百个细胞的样本量，算是一个非常明显的高维问题。对高维数据进行变量筛选的方法有如下几类：

1. 逐步回归法。但不用STEP是因为维度太大，STEP极易陷入局部最优解。

2. 主成分法。但本文也不打算使用。不用主成分降维是因为变量筛选要具体到某一个基因，所以主成分出来的意义不大。当然除非有非常强的生物学解释意义，不排除这种做法。可后续尝试。

3. 各种shrinkage + LR(Logistics)方法。本文主要采用的方法，使用了SCAD及Lasso族，如elastic net，adaptive Lasso，group lasso等。

4. 决策树（带剪枝）。决策树用于变量选择时，在生物基因表达数据上拥有

更加便于解释的优点。但在非平衡问题上缺点暴露地一览无遗。因此要在SMOTE的基础上进行建模。后续还要考虑进行数据降噪。

5. 本文还将集成方法：如RandomForest、Adaboost里按照Gini系数计算的变量Importance排序选出来的基因加入参考。

本文主要使用是后三种方法，并将选出的变量结合生物学知识进行综合考虑。

3. 数据说明

3.1 原始数据说明

本次分析本次实验单细胞测序的基因表达数据，420个细胞是单细胞，共16927个基因。402个非正常实验组（病变），即CLP，18个正常组（不病变），即sham。合起来是一个420*16927的高维宽矩阵。

具体变量说明如表 1 所示。

表 1 数据说明表

	变量名	详细说明	取值范围	变量个数
因变量	Condition	二分类数据	CLP、sham	**420，比例达402:18**
自变量（基因表达数值）	**X0610005C13RIK、NDC80等**	**连续型数据（零膨胀）**	**≥0的正实数**	**（维度）16926**

对420个细胞的基因表达数值的均值和标准差计算如图1所示。其中红点为少样本（sham）。

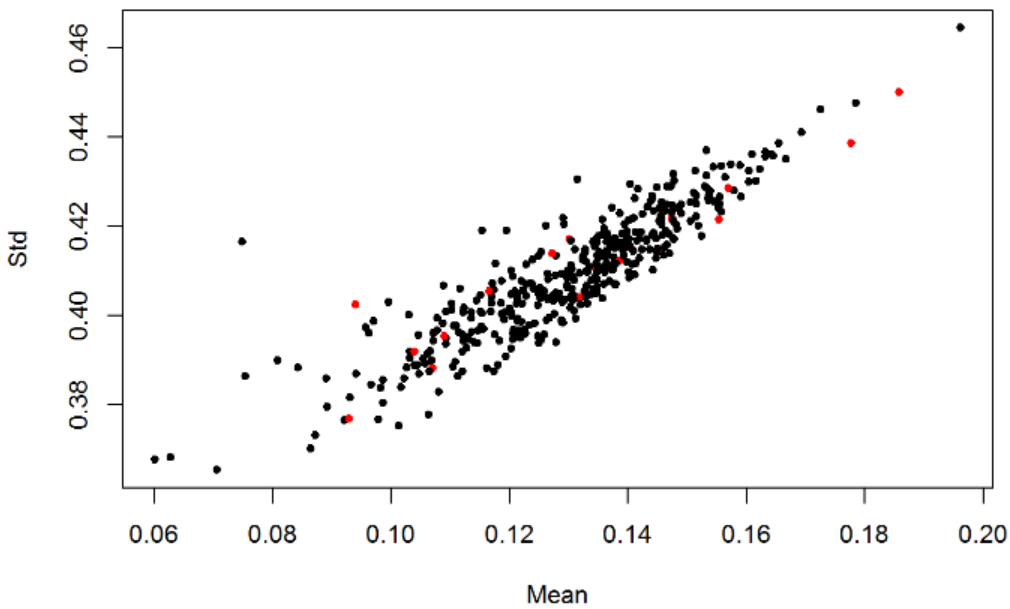

图 1 样本均值及方差分布

3.2 数据预处理与划分

- 去掉基因数据中，非0元素个数小于样本数量5%（即约420*5%＝20为阈值）

的列。适当减轻零膨胀问题。维度从16926变到6593。对该数据进行TSNE非线性聚类的可视化后如图2.其中绿色为sham，少样本点

- 检查并发现没有缺失值
- 数据划分：拆分成训练集和测试集。采用分层拆分，分别把CLP和sham两部分的三分之一分给测试集，三分之二为训练集。此时以测试集为例，共有140个样本，其中sham（少样本）仅有6个。
- 将sham设为0，CLP设为1
- 为防止数据随机性，本文所有种子都设置为1
- 调整参数，采用SMOTE重抽样后产生SMOTE数据集，正负样本CLP：sham变为240：239近似于1:1

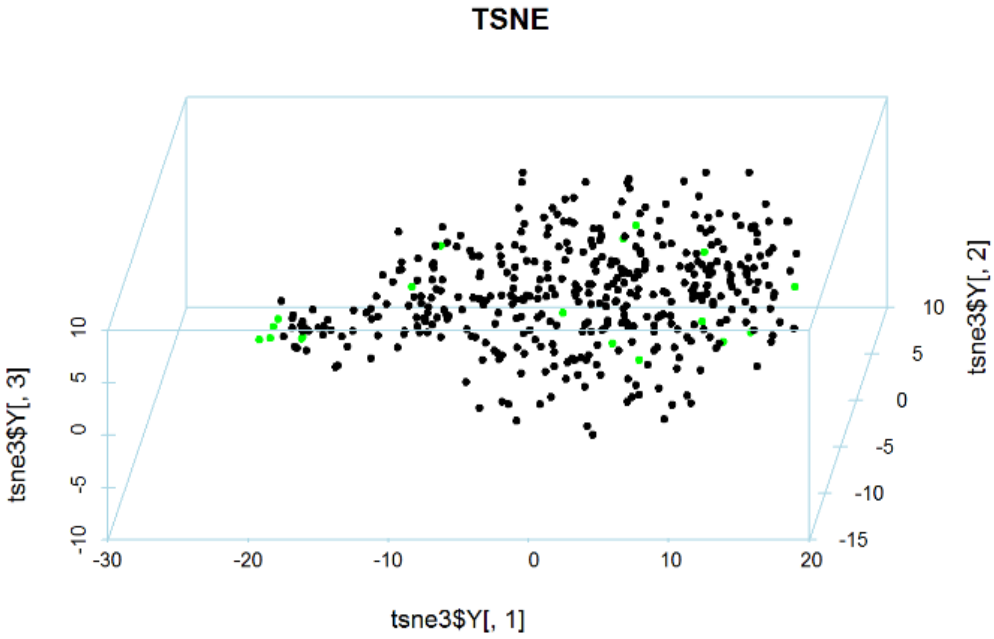

图 2 使用TSNE非线性聚类后的可视化

4. 非平衡数据建模

4.1 Shrinkage 方法

首先对原始数据做分类建模。首选Shrinkage+LR方法，包括SCAD及Lasso族，如elastic net，adaptive Lasso，group lasso。数据是标准化后的数据（自变量。因变量0-1不做标准化）。使用Cross Validation选择最佳压缩参数α。对于elastic net，在同一个随机种子下，调整超参使得每次压缩得到的非零系数稳健且在5~10个左右。Adaptive Lasso的权重由岭回归得到。通过ROC包可同时画出对测试集测试的ROC图、AUC值和基于Yuden index给出的最优分类阈值。如图3所示：

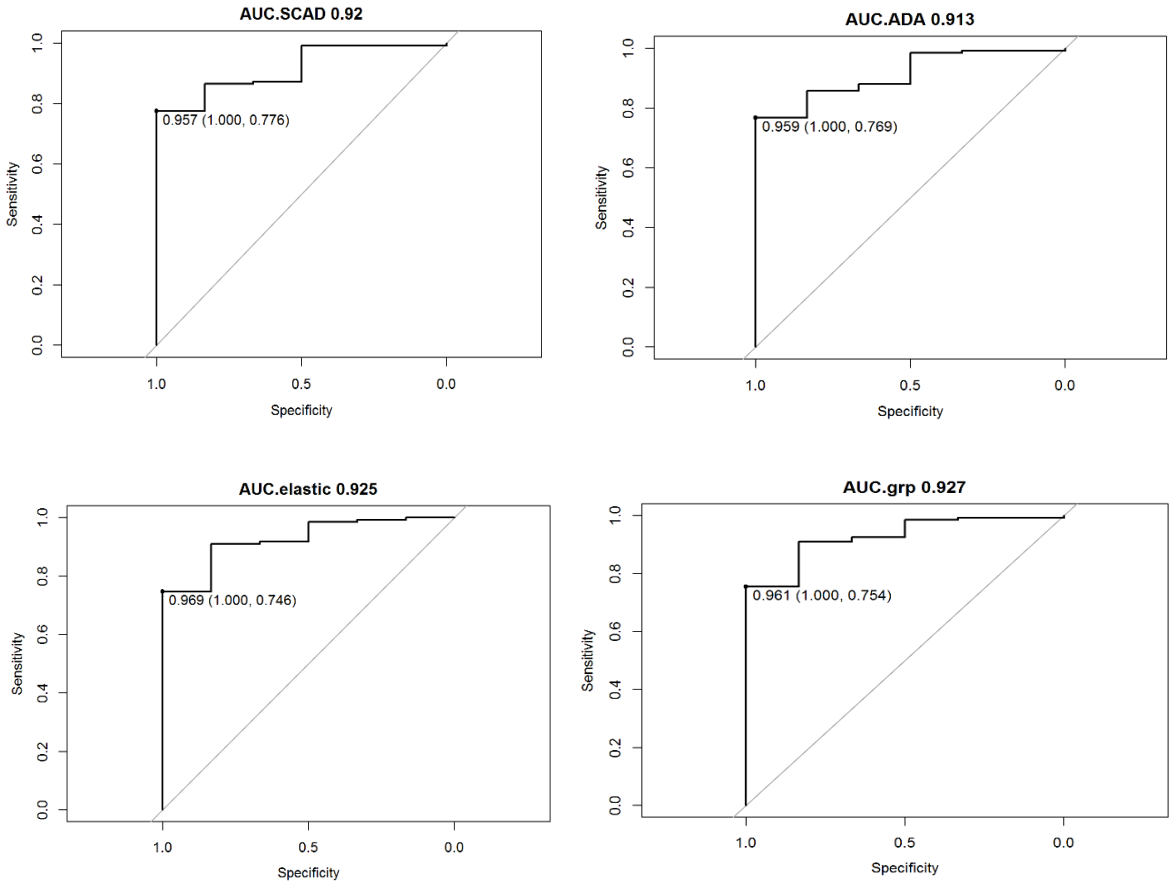

图 3 非平衡样本下的Shrinkage效果

对测试集的分类效果还算不错，其中少数类样本（6个）都是全对（1-FPR=0）；当然对正样本的分类效果还不够优（sensitivity即TPR）。变量选择方面，SCAD压缩和三种LASSO族筛选出来的完全不同（后三者重叠度很大；且也尝试了随机种子为其他的情况下，结果仍然相似。显示了稳健）。对于这种不同方法选出不同变量的情况，需要结合生物学知识做进一步分析。

注：
LASSO族选出：
"X0610009B22Rik" "X1110004F10Rik" "X0610009O20Rik" "X0610030E20Rik"
"X0610007P14Rik" "X0610037L13Rik" "X0610010K14Rik" "X1110037F02Rik"
"X1110012L19Rik""X1110008F13Rik" "X1110008P14Rik" "X0610012G03Rik"等。
三个 LASSO 筛选出来的基因都是 RIK 族，只是多与少的问题

SCAD 方法选出：
Calr Gzma Ccl5 Hspa5 Snx19 Eif1
Axin1 Nt5c3 Hacd4
所有基因皆按系数绝对值大小排序。在标准化数据下即为其重要性

4.2 树与集成方法

前文已说明，对于非平衡数据，树方法和集成方法都是不能使用的，不过基于GINI系数计算的变量Importance排序也可以给我们一个变量筛选的启示。对现有数据做剪枝后的决策树分类及随机森林的Importance结果如图4所示：

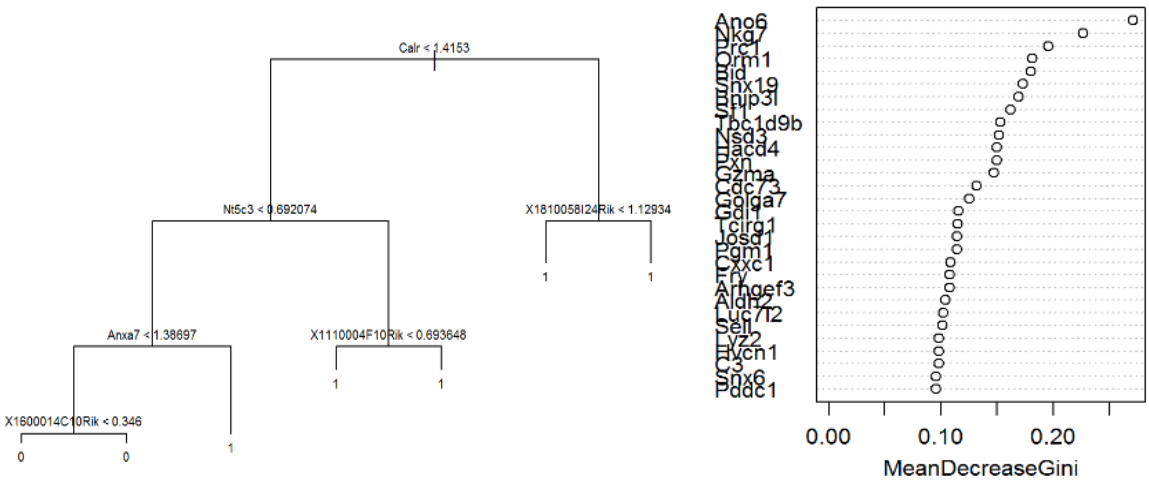

图 4 非平衡样本下的树与集成效果

AUC值毫无疑问的为0.5，侧面反映出非平衡数据下分类是很有问题的。不过，结合这里的Importance排序（不过分类都只有0.5，这里的排序和决策树的变量选择其实也没啥意义），我们尝试了一下，将这里出现的基因都选入Logistics后再进行STEP回归看看结果。最终的AUC达到0.75左右。虽然比直接决策树的0.5要高不少，但还是比之前的Shrinkage+LR的方法低得多。这个尝试至少给了个启示，就是这两个方法的变量选择，不是一个体系的。树和集成方法选出的变量是非线性的关系，不能直接放在LR这个线性模型里。

注：这其实有个XGBOOST非线性one hot encoding变换后再加入到LR的方法。（业界常用）后续可以一试。

5. 解决非平衡问题后的建模

使用SMOTE方法，原数据成功变成了240：239，近似于1:1的数据。使用TSNE对该数据进行可视化如图5（绿色为sham，黑色为CLP），可以发现此时两类数据分的比较开了：

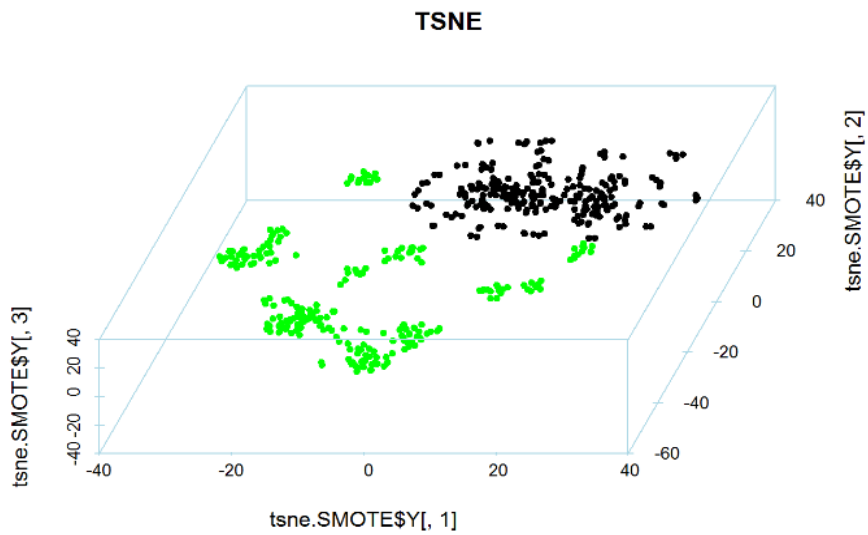

图 5 对SMOTE数据使用TSNE的可视化

正常组更进一步出现了多聚类的问题，这和后续要做的对病变组做多分类的思路有所对立。后续准备进一步讨论。

5.1 Shrinkage 方法

对SMOTE重抽样数据做分类建模。同样先尝试Shrinkage方法。将前面几种方法用于SMOTE数据后，得到的AUC结果如图6：

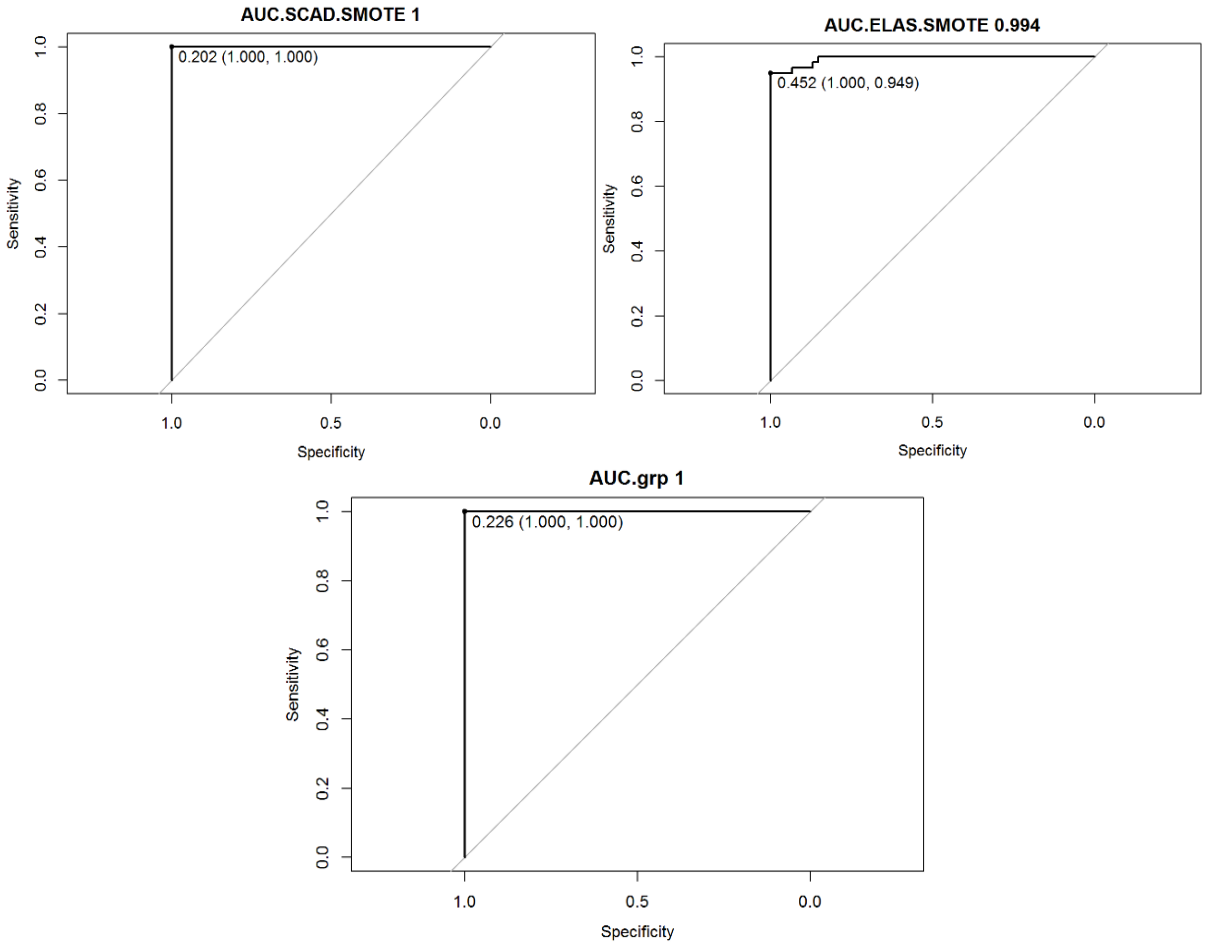

图 6 平衡样本下的Shrinkage效果

对测试集的分类效果已经近乎完美了。然而在变量选择方面，由于不清楚SMOTE后数据的改变会不会对本来是非平衡条件下的变量选择有所影响，因此我们还不能直接放弃SMOTE前的选择变量。SMOTE后的变量选择如下：同样存在SCAD压缩和两种LASSO族筛选出来的完全不同的情况（后两者重叠度很大，显示了稳健）。需要结合生物学知识做进一步分析。

注：
LASSO族选出：
"X0610012G03Rik" "X1700047M11Rik" "X1600014C10Rik" "X0610030E20Rik"
"X0610010K14Rik" "X1700123O20Rik" "X2210016L21Rik" "X1110038F14Rik"
"X1700020I14Rik" "X0610037L13Rik" "X1810022K09Rik" "X1600010M07Rik"等
三个 LASSO 筛选出来的基因都是 RIK 族，只是多与少的问题

SCAD 方法选出：
Ccl5 Calr Lgals3 Hspa5 Gpatch2l Axin1
Ipo8 Klf2 Chmp5 Naca Lilrb4a Nt5c3 等
所有基因皆按系数绝对值大小排序。在标准化数据下即为其重要性。

5.2 树与集成方法

在SMOTE后的数据中，树与集成方法表现出其优良性。普通的决策树经过剪枝下已能达到测试集AUC为0.967的效果（其他所有集成方法、包括另外尝试的SVM都能达到1），且在不同随机种子下剪枝AUC和变量选择甚至阈值都保持稳健。因此该决策树选择出来的基因已有一定的意义。且由于前文所述，决策树在基因数据中可解释性更优，所以后续会作为首要参考。

不过，决策树仍然存在着极容易受异常点干扰的影响，在未去除本数据噪音的前提下还存在一定缺陷。我们用更加稳健的集成方法（这里以Random Forest为例）的Importance Gini排序作为参考。

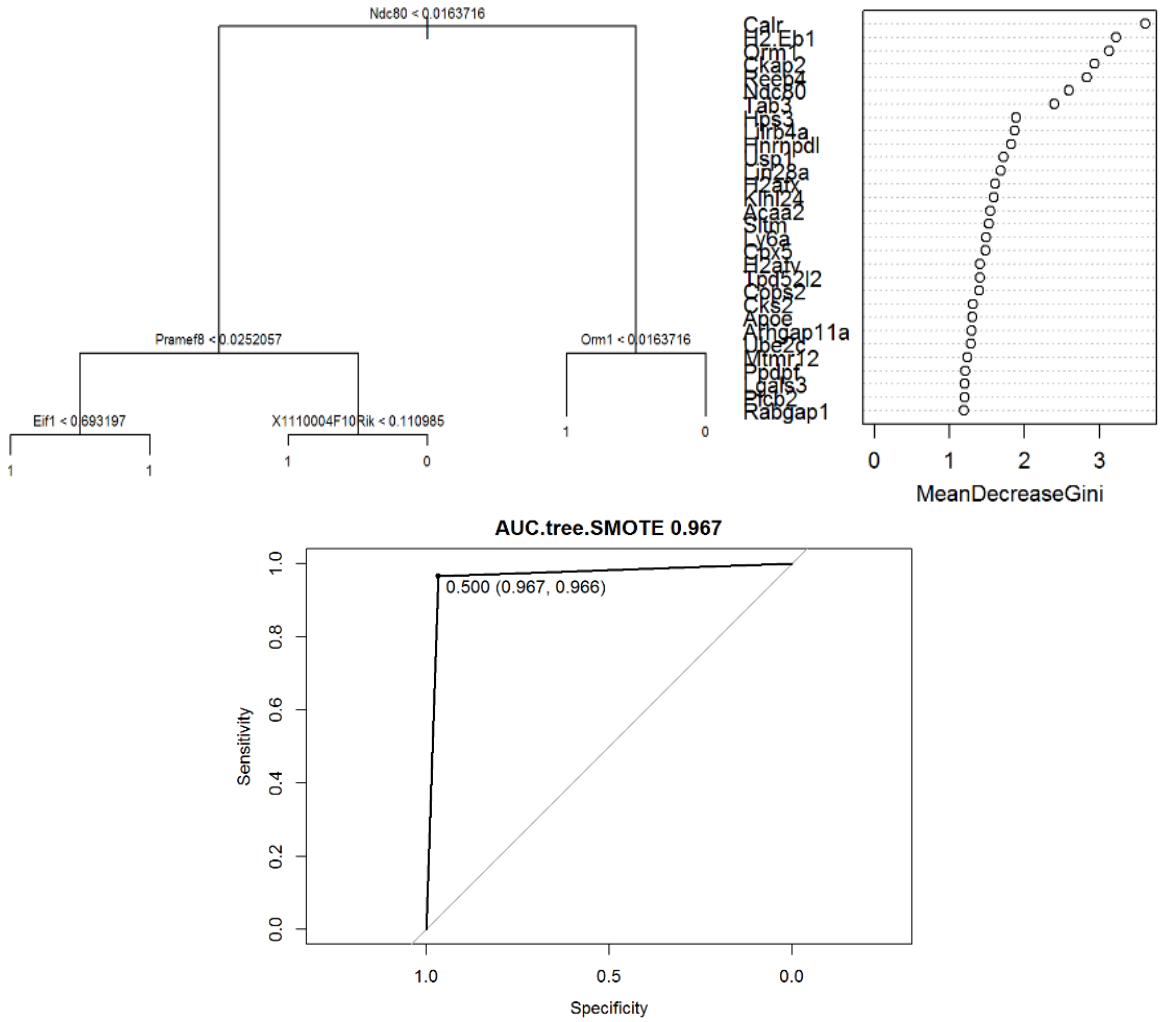

图 7 平衡样本下的树与集成效果

注：
决策树选出：
“Ndc80” “Pramef8” “Eif1” “Arl6ip1” “X1110004F10Rik” “Orm1”等
随机森林 Importance-Gini 的排序选出：
Calr Orm1 Ndc80 Usp等

6. 总结

（1）分类问题已基本解决

（2）变量选择问题：有可能的特别基因：

TREE方法："Ndc80" "Pramef8" "Eif1" "X1110004F10Rik" "Orm1“

RF方法：Calr Orm1 Ndc80 Usp等

SCAD: Ccl5 Calr Lgals3 Hspa5 Gpatch2l等

LASSO: RIK族

（3）仍旧有挺多数据假设和方法上的BUG需要后续多加验证和补充

6.1 基因变量总结

本文从Gene Card官网搜寻和基因有关的资料，有个初等印象。后续还得多和生化那边的老师沟通学习。举例如下

- NDC80:与染色体分离有关。This gene encodes a component of the NDC80 kinetochore complex. The encoded protein consists of an N-terminal microtubule binding domain and a C-terminal coiled-coiled domain that interacts with other components of the complex. This protein functions to organize and stabilize microtubule-kinetochore interactions and is required for proper chromosome segregation. [provided by RefSeq, Oct 2011]
- Calr：转录调控有关。Calreticulin is a multifunctional protein that acts as a major Ca(2+)-binding (storage) protein in the lumen of the endoplasmic reticulum. It is also found in the nucleus, suggesting that it may have a role in transcription regulation. Calreticulin binds to the synthetic peptide KLGFFKR, which is almost identical to an amino acid sequence in the DNA-binding domain of the superfamily of nuclear receptors. Calreticulin binds to antibodies in certain sera of systemic lupus and Sjogren patients which contain anti-Ro/SSA antibodies, it is highly conserved among species, and it is located in the endoplasmic and sarcoplasmic reticulum where it may bind calcium. The amino terminus of calreticulin interacts with the DNA-binding domain of the glucocorticoid receptor and prevents the receptor from binding to its specific glucocorticoid response element. Calreticulin can inhibit the binding of androgen receptor to its hormone-responsive DNA element and can inhibit androgen receptor and retinoic acid receptor transcriptional activities in vivo, as well as retinoic acid-induced neuronal differentiation. Thus, calreticulin can act as an important modulator of the regulation of gene transcription by nuclear hormone receptors. Systemic lupus erythematosus is associated with increased autoantibody titers against calreticulin but calreticulin is not a Ro/SS-A antigen. Earlier papers referred to calreticulin as an Ro/SS-A antigen but this was later disproven. Increased autoantibody titer against human calreticulin is found in infants with complete congenital heart block of both the IgG and IgM classes. [provided by RefSeq, Jul 2008]
- Pramef8 也为蛋白质编码基因

6.2 后续继续深入的思路

如果仅是project，本文可以算是结束了。如果是针对问题的科研，本文才是起了个头。漫漫无期啊。

列举了一些目前已有的文献。对后续的研究思路做一些参考：

（1）解决非平衡问题的方法（SMOTE， V-synth 合成方法等）解决非平衡的合成采样方法非常多。SMOTE只是其中特别常用的一种。可以考虑该文献提到的新方法

（2）解决测序数据噪声。

————Multiplatform single-sample estimates of transcriptional activation

————Bayesian approach to single-cell differential expression analysis　这一篇被引用得较多

（3）新的聚类方法（同时也提供高位数据下的可视化工具）————sc3: consensus clustering of single-cell rna-seq data。提供了一种代替多分类问题的聚类方法

（4）针对零膨胀问题的数据降维（解决高维问题，大量dropout导致零膨胀，之前都没有做过处理）————ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis

（5）针对数据本身，可能来自于不同细胞阶段，需要鉴定亚群（即不是IID数据）————Computational analysis of cell-to-cell heterogeneity in single-cell RnA-sequencing data reveals hidden subpopulations of cells。提供了一种解决非IID假设的方法

参考文献

[1] Nykl S L, Weckman G R, Chelberg D M. Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets[J]. Neural Computing & Applications, 2015, 26(5):1041-1054.

[2] Piccolo S R, Withers M R, Francis O E, et al. Multiplatform single-sample estimates of transcriptional activation.[J]. Pnas, 2013, 110(44):17778-17783.

[3] Kharchenko P V, Silberstein L, Scadden D T. Bayesian approach to single-cell differential expression analysis[J]. Nature Methods, 2014, 11(7):740.

[4] Kiselev V Y, Kirschner K, Schaub M T, et al. SC3 - consensus clustering of single-cell RNA-Seq data[J]. Nature Methods, 2017, 14(5):483-486.

[5] Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis[J]. Genome Biology, 2015, 16(1):241.

[6] Buettner F, Natarajan K N, Casale F P, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.[J]. Nature Biotechnology, 2015, 33(2):155-60.

[7] Galar M, Fernandez A, Barrenechea E, et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches[J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2012, 42(4):463-484.

[8] Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291(5):184-203.