数学建模实践——课程期末大作业

小组成员：1952241 张家瑞　2051495 徐奕　2053410 胡孝博

0.1 研究课题

有效地分析基因数据有利于疾病的诊断和预防，改变基因表达的突变可能在复杂疾病中发挥重要作用。如果健康人群检测到携带疾病的易感基因，可以尽早采取预防措施以降低患病风险。

为了获取与人类的红眼病相关的基因，生物学家在 120 只 12 周大的小鼠上进行了实验，获得了 18976 个基因位点的遗传数据，使用实验室小鼠基因位点对应的表达数量性状基因座作图来获得哺乳动物眼中基因调控的广阔视角，以深入了解人类眼部疾病相关的遗传变异调控机制。研究发现位于 1389163_at 上的基因对造成老鼠红眼病至关重要，并且跟其他基因位点上的基因关系是非线性的。

本题要求要求根据附录中的数据，建立合适模型来解决以下问题：

（1）分析哪些位点上的基因对 1389163_at 上的基因表达最有影响，并给出选择有影响的基因位点的衡量方法。

（2）确定如何对有影响的基因位点进行排序。

（3）对 1389163_at 上的基因建立一个合适的模型，分析所需要最少的自变量个数，并确定一个合理的决定自变量个数的准则。

0.2 课题说明

主题源于 2022 年同济大学数学建模竞赛 A 题。

我们三人组队参赛，得分排名前 10%，获本科生二等奖。

在此基础上，我们根据大作业要求，具体结合课程的知识体系，又进行了相应的问题设计与建模求解。

0.3 小组成员及分工情况

1952241 张家瑞：问题分析、算法与建模思路、模型评价、人员协调。

2053410 胡孝博：建模方案设计、代码、数据处理与可视化、结果分析。

2051495 徐奕：文献资料搜集、论文与报告撰写、算法优化、项目进度管理。

0.4 附件说明

code 文件夹：源代码 (编程语言：python 与 matlab)，以 main.ipynb 为核心。

data 文件夹：建模过程中的数据记录、模型参数等。

figure 文件夹：流程图、数据图表等。

基于 XGBoost 和核主成分分析的基因位点选择模型

摘要

基因数据的有效分析有利于疾病的诊断和预防，并且可以对携带易感基因的健康人群起到提示作用。生物学家通过研究老鼠红眼病基因以及基因间表达的相关性，以获取人类红眼病有关基因。本文针对相关性分析与特征选择问题，首先对数据进行了奇异值处理，KS 正态检验和标准化，通过 Spearman 系数、互信息与最大互信息系数进行相关性初步筛选。随后基于 XGBoost 模型进行回归预测，给出不同特征对目标基因影响的量化权值并进行排序。最后利用核主成分分析法对数据进行降维筛选，并通过递归特征消除进行特征选择，利用交叉验证评估模型质量，给出最少自变量个数及基因位点。

针对问题一，我们通过阅读相关文献，探究了附表的数值含义和研究基因表达的现有方法。首先，为初步了解基因表达量的数据分布情况，我们进行 **KS 正态检验**，设置合理的方差阈过滤掉极少数基因，并对剩余基因进行 **Zscore 标准化处理**。在此基础上计算各基因与 1389163_at 上基因的 **Spearman 系数**、**互信息以及最大互信息系数**，共同作为判据，检验基因间相关性是否显著从而进行初步筛选。此外，我们采用**典型相关分析**的方法作为补充，对筛选结果进行了有效验证。最终，我们从数据中挖掘出对 1389163_at 上基因表达**影响显著的基因位点 224-510 个**。

针对问题二，结合问题一初步筛选的基因位点，我们采用 **R 型聚类**方法初步判断相似基因的分布特征，并对聚类结果进行了可视化。之后，我们基于 **XGBoost 算法**对不同特征（基因）进行回归拟合，并给出筛选出高相关性的特征并给出相关权值。跟据此权值进行重要度排序预筛选，最终发现 **1382223_at，1368136_at，1388404_at，1393231_at，at1383249_at 等 25 个基因**对表达有显著影响。

针对问题三，我们对 XGBoost 由重要度排序所得到的 25 个基因进一步探讨。首先，我们基于传统的**主成分分析（PCA）**的思想，采取**核主成分分析 (KPCA)** 进行**非线性降维**。在降维分析的过程中，我们创新性地以 **Adjusted R-square** 为评价指标，基于样本数量的增加，给予一个正则项，最终发现**自变量个数的较合理范围是 10-15**。在此基础上，我们采用**递归特征消除 (RFE)**，以训练好的 XGBoost 模型作为评估器，结合**交叉验证得分**进行变量的逐一消除，同样得到自变量个数为 **11-17** 时，对模型有最小的性能损失。因此，依据上述两种算法与两种标准，我们认为模型建立所需的**最小自变量个数为 11**。由此重新建立 XBGoost 回归模型，采用 **k 折交叉验证**，预测误差小于 5%，验证了模型的有效性。

关键字：**Spearman 系数　典型相关分析　R 型聚类　XGBoost 回归　核主成分分析　递归特征消除　交叉验证**

一、 问题概述

1.1 项目背景

有效地分析基因数据有利于疾病的诊断和预防，改变基因表达的突变可能在复杂疾病中发挥重要作用。如果健康人群检测到携带疾病的易感基因，可以尽早采取预防措施以降低患病风险。

为了获取与人类的红眼病相关的基因，生物学家在 120 只 12 周大的小鼠上进行了实验，获得了 18976 个基因位点的遗传数据，使用实验室小鼠基因位点对应的表达数量性状基因座作图来获得哺乳动物眼中基因调控的广阔视角，以深入了解人类眼部疾病相关的遗传变异调控机制。研究发现位于 1389163_at 上的基因对造成老鼠红眼病至关重要，并且跟其他基因位点上的基因关系是非线性的。

1.2 数据描述

数据源于附件中 data 目录下的 rat_eye.xlsx。列标为基因名称，无行标。数据共 18975 行、120 列；一行表示某基因在 120 只样本小鼠上的表达量，一列表示某小鼠 18975 个基因的表达量。数据完整，无缺失。

1.3 问题介绍

（1）分析哪些位点上的基因对 1389163_at 上的基因表达最有影响，并给出选择有影响的基因位点的衡量方法。

（2）确定如何对有影响的基因位点进行排序。

（3）对 1389163_at 上的基因建立一个合适的模型，分析所需要最少的自变量个数，并确定一个合理的决定自变量个数的准则。

二、 问题分析

2.1 问题一分析

问题一要求分析哪些位点上的基因对 1389163_at 上的基因表达最有影响，并给出选择有影响的基因位点的衡量方法。这要求我们根据附件所给的对 120 只老鼠实验获取的 18,976 个基因位点的基因数据表达量数据，进行相关性分析。

作者在 eQTL 方法的基础上，假设基因表达的成对相关性可能揭示生物学上相关的功能关系。他们通过对基因表达的配对分析（Pairwise Analysis of Gene Expression），确

定以协调方式调节的基因。我们延续作者的研究思路。首先通过正态性检验，对各个基因表达量的分布情况有了初步把握。又知红眼基因与各位点对应的基因是非线性关系，于是在对数据进行标准化等预处理操作之后，用 Spearman 相关系数法、互信息和最大互信息系数等方法筛选相关性高的基因，再根据得到的结果进行综合分析，结合典型相关分析方法加以对比验证，从而初步筛选出若干个对 1389163_at 上的基因表达最有显著影响的基因位点。

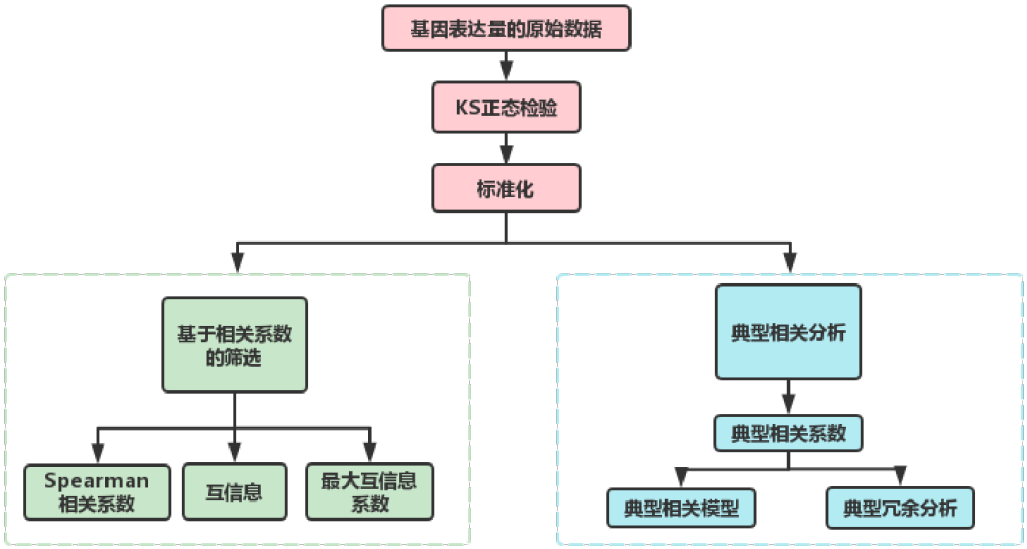

图 1　问题一流程图

2.2 问题二分析

问题一中，我们筛选出了对 1389163_at 上的基因表达有较大影响的基因座数据，本题在此基础上对数据进行分类处理量化，研究各位点对应基因与红眼基因之间的关联性，对有影响的基因位点进行排序。

首先，我们尝试通过 R 型聚类方法，对相似基因的分布情况进行初步分析，并借助降维方法将聚类结果可视化。

在多因素之间的关联性分析的基础上，我们需要进一步通过多元非线性回归以及特征选择的方法，寻求多因素之间，贴近数值的函数关系，进而评价各因素的影响力贡献程度。对于寻求多个解释变量与被解释变量的关系时，我们选用 XGBoost 建立回归模型，其核心思想是通过不断地增减解释变量并反复检验显著性以得到与被解释变量关联最为显著的因素组合，从而计算出各位点基因对红眼基因表达的影响预测值，最终基于该影响预测值排序得到结果。

2.3 问题三分析

问题三分为两个子问题：子问题一是预估对 1389163_at 上的基因建立合适模型所需要最少的自变量个数，子问题二是给出一个决定自变量个数的合理准则。

在问题一和问题二中我们已经完成了 1389163_at 上的基因表达有较大影响的基因座数据的筛选，并根据他们的影响特征大小进行量化排序，在前面的基础上，本问基于主成分分析（PCA）的思想，采用核主成分分析（KPCA）方法对数据进行降维处理，在此同时使用 Adjusted R-square 进行误差计算，根据非显著的变量给出惩罚。通过最终形成的聚簇个数预估所需要最少的自变量个数。接着我们基于自变量个数利用 RFE（递归特征消除）配合交叉验证完成变量的选择。最后对模型进行检验，判断其正确性与合理性。

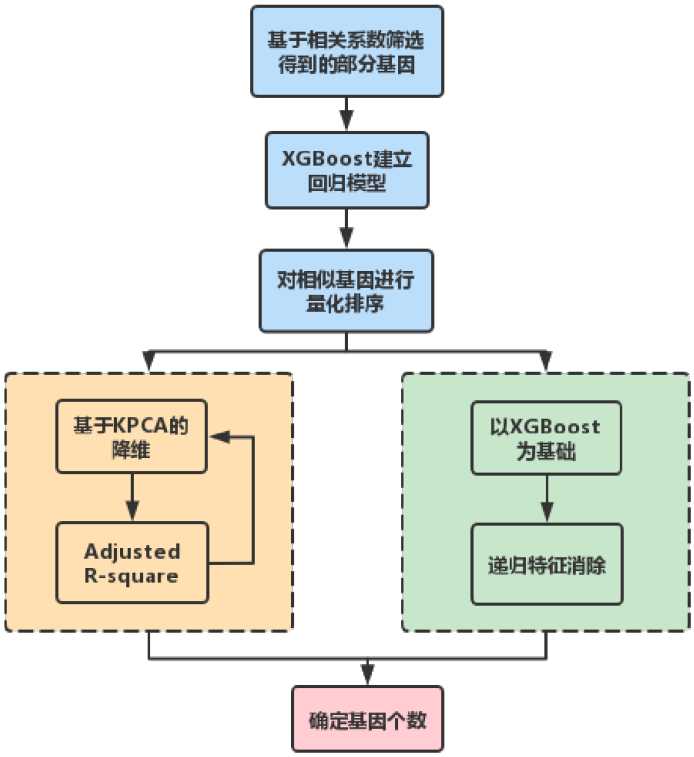

图 2　问题三流程图

三、 模型假设

1. 假设每个位点均可通过孟德尔错误控制检测标准。
2. 假设任何一组与单个转录物的表达有联系的连续遗传标记都是单个 eQTL。
3. 假设位于 1389163_at 上的基因跟其他基因位点上的基因关系是非线性的。

四、 符号说明

表 1　符号说明

x_{ij}	第 i 个小鼠的第 j 基因表达量
$\overline{x_i}$	第 i 个基因在所有老鼠上表达量的平均值
σ_i	第 i 个基因在所有老鼠上表达量的标准差值
${x_{ij}}^*$	第 i 个小鼠的第 j 基因表达量标准化后的值
∂	显著性水平
r_{ij}	第 i 个基因与第 j 个基因之间的 Spearman 相关系数
p	Spearman 显著性统计量
sp	Spearman 相关性系数的指定阈值
I_{ij}	第 i 个基因与第 j 个基因之间的互信息值
IC_{ij}	第 i 个基因与第 j 个基因之间的最大互信息系数
w_i	节点 i 的权值
T	叶子的个数
$\hat{y}^{(t)}$	第 t 轮的模型预测
z_i	原始基因位点
b	样本个数
$k_{AdjustedR^2}$	调整 R^2 的变化率
CVScore	交叉验证得分

五、 问题一模型建立与求解

5.1 数据分布的正态性检验

数据共含 18976 行，表示不同基因位点；含 120 列，表示不同的老鼠样本。通过初步分析，原始数据中不存在缺失项、格式错误与异常值，且度量单位一致（基因表达量）。在此基础上，我们读取 excel 文件，进行了数据矩阵的转置，并将红眼基因座数据移动到最后一列，作为标签数据列。

5.1.1 Kolmogorov-Smirnov 检验

针对原始数据，我们通过 Kolmogorov-Smirnov 检验（简称 KS 检验）判断连续性变量 (基因表达量) 是否严格服从或近似服从正态分布。

由于本题的关键在于基因表达量之间进行相关性分析，而许多常见的统计学方法都要求数据满足正态性，如方差分析（ANOVA）、Pearson 相关系数。在考虑采用这些检验或分析方法时，首先对数据进行正态性检验。如果数据不服从正态分布，直接采用上述参数检验的方法，有可能导致统计效能下降和假阴性风险增加。

KS 检验直接针对原假设 $H0 : F_n(x) = F(x)$，这里分布函数 $F(x)$ 必须是连续型分布。KS 检验基于经验分布函数作为检验统计量，检验理论分布函数与样本分布函数的拟合优度，或比较两个经验分布是否有显著差异。。

5.1.2 KS 检验的步骤

设总体 X 服从连续分布，$X_1, X_2, ..., X_n$ 是来自总体 X 的简单随机样本，F_n 为经验分布函数，根据大数定理，当 n 趋于无穷大时，经验分布函数 $F_n(x)$ 依概率收敛总体分布函数 $F(x)$。因此检验统计量为：$D_n = \sup_x |F_n(x) - F(x)|$ ，其中 $F_n(x)$ 为观察序列值，$F(x)$ 为理论序列值或另一观察序列值。

（1）提出原假设与备择假设

$$H_0 : F_n(x) = F(x)$$

$$H_1 : F_n(x) \neq F(x)$$

（2）计算样本累计频率与理论分布累计概率的绝对差，令最大的绝对差为

$$D_n = max[F_n(x) - F(x)]$$

当 H_0 为真时，D_n 偏小趋势，则拟合得越好；当 H_0 不真时，D_n 偏大趋势，则拟合得越差。

（3）确定拒绝阈。用样本容量 n 和显著水平 α 查 D_n 极限分布表，求出 t_α 满足

$$P\left\{\sqrt{n}D_n \geq t_\alpha\right\} = \alpha$$

作为临界值，即拒绝域为 $[t_\alpha, +\infty)$

(4) 计算统计量的观察值，如果检验统计量 $\sqrt{n}D_n$ 的观察值落在拒绝域中，则拒绝原假设，否则不拒绝原假设。

5.1.3 KS 检验结果与分析

首先，我们逐基因对其表达量在 120 只实验老鼠上的分布进行 KS 检验。

我们对全样本进行基因表达量的带权加和以及 KS 检验，发现结果近似服从正态分布。基因表达量数据的分布情况与正态检验情况如图 3 所示：

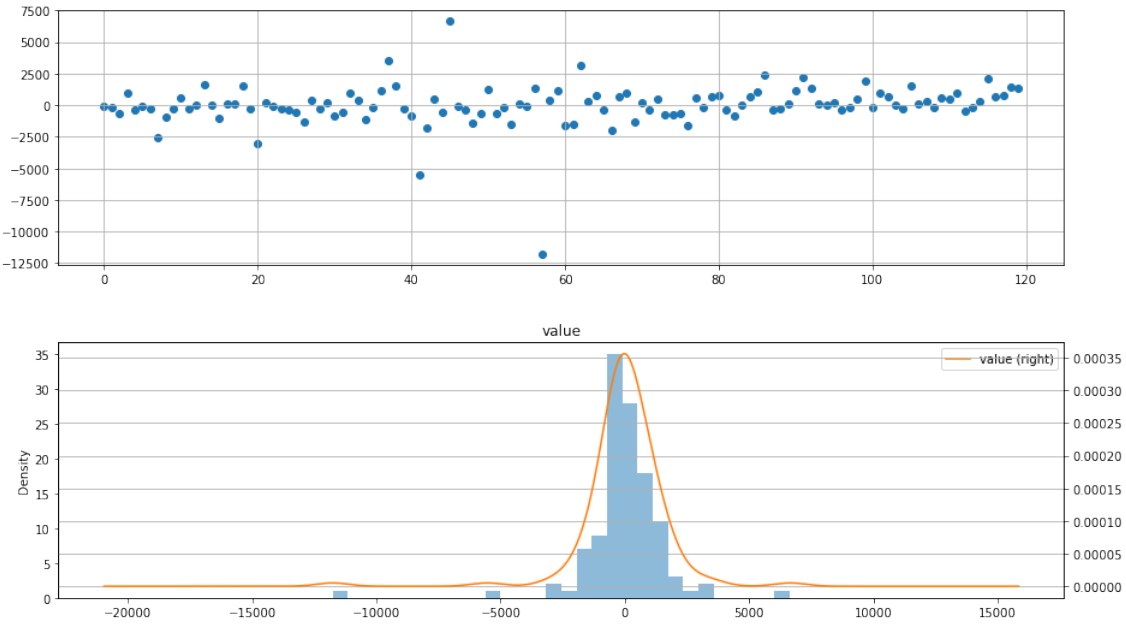

图 3　基因表达量数据分布与直方图

在 *pvalue*>0.01 的显著水平下，18796 个基因位点中，有 17909 个不拒绝原假设，可以认为服从正态分布。

在 *pvalue*>0.05 的显著水平下，18796 个基因位点中，有 16834 个不拒绝原假设，可以认为服从正态分布。

部分数据结果以及 pvalue 如表 2 所示

表 2　KS 检验正态性

pvalue 值	标号	位点	是否拒绝原假设	总计不拒绝原假设个数	正态性
pvalue>0.01	1	1393231_at	否	17909	符合正态分布
	……	……	……		
	20	1367526_at	否		
	21	1367527_at	否		
	……	……	……		
	36	1367553_x_at	否		
	37	1367555_at	是		
	……	……	……		
	18973	1398744_at	否		
	18795	1398745_at	否		
pvalue>0.05	1	1393231_at	否	16834	符合正态分布
	……	……	……		
	20	1367526_at	是		
	21	1367527_at	否		
	……	……	……		
	36	1367553_x_at	是		
	37	1367555_at	是		
	……	……	……		
	18973	1398744_at	是		
	18795	1398745_at	否		

在上述检验的基础上，我们基于特征选择中 Filter 方法——当一个特征本身的方差很小，则表明样本在这个特征上基本没有差异，那这个特征对于样本区分不能提供更多的有效信息。因此，我们采用方差去噪，设置方差阈值为 0.02，去除了部分在全样本上方差表现过小的基因位点。

5.2 数据标准化

各基因表达量之间由于生物学规律以及自身特性，存在着不可公度性，这会影响到数据分析的结果。为了消除基因表达量对基因作用的影响与偏差，我们对数据进行标准化处理，以解决数据指标即基因表达之间的可比性。原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。

Z-score 标准化方法

$$x_i^* = \frac{x_i - \mu_i}{\sigma_i}$$

数据标准化后的结果

数据处理后每个基因的表达量均符服从标准正态分布，即均值为 0，标准差为 1，计算所得部分标准化数据如表所示，全部标准化数据见附件 std.xlsx。

表 3　数据标准化后的结果部分截选

位点	小鼠 1	小鼠 2	小鼠 3	小鼠 4	小鼠 5	...
1367458_at	-0.4917	0.4377	-0.4684	-0.8378	0.2082	...
1367462_at	0.0843	-0.2354	0.3749	-0.252	-0.4678	...
1367467_at	0.5275	-0.7345	-1.0991	-0.0521	-1.2954	...
1367474_at	1.4783	-0.5942	0.1214	1.8564	-1.5798	...
1367478_at	0.6514	-1.0585	0.5133	0.7172	-0.5324	...
1367484_at	-0.1499	0.1693	-0.4802	1.0554	-0.2985	...
1367491_at	-0.213	0.6432	0.0287	0.6501	0.5328	...
1367495_at	-0.3075	0.375	-0.8959	0.8942	0.2909	...
...	...	...	...	...	...	...

5.3 基于相关系数方法的筛选

在文献一中，作者在 eQTL 方法的基础上，假设基因表达的成对相关性可能揭示生物学上相关的功能关系。他们通过对基因表达的配对分析（Pairwise Analysis of Gene Expression），确定以协调方式调节的基因。这种方法背后的生物学原理是，随着生物体的进化，将功能相关基因的表达与需要其功能的生物学情况联系起来具有进化优势。

在本题中，我们采用相关系数分析的方法探究各位点上基因对 1389163_at 上基因表达的影响。题目中确认 1389163_at 上基因与其他基因位点的关系是非线性的，这不符合 Pearson 相关系数、方差分析法（ANOVA）的适用前提。

本文为了保证结果的稳健性，将采用 Spearman 相关系数、互信息（MI）和最大互信息系数（MIC）等方法共同筛选与 1389163_at 相关性高的基因，再根据得到的结果进行综合分析，即对所得数据集取交集。这样既能在最大限度内筛选与 1389163_at 基因具有高度相关性的基因位点，又避免了采用单一方法的过高阈值产生的偏差，使得得到的结论更全面有效。

5.3.1 Spearman 相关系数

相关分析是描述两个变量间关系的密切程度的方法。双变量系数测量的主要指标有卡方类测量、Spearman 相关系数、pearson 相关系数等。Spearman 相关系数侧重检测变量间的单调性关系，这符合基因表达相互影响而导致的表达量的高度正相关或负相关。而在本题中，由于 1389163_at 上的基因跟其他基因位点上的基因关系是非线性的，我们以于 1389163_at 上的基因为固定参考量，其他各位点上的基因作为另一个变量，使用 Spearman 相关系数来判断各位点上的基因对于 1389163_at 上基因影响大小。其计算公式为：

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

相关系数 r 的取值范围为 $[-1, 1]$。相关系数越接近于 1 或-1，两者关联度越高，相关系数越接近于 0，两者关联度越低。

Spearman 关系数假设检验

Spearman 相关系数的假设检验分为两类，一个是小样本的情况，即样本的数量小于 30 的情况下时，可以直接使用查表的方式进行验证。当我们的相关系数大于等于表中的临界值的时候，可以认为相关系数是有显著性差异的，即有相关性，相关性不为 0。明显本题不属于这种情况。又知在大样本的情况下，我们可以通过构建统计量的方式进行假设检验，查询资料总结了在以下的统计量是符合正态分布的。

其公式表示为：

$$r_s\sqrt{n-1} \sim N(0,1)$$

因此当样本数量大于 30 的时候我们可以用如下的方法构建统计量，计算 p 值：

$$p = r_s\sqrt{n-1}$$

如果 p 值大于 0.05，则不认为存在显著性差异，即变量间没有相关性。如果 p 值小于 0.05 的，我们可以认为变量间存在显著性的差异。

结果

在显著性 0.05 的条件下，设置阈值 0.5，初步筛选出和对 1389163_at 上的基因表达较有影响的基因对应的位点数据 1152 个。

5.3.2 互信息 MI)

互信息（Mutual Information，MI）实际上是更广泛的相对熵的特殊情形，用以度量一个随机变量由于已知另一个随机变量而减少的不确定性。

设两个随机变量 (X,Y) 的联合分布为 $p(x,y)$，边缘分布分别为 $p(x)$ 与 $p(y)$，则互信息 $I(X;Y)$ 是联合分布 $p(x,y)$ 与边缘分布 $p(x)p(y)$ 的相对熵，公式如下：

$$I[x;y] = \int dxdyp(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

$$I[x;y] \approx I[X;Y] = \sum_{X,Y} p(X,Y)\log_2\frac{p(X,Y)}{p(X)p(Y)}$$

在 MI 计算中，可以使用 sklearn 中 mutual_info_regression 来计算各个输入特征和输出值之间的互信息。使用 feature_selection 库的 SelectKBest 类进行选择特征，估计其他基因与 1389163_at 上基因之间的相关性。

5.3.3 最大互信息系数（MIC）

在实验过程中，我们发现互信息直接用于特征选择存在一些问题：它不属于度量方式，也没有办法归一化，在不同数据及上的结果无法做比较，通常变量需要先离散化，而互信息的结果对离散化的方式很敏感。

借鉴蒙特卡洛思想，我们用最大互信息系数（Maximal Information Coefficient，MIC）即最大互信息系数克服了这两个问题。使用 MIC 来衡量两个基因之间的关联程度，线性或非线性关系，相较于互信息而言有更高的准确度。它首先寻找一种最优的离散化方式，然后把互信息取值转换成一种度量方式，取值区间在 $[0,1]$。

结果

在 MIC 计算中，可以使用 minepy 中提供的相应接口，逐一计算其他基因与 1389163_at 上基因之间的最大互信息系数。

利用互信息和最大互信息系数，通过对 feature importance 排序初步筛选出和对 1389163_at 上的基因表达较有影响 1000 个的基因对应的位点数据。

5.3.4 结果的分析与讨论

本文选取整理有关表达数量性状基因座与多基因之间数据相关资料，并结合本问题实际情况，最终采取了两套筛选参数：

(1)Spearman 系数，设置阈值 0.5；互信息和最大互信息系数，取得分前 1500，两种方法结果取交集，筛选出了对 1389163_at 上基因影响最高 510 个基因位点。具体见附件中 q1_511_p.xlsx 与 q1_511_s.xlsx，分别为原始基因数据和标准化的基因数据。

(2)Spearman 系数，设置阈值 0.55；互信息和最大互信息系数，取得分前 1000，两种方法结果取交集，筛选出了对 1389163_at 上基因影响最高 224 个基因位点。具体见附件中 q1_224_p.xlsx 与 q1_224_s.xlsx，分别为原始基因数据和标准化的基因数据。

利用 Spearman 求出的相关系数绘制多个基因位点与 1389163_at 基因相关系数，用热力图中不同方块颜色对应代表相关系数的大小，由此直观判断出变量之间相关性的大小，14 代表 1389163_at 基因，如图 4 框选出来的部分可直观表示各个基因位点与红眼基因相关系数关系。

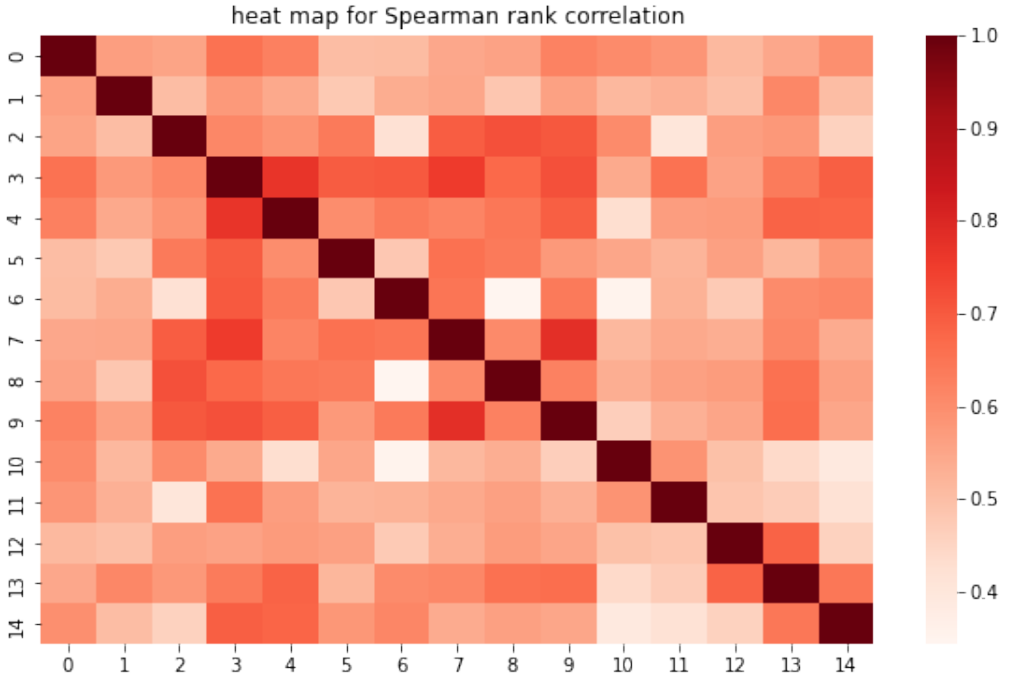

图 4　Spearman 相关系数热力图表示

两种方法求得各个基因位点与红眼基因之间的相关系数，部分截取如表 4 所示。

表 4　相关系数表的部分截选

标号	位点	Spearman	MI	MIC
0	1382223_at	0.565790616	0.260537247	0.382228251
1	1368136_at	0.555408355	0.307119327	0.407314038
2	1393231_at	0.653913536	0.317365224	0.437164662
3	1376180_at	0.626331096	0.25503561	0.491931524
4	1374692_at	0.503082813	0.186519964	0.365130992
5	1391126_at	0.506962555	0.248211342	0.373367024
6	1387018_at	0.548304314	0.280006037	0.390376107
7	1398795_at	0.557354563	0.231617156	0.415403154
8	1372921_at	0.622739072	0.225009148	0.408987962
9	1377836_at	0.604391483	0.276795197	0.568739956
10	1390653_at	0.583792437	0.185428467	0.406140826
...	...	...	...	...

5.4 典型相关分析

通常情况下，为了研究两组变量

$$[x_1, x_2, \cdots, x_p], \quad [y_1, y_2, \cdots, y_q]$$

的相关关系，可以用比较原始的方法，分别计算两组变量之间的全部相关系数，一共有 pq 个简单相关系数，即 5.3 中采用的相关系数筛选的方法。但这样即繁琐又不能有效抓住问题的本质。如果能够采用类似于主成分的思想，在两组变量中，分别选取若干有代表性的变量组成有代表性的综合指标，通过研究这两组综合指标之间的相关关系，

来代替这两组变量间的相关关系，这些综合指标称为典型变量。。首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，

$$\begin{cases} u_1 = \alpha_{11}x_1 + \alpha_{21}x_2 + \cdots + \alpha_{p_1}x_p, \\ v_1 = \beta_{11}y_1 + \beta_{21}y_2 + \cdots + \beta_{q1}y_q. \end{cases}$$

然后再在每组变量中找出第二对线性组合, 使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性。

$$\begin{cases} u_2 = \alpha_{12}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{p_2}x_p, \\ v_2 = \beta_{12}y_1 + \beta_{22}y_2 + \cdots + \beta_{q2}y_q. \end{cases}$$

u_2 与 u_1、v_2 与 v_1 不相关，但 u_2 和 v_2 相关。如此继续下去，直至进行到 r 步，两组变量的相关性被提取完为止，可以得到 r 组变量，这里 $r \leq \min(p, q)$。

5.4.1 数学描述

给定两个带有限矩的随机变量的列向量 $X = (x_1, \ldots, x_n)'$ 和 $Y = (y_1, \ldots, y_m)'$，我们可以定义互协方差矩阵 $\Sigma_{XY} = \operatorname{cov}(X, Y)$ 为 $n \times m$ 的矩阵，其中 (i, j) 是协方差 $\operatorname{cov}(x_i, y_j)$。实际上，我们可以基于 X 和 Y 的采样数据来估计协方差矩阵。

典型相关分析求出向量 a 和 b 使得随机变量 $a'X$ 和 $b'Y$ 的相关性 $\rho = \operatorname{corr}(a'X, b'Y)$ 最大。随机变量 $U = a'X$ 和 $V = b'Y$ 是第一对典型变量。然后寻求一个依然最大化相关但与第一对典型变量不相关的向量；这样就得到了第二对典型变量。这个步骤会进 $\min\{m, n\}$ 次。

5.4.2 典型相关系数的检验

在实际应用中，总体的协方差矩阵常常是未知的，类似于其他的统计分析方法，需要从总体中抽出一个样本，根据样本对总体的协方差或相关系数矩阵进行估计，然后利用估计得到的协方差或相关系数矩阵进行分析。由于估计中抽样误差的存在，所以估计以后还需要进行有关的假设检验。

(1) 计算样本的协方差阵

(2) 整体检验

$H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_s = 0, (s = \min(p, q))$

$H_1 : \lambda_i (i = 1, 2, \cdots, s)$ 中至少有一非零。记

$$\Lambda_1 = \frac{|\widehat{\boldsymbol{\Sigma}}|}{\left|\widehat{\boldsymbol{\Sigma}}_{XX}\right| \left|\widehat{\boldsymbol{\Sigma}}_{YY}\right|}$$

经计算得

$$\Lambda_1 = \left| \boldsymbol{I_p} - \widehat{\boldsymbol{\Sigma}}_{XX}^{-1} \widehat{\boldsymbol{\Sigma}}_{XY} \widehat{\boldsymbol{\Sigma}}_{YY}^{-1} \widehat{\boldsymbol{\Sigma}}_{YX} \right| = \prod_{i=1}^{s} \left(1 - \lambda_i^2\right)$$

在原假设为真的情况下，检验的统计量

$$Q_1 = -\left[n - 1 - \frac{1}{2}(p + q + 1)\right] \ln \Lambda_1$$

近似服从自由度为 pq 的 χ^2 分布。在给定的显著水平 α 下，如果 $Q_1 \geq \chi_\alpha^2(pq)$，则拒绝原假设, 认为至少第一对典型变量之间的相关性显著。

(3) 部分总体典型相关系数为零的检验

$H_0 : \lambda_2 = \lambda_3 = \cdots = \lambda_s = 0, H_1 : \lambda_2, \lambda_3, \cdots, \lambda_s$ 至少有一非零。

若原假设 H_0 被接受，则认为只有第一对典型变量是有用的；若原假设 H_0 被拒绝，则认为第二对典型变量也是有用的，并进一步检验假设 $H_0 : \lambda_3 = \lambda_4 = \cdots = \lambda_s = 0, \quad H_1 : \lambda_3, \lambda_4, \cdots, \lambda_s$ 至少有一非零。如此进行下去, 直至对某个 k $H_0 : \lambda_k = \lambda_{k+1} = \cdots = \lambda_s = 0, H_1 : \lambda_k, \lambda_{k+1}, \cdots, \lambda_s$ 至少有一非零。

$$\Lambda_k = \prod_{i=k}^{s} \left(1 - \lambda_i^2\right)$$

在原假设为真的情况下，检验的统计量

$$Q = -\left[n - k - \frac{1}{2}(p + q + 1)\right] \ln \Lambda_k$$

近似服从自由度为 $(p - k + 1)(q - k + 1)$ 的 χ^2 分布。在给定的显著水平 α 下，如果 $Q \geq \chi_\alpha^2((p - k + 1)(q - k + 1))$，则拒绝原假设，认为至少第 k 对典型变量之间的相关性显著。

5.4.3 基因的典型相关性分析

我们以 5.3 中已筛选出的 510 基因作为对象，分析其与 1389163_at 上的基因的典型相关性。因此 x 组中包含 500 个基因的数据（120 条样本），y 组中有包含 1389163_at 在内的 10 个基因数据（120 条样本）。y 组基因涉及了从相关系数角度最强的 9 个基因以及 1389163_at 本身，以它们本身的表达量数据作为 10 个关键指标。我们希望通过典型相关性分析，对 5.3 中筛选的基因结果的合理性进行一定程度的检验，并尝试进一步剔除掉 x 组中相关性较弱的若干个基因。

I. 典型相关系数及其检验

首先，我们希望借助典型相关系数，验证 5.3 中相关的筛选方法得到的 9 个基因与 1389163_at 基因本身相关性显著。这是我们分析后续问题二、三的重要基础。

序号	1	2	3	4	5	6	7	8	9	10
典型相关系数	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.97

图 5　典型相关系数

但要确定典型变量相关性的显著程度，尚需进行相关系数的 χ^2 统计量检验，即比较统计量 χ^2 计算值与临界值的大小。其结果如图 6 所示：

序号	自由度	卡方计算值	卡方临界值
1	1190	Inf	0
2	1062	Inf	0
3	936	Inf	0
4	812	Inf	0
5	690	Inf	0
6	570	9.61e+03	0
7	452	7.67e+03	0
8	336	5.72e+03	1.47e-97
9	222	3.77e+03	5.95e-72
10	110	1.89e+03	3.43e-26

图 6　典型相关性的显著检验

所有的 10 个典型相关系数均较高，表明相应典型变量之间密切相关，且都通过了 χ^2 统计量检验，表明相应典型变量之间相关关系显著。上述这验证了我们在 y 组中所设置的关键基因是合理的。

II. 典型相关模型

我们采用标准化的典型系数，给出了典型相关模型的系数，具体可见附件中 u_coef.mat 与 v_coef.mat。与 I 中结论类似的，y 组中的关键因素涵盖了 y 组中所有的 10 个基因；而 x 组中，则被筛选出来了若干系数极小的变量（即对应系数矩阵中的某个全零行），由此可以从 x 组的 500 个基因中，按重要程度筛选出主要因素共 182 个，即 182 个基因。

经后续检验，182 个基因中，与 5.3.4 中所述的第二种筛选策略得到 224 个基因，重合度达到了 96.7%。

III. 典型冗余分析与解释能力

典型冗余分析用来表示各典型变量对原始变量组整体的变差解释程度，分为组内变差解释和组间变差解释，典型冗余分析的结果见图 7。

被本组的典型变量解释			被对方y组典型变量解释		
	比例	累计比例		比例	累计比例
u1	0.037	0.037	v1	0.068	0.068
u2	0.006	0.043	v2	0.027	0.095
u3	0.018	0.061	v3	0.038	0.132
u4	0.152	0.212	v4	0.144	0.276
u5	0.080	0.293	v5	0.080	0.356
u6	0.013	0.306	v6	0.062	0.418
u7	0.012	0.318	v7	0.021	0.439
u8	0.018	0.336	v8	0.039	0.479
u9	0.006	0.342	v9	0.022	0.501
u10	0.399	0.741	v10	0.499	1.000

图 7　被典型变量解释的 x 组原始变量的方差

六、 问题二模型建立与求解

问题一中，我们筛选出了对 1389163_at 上的基因表达有较大影响的基因座数据，本题在此基础上对数据进行分类处理量化，根据对红眼基因影响力大小对基因位点进行回归模型的建立。基于 XGBoost 在分类回归、特征选择问题上的良好表现，本问将采用 XGBoost 对基因位点进行排序。

6.1 基于聚类分析方法预估相似基因的分布

R 型聚类（R-type Cluster），聚类分析方法的一种。根据不同变量之间相关程度高低进行分类。研究中，若变量较多且相关较强时，可以使用 R 型聚类法把变量聚为几个大类，同一类变量之间有较强相关性，不同类变量之间相关程度低，并可以从同类变量中找出一典型性变量作为代表，最终减少变量个数达到降维目的。

本题由于 1389163_at 上的基因跟其他基因位点上的基因关系是非线性的，则可以从其余的位点中筛选高相似度变量，了解个别变量之间的关系的亲疏程度，以达到缩减变量个数。我们通过计算变量（基因）间的夹角余弦，针对所有特征位点，利用最短距离法对变量进行聚类。

6.1.1 变量相似度度量

利用两变量 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_k$ 的夹角余弦 r_{jk} 来定义它们的相似性度量，有

$$r_{jk}=\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\left(\sum_{i=1}^{n}x_{ij}^{2}\sum_{i=1}^{n}x_{ik}^{2}\right)^{\frac{1}{2}}}$$

变量的相似度量均应具有以下两个性质:

(1) $|r_{jk}|\leqslant 1$，对于一切 j,k;

(2) $r_{jk}=r_{kj}$，对于一切 j,k 。

$|r_{jk}|$ 越接近 1，$\boldsymbol{x}_j$ 与 $\boldsymbol{x}_k$ 越相关或越相似。$|r_{jk}|$ 越接近 0，$\boldsymbol{x}_j$ 与 $\boldsymbol{x}_k$ 的相似性越弱。

6.1.2 变量聚类法

针对本次基因选择的变量聚类问题，我们采用最长距离法。两类变量的距离定义为

$$R\left(G_1,G_2\right)=\max_{\substack{x_j\in C_1\\ x_k\in G_2}}\left\{d_{jk}\right\}$$

式中 $d_{jk}=1-|r_{jk}|$ 或 $d_{jk}^2=1-r_{jk}^2$，这时，$R\left(G_1,G_2\right)$ 与两类中相似性最小的两变量间的相似性度量值有关。

6.1.3 聚类结果

通过上述变量聚类方法，我们通过调整聚类参数，以 1389163_at 作为一类簇中心，并结合层次聚类的思想进行迭代，最终将 223 个变量（基因）划分为 4 类。

为了将聚类结果进行可视化，我们采用 t-SNE 方法对各基因数据降至三维。可以发现，1389163_at 周围存在若干同类点，相似基因的分布整体呈现“多层”趋势，这为我们后续对相似基因进行量化排序提供了一定分析基础。

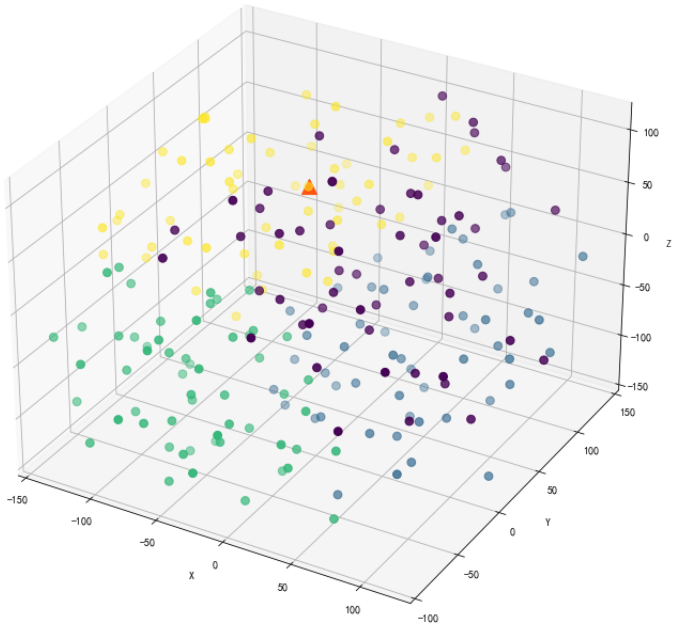

图 8　聚类结果可视化

6.2 基于 XGBoost 建立回归模型的对数据进行量化排序

XGBoost 是一种 Boosting 型的树集成模型，在梯度提升决策树 GBDT 基础上扩展，能够进行多线程并行计算，通过迭代生产新树，即可将多个分类性能较低的弱学习器组合为一个准确率较高的强学习器。XGBoost 采用对字段抽样，将正则项引入损失函数中，从而防止模型过拟合，并降低模型计算量。

XGBoost 的目标函数:

$$Obj(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)} + f_t(x_i)) + \Omega(f_t) + constant$$

其中:

第一项中的 l 即为损失函数（比如平方损失函数）。

第二项是正则项（包括 L1 正则、L2 正则），用来定义复杂度，其作用是限制树的叶子节点的个数，防止树变得过于庞大，该项值越小复杂度越低，泛化能力越强。其表达式为: $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2$

第三项为常数项最终的目标是要使得树群的预测值 $y_i{}'$ 尽量接近真实值 y_i，而且有尽量大的泛化能力。

6.2.1 XGBoost 的算法流程

（1）不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数 f(x)，去拟合上次预测的残差。

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

$$whereF = \{f(x) = \omega_{q(x)}\}q : R^m \to T, \omega \in R^T$$

注: $\omega_{q(x)}$ 为叶子节点 q 的分数，对应了所有 K 棵回归树的集合 $f(x)$ 为其中一棵回归树。

（2）当训练完成得到 k 棵树，要预测一个样本的分数时，其实就是根据这个样本的特征，在每棵树中落到对应的一个叶子节点，每个叶子节点就对应一个分数。

（3）最后只需要将每棵树对应的分数加起来就是该样本的预测值。

6.2.2 XGBoost 的优势

相较于传统的 GBDT 算法，XGBoost 的优势如下:

（1）GBDT 将目标函数泰勒展开到一阶，而 XGBoost 将目标函数泰勒展开到了二阶。保留了更多有关目标函数的信息，对提升效果有帮助;

（2）GBDT 是给新的基模型寻找新的拟合标签（加法模型的负梯度），而 XGBoost 是给新的基模型寻找新的目标函数（目标函数关于新的基模型的二阶泰勒展开）;

（3）XGBoost 加入了和叶子权重的 L2 正则化项，因而有利于模型获得更低的方差;

（4）XGBoost 增加了自动处理缺失值特征的策略。通过把带缺失值样本分别划分到左子树或者右子树，比较两种方案下目标函数的优劣，从而自动对有缺失值的样本进行划分，无需对缺失特征进行填充预处理。

6.3 结果分析与讨论

（1）根据问题一中的筛选，保留下高相关性基因对应列数据，重组成 120 行 510 列的矩阵;

（2）使用 XGBoost 中封装好的回归器（XGBRegressor）初始化模型;

（3）输入训练数据，让回归器自动进行进行拟合，并通过 feature_importances 返回变量的重要性排序;

（4）采用网格搜索调参，对 XGBRegressor 的参数中最大树高度、学习率、最小叶子权重（即预测得分）、树的数量等进行动态调整，定位最优参数;

（5）重新输入训练数据进行拟合，并通过 feature_importances 返回变量的最终重要性排序。

XGBoost 调参结果如下表所示

表 5　XGBoost 调参结果

参数名	对应值
learning_rate	0.08
max_depth	5
n_estimators	100
base_score	0.5
booster	gbtree

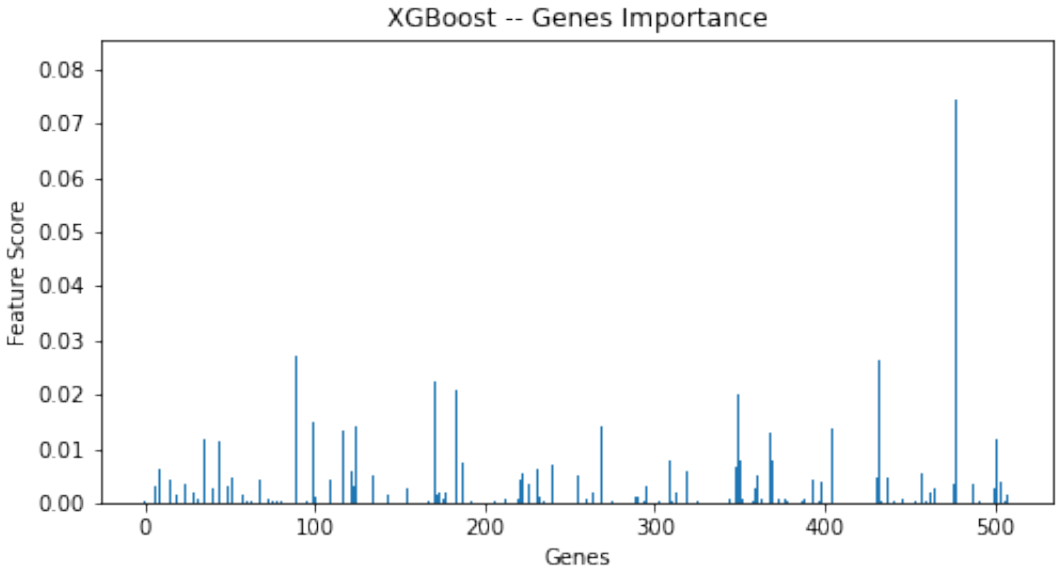

图 9　XGBoost 回归模型对 510 个基因的重要性评估

图 9 是 XGBRegressor 的回归结果中各基因的 feature _importances，横坐标代表筛选得到的 510 个基因。可以发现，在众多基因中，存在 5-15 个基因对 1389163_at 上基因的表达有着显著影响，而绝大多数对 1389163_at 上基因的影响可以忽略不计。

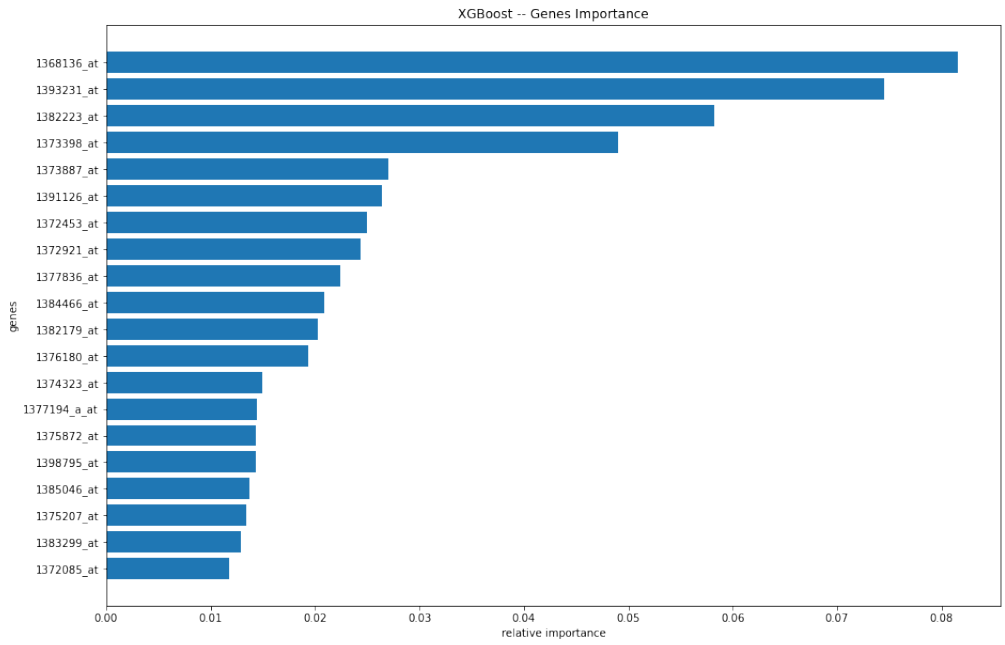

图 10　XGBoost 回归模型拟合后，相对重要性前 20 的基因

同时根据各个基因的重要程度进行排序，截取前 20 个基因，即得到图 10。

七、 问题三模型建立与求解

问题三分为两个子问题：子问题一是预估若对 1389163_at 上的基因建立合适的模型所需要最少的自变量个数有多少，子问题二是给出一个决定自变量个数的合理准则。

在问题一和问题二中我们已经完成了 1389163_at 上的基因表达有较大影响的基因座数据的筛选，并根据他们的影响特征大小进行量化排序，筛选得到 relative importance 高于两倍平均值的基因共 25 个，1368136_at，1393231_at，1382223_at，1373398_at，1373887_at，1391126_at，1372453_at，1372921_at，1377836_at，1384466_at，1382179_at，1376180_at，1374323_at，1377194_a_at，1375872_at，1398795_at，1385046_at，1375207_at，1383299_at，1372085_at，1395581_at，1372463_at，1372163_at，1384029_at，1390539_at。

在已有基础上，本问题中我们将以主成分分析法为基础，采用核主成分分析（KPCA）对数据进行降维处理，通过结果形成的聚簇分析预估所需要最少的自变量个数。接着，我们使用递归特征消除（RFE）优化 XGBoost 模型，排除拟合度过高的问题，并结合交叉验证得分情况对基因重要性进行更加精准的排序与估计，进而确定所需要最少的自变量个数。

7.1 基于 KPCA 对数据进行降维预估所需要最少的自变量个数

降维顾名思义就是把数据或特征的维数降低，一般分为线性降维和非线性降维，本题由于 1389163_at 上的基因跟其他基因位点上的基因关系是非线性的。针对非线性降维，主要使用的方法有 t-SNE 和 KPCA，而基于 KPCA 算法的降维，当核函数取高斯核时，随着 γ 的增大，使用 KPCA 降维至低维的数据仍可充分地保留数据的原始特性，且针对不同的目标维数均可保证结果的稳定性，因此我们选用 KPCA 数据进行降维从而进行所需要最少的自变量个数的预估。

KPCA 是对传统 PCA 算法的非线性扩展。PCA 主要用于解决线性问题，而 KPCA 能够挖掘到数据集中蕴含的非线性信息。在 PCA 的基础上，KPCA 把原始的非线性的数据映射到高维空间变成线性的，然后用 PCA 来处理映射后的高维数据。

由于经过两轮筛选的基因集合仅包含 25 个基因，且每个基因均与 1389163_at 高度相关，因此我们将 1389163_at 加入原有的基因集合，使得最终的降维结果可以在更高程度上表征对 1389163_at 模型的最少自变量个数。

7.1.1 基础：PCA 算法

主成分分析，实质上是一种线性降维方法，主要目的是希望用较少的变量去解释原来资料中的大部分变异，将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量。通常是选出比原始变量个数少，能解释大部分资料中的变异的几个新变量，即所谓主成分，并用以解释资料的综合性指标。

7.1.2 KPCA 算法原理

采用的核函数为径向基核函数 (RBF):

（1）我们将在高维空间中把数据投影到由 W 确定的超平面上，即 PCA 欲求解

$$(\sum_{i=1}^{m} z_i {z_i}^T)W = \lambda W$$

（2）假定是由原始属性空间中的祥本点 Z_i，通过映射 ϕ，即 $Z_i = \phi(x_i), i = 1, ..m.$ 若能被 ϕ 表达出来，则通过它把样本映射到高维度特征空间，再在特征空同中实施 PCA 即可，上面两式变为

$$(\sum_{i=1}^{m} \phi(x_i)\phi(x_i)^T)W = \lambda W$$

（3）引入核函数

$$z_j = {w_j}^T \phi(x) = \sum_{i=1}^{m} {\alpha_j}^j k(x^i, x^j)$$

针对本题，我们采用高斯核函数:

$$\mathrm{k}(x^i, x^j) = \exp(-\frac{||x^i - x^j||^2}{2\sigma^2})$$

7.1.3 基于 Adjusted R-square 进行筛选评价

R-Square 作为回归模型中常用来衡量回归方程与真实样本的拟合程度的标准，有其自身的局限性，只要增加了更多的变量，无论增加的变量是否和输出变量存在关系，R-square 均呈现非减状态，即检验并不能发现特征变量对于模型本身的影响程度。所以我们选择 Adjusted R-square 进行修正，它会增加一个惩罚项，基于样本数量的增加，给予一个正则项。对加入模型但不会显著改善模型效果的变量，给予数值调整，从而抵消样本数量对 R-square 的增益，寻找最优特征数量。其数学表达式为:

$$R^2\_adjusted = 1 - \frac{(1 - R^2)(n - 1)}{n - b - 1}$$

上式中，b 为变量个数，n 为样本个数。

惩罚项用以评估增加的基因是否能 “有效” 降低回归误差，如果新增基因并不能为 1389163_at 模型建立提供更多信息，则 R-square 和 Adjusted R-square 之间的差距会越来越大，Adjusted R-square 会下降。

7.1.4 KPCA 结果与分析

最终，我们利用 KPCA 实现了从高维到低维的映射，由结果可以看出，随着变量选取的个数不断增大，用以评判误差的 Adjusted R-square 会不断减小。在自变量数取得 11 时，Adjusted R-square 到达临界值。由此我们预估出所需要最少的自变量个数在 11-15。

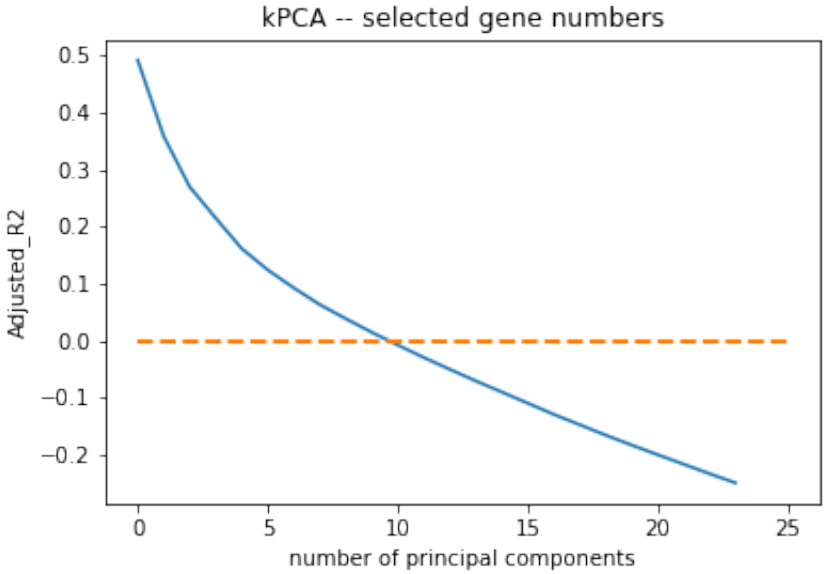

图 11　Adjusted_R2 随变量个数的变化

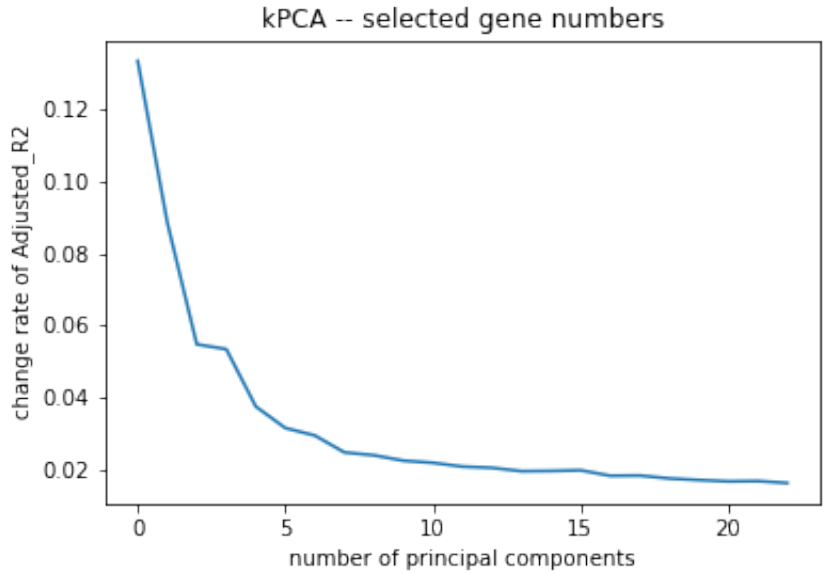

图 12　Adjusted_R2 变化率随变量个数的变化

7.2 基于递归特征消除（RFE）选择所需要最少的自变量个数

递归特征消除（Recursive feature elimination，RFE），是一种典型的包裹式（wrapper）特征选择的方法。RFE 的主要思想是反复构建模型，然后选出最好的（或者最差的）特征（根据系数来选），把选出来的特征放到一边，然后在剩余的特征上重复这个过程，直到遍历了所有的特征。在这个过程中被消除的次序就是特征的排序。RFE 的稳定性很大程度上取决于迭代时底层所选择的模型。

7.2.1 算法步骤

RFE 通过之前训练好 XGBoost 的模型及赋予特征的重要性，选出最差的特征，在剩余特征组成的数据上进行交叉验证分析并在剩余的特征上重复上述过程，直到递归遍历了所有的特征，来获取特征的重要性排序及特征的不同组合的交叉验证得分情况；如果减少特征会导致性能损失，那么递归将会停止，RFE 不会再去除任何特征。

经过上文的特征筛选，我们对 25 个基因对应的表达量信息重组成矩阵，采用 step 为 1、cv 为 5，进行递归特征消除。

RFECV 参数如下表所示

表 6　RFECV 参数

参数名	对应值
estimator	XGBmodel
step	1
cv	5
min_features_to_select	5

7.2.2 结果与分析

最终，结合交叉验证（CV）的递归特征消除得到上述 25 个基因的得分排名为：

[1 1 13 12 6 1 3 1 14 11 10 7 2 9 1 8 5 1 4 16 1 15 1 1 1]

基因的特征得分排名越低（1 最好），表示该基因对 1389163_at 基因的影响越强。

因此，该过程中成功消除表现较差的基因 15 个，最终取得的特征得分排名为 1 的基因，共 10 个，这与 7.1 中采用 KPCA 降维得到的结果基本一致。

根据这 10 个基因的相对特征得分进行排序，结果为图 13。

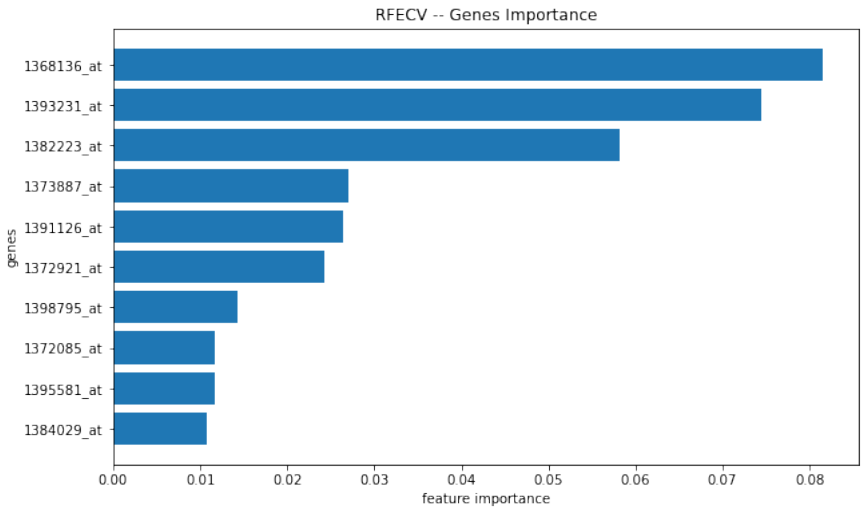

图 13　相对特征得分前十名的基因

交叉验证得分在变量选择过程中的变化

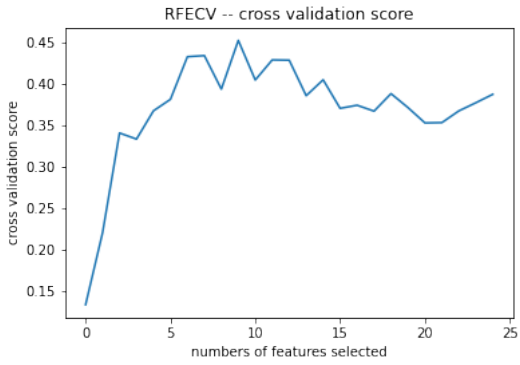

图 14　交叉验证得分

八、 模型评价、改进与推广

8.1 模型综合评价

本文通过数理统计方法和机器学习方法，依次完成了从数据表现对影响 1389163_at 上基因表达起共同作用基因位点的筛选，量化不同基因对目标基因表达影响的权重，以及通过机器学习算法完成对特征数量的筛选与具体特征的选择。

首先，本文在数据检查中未发现数据缺失与奇异值，因此在原始数据的基础上通过 Kolmogorov-Smirnov 正态检验验证样本分布状态，并对所有基因位点分别进行了标准化处理。同时，将不同的基因位点进行离散化处理，将位点名转换为离散变量，方便后续算法处理。在此基础上，通过 Spearman 系数、互信息与最大互信息系数来检验样本之间的相关性是否显著，绘制热力图以展示变量间相互关系，直观表现相关程度。此外，我们以典型相关分析方法作为补充，验证了上述筛选过程的有效性与合理性。

其后，我们采用 R 型聚类方法，以夹角余弦作为基因相似度度量，以最长距离作为判据进行变量聚类，初步分析了相似基因在高维空间的分布情况。在此基础上，我们基于 XGBoost 对数据建立量化回归模型，利用该强学习器将问题一中初步筛选出的基因型进行回归，同时引入正则项防止过拟合，最终得到在问题一的基础上得到特征影响权值的分数，达到量化排序的目的。

最后，我们通过核主成分分析法 (KPCA) 对数据进行降维分析，通过分析非线性组合成的类簇来预估最少的特征个数，综合考虑损失最少和特征数最少，预估出特征个数 10-15 个，并利用递归特征消除和交叉验证递归遍历所有特征，获得一致结果。重新构建 XGBoost 回归模型，进行 k 折交叉验证，最终预测回归误差小于 5%。

因此，模型建立所依赖的部分重要基因如表 7 所示：

表 7　建立模型的过程中重要性最高的基因

序号	基因位号	Feature Importance
0	1368136_at	0.08158867
1	1393231_at	0.07450318
2	1382223_at	0.058222797
3	1373887_at	0.027062949
4	1391126_at	0.026390161
5	1372921_at	0.02432973
6	1377194_a_at	0.014441502
7	1398795_at	0.01432289
8	1372085_at	0.011720907
9	1395581_at	0.011698255
10	1384029_at	0.010822368

8.2 基于相关系数分析研究各位点对应基因对红眼病基因的影响

8.2.1 模型优点

（1）本模型通过查阅文献并结合实际情况，对老鼠基因分布与 1389163_at 上基因与各位点对应的基因的关系进行了一定的探索，发现老鼠的各个基因绝基本服从正态分布，且 1389163_at 上基因与各位点对应的基因是非线性关系。

（2）在使用模型之前对数据进行标准化处理，避免了因为量纲差异引起的评分误差。

（3）综合使用 Spearman 相关系数、互信息、最大互信息系数做相关性分析，结果更具说服力。以典型相关分析作为补充，验证了上述筛选过程的合理性与有效性。

（4）在分析各个位点对应的基因对红眼基因表达影响的时候对数据进行量化处理，综合考虑各方面因素，合理选取阈值，完成对数据的初筛。

8.2.2 模型的缺点

（1）在相关性的分析中，计算成对基因间的相关系数，对计算机资源耗费较高。从基因间的非线性关系出发，可能忽略了的其他因素对基因相关性判定的影响。

（2）相关性分析存在误判的因素，有可能导致少数重要基因在此过程中被筛除，这需要我们在阈值设置时充分谨慎——既能起到有效的筛选作用，又不能对模型的重要特征产生误判。

8.2.3 模型的改进

在非线形关系相关性分析方面，合理选择增加相关系数计算方法，弥补 Spearman 与 MIC 的不足，并量化分析数据，使得评判更加全面；对基因正态性做更细致的讨论，增强其合理性。

8.2.4 模型的推广

本模型结合 Spearman、互信息和最大互信息系数做相关性分析，十分适合应用于研究基因组间非线形关系的相关性，量化其影响程度。同时，不仅仅局限于红眼基因组研究，本模型可以扩展至人类基因组乃至生物医学研究以外、政治经济生活中各种角色的评估。

8.3 基于 XGBoost 对各位点数据对红眼基因表达量影响重要性量化模型

8.3.1 模型的优点

（1）在传统 GBDT 的基础上，我们使用 XGBoost 完成对树形模型的构建，对于特征的值有缺失的样本，XGBoost 还可以自动学习出它的分裂方向，对数据特征的挖掘更加深入，扩展性更强。

（2）代价函数中引入正则化项，控制了模型的复杂度，正则化项包含全部叶子节点的个数、每个叶子节点输出得分、L2 的平方和。从贝叶斯方差角度考虑，正则项降低了模型的方差，防止模型过拟合。

（3）XGBoost 支持并行。迭代之前，先进行预排序，存为 block 结构，每次迭代，重复使用该结构，降低了模型的计算；block 结构也为模型提供了并行可能，在进行结点的分裂时，计算每个特征的增益，选增益最大的特征进行下一步分裂，那么各个特征的增益可以开多线程进行。

8.3.2 模型的缺点

（1）对噪声敏感，模型的鲁棒性依赖于数据预处理的效果。

（2）RFE 递归结果极其依赖于评估器的选择及其参数，有一定的不稳定性。

8.3.3 模型的改进

由于本题是针对研究多对基因对单个基因的影响而建立的 XGBoost 模型，交叉验证并不能从实质上解决模型有可能过拟合的问题，后续可以考虑寻找更适合多对一数据研究的优化模型方法。

8.3.4 模型的推广

XGBoost 模型不仅适用于非线性关系，也同样适用于线性关系之间的研究，作为优秀的分析算法模型，XGBoost 不仅可以用于基因的关系量化排序探索，也可以很好的运用于预测性维护方面，增强传统制造业安全系数，内的供货量预测上，此外，产量预测、客户需求分析和服务类型识别等也是广泛存在的场景。

8.4 基于核主成分分析与递归特征消除对自变量选择的模型

8.4.1 模型的优点

（1）本模型基于主成分分析的思想，具备有严格的数学理论作基础。由于传统的主成分分析（PCA）方法不利于求解非线性问题，我们在实践中采用核主成分分析（KPCA）的方法。

（2）本模型把原始的非线性的多个基因位点对 1389163_at 上基因的影响关系，通过 KPCA 方法，尝试向低维空间进行非线性映射，进而以主成分研究来探索基因集合的合理自变量个数。

（3）本模型创新性地采取 Adjusted R-square 来完成对 KPCA 预测的评价，这一修正后的误差指标增加了惩罚项，使得我们在自变量筛选的过程中趋向于放弃对模型预测效用较小的基因。

（4）本模型集合递归特征消除与交叉验证得分，使得模型在减少可能的自变量数目的过程中，不会损失预测性能，并在最大程度对可能的自变量个数进行压缩。

（5）本模型在确定自变量个数后，通过 k 折交叉验证，确保回归预测的误差在合理范围内，从而验证了模型的有效性。

8.4.2 模型的改进

本文没有深入分析各个位点彼此之间的多重影响关系，由于计算位点之间的相关性的运算量过于庞大，受限于计算机的计算性能，需要考虑更加高效的方法来实现，此项工作或许可以在日后的学习中继续推进并完善。

8.4.3 模型的推广

本模型基于相关系数筛选、典型相关分析、R 型聚类分析等方法，并通过 KPCA 对数据进行尝试化降维，使用调整后的 Adjusted R-square 计算对非显着惩罚给出的变量，最后再估计出变量值的基础上找到建立红眼基因模型所需要的最少自变量个数，从而完成对多基因对红眼性状影响的初步探索。这种数理统计方法与集成学习相结合的研究方法，可以为生物学领域基因相关性的研究提供更多支持，帮助研究人员缩小相关基因的范围，指明探索方向，节约研究成本。而 XGBoost、RFE 也可以适用于高维数据的分类、回归分析与特征选择，结合 RNN 等深度学习方法，我们可以去探索多顺反子表达载体的构建策略，研究外源基因，针对疾病发生发展的各个环节设计联合基因治疗方案等。不仅限于生物医学，我们也可以将上述建模思路应用在经济、政治等各诸多方面。

参考文献

[1] Akossou A Y J, Palm R. Impact of data structure on the estimators R-square and adjusted R-square in linear regression[J]. Int. J. Math. Comput, 2013, 20(3): 84-93.

[2] Ryu S E, Shin D H, Chung K. Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization[J]. IEEE Access, 2020, 8: 177708-177720.

[3] Granitto P M , Furlanello C , Biasioli F , et al. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products[J]. Chemometrics Intelligent Laboratory Systems, 2006, 83(2):83-90.

[4] Dy J G , Brodley C E . Feature Selection for Unsupervised Learning[J]. Journal of Machine Learning Research, 2004.