中期报告——观菌识人

尹昊宇 苗壮 张运兴

Group：nk_22_B_406_4

最后更新时间：*2019* 年 *2* 月 *10* 日

摘要

本文首先对原始数据集进行了初步清晰和可视化的工作。三个原始数据集总共包含了 2054 个观测。筛选完推荐样本后 (重庆泉州样本)，一共包含了 1638 个观测。其中背景信息中的所有 21 个特征均可以作为响应变量进行预测，本文目前仅在年龄作为响应变量的情况下[1]进行研究。genus 数据集总共包含 287 个特征，otu 数据集总共包含 2101 个特征。本文利用最大信息系数和距离相关系数这两个统计量对变量进行初步筛选，设定阈值后，将两个数据集的变量分别压缩至 18 和 70 个变量。再利用 xgboost 的基准模型，利用递归的方式将两个数据集的变量继续压缩至 7 和 32 个。虽然特征工程在一定程度上会丢失一定的信息，但整体而言这项工作降低了数据集的噪声，避免了过拟合，也使预测模型的运算速度和效率有了很大的提升。

关键字：特种工程, 最大信息系数, 距离相关系数, 递归 xgboost

[1]对其余响应变量的研究如 BMI 等将会在最终报告中讨论

目录

1 数据预处理

1.1 数据清洗

依据数据说明，将第一个背景数据文件中泉州和重庆的样本筛选出来进行研究，总共得到 1638 个观测，以变量 SampleID 为键，将背景数据与 genus 数据集和 otu 数据集进行横向合并，继而研究缺失值。

统计发现数据集仅 Waistline、Height 以及 Weight 变量分别有 15，6 以及 6 个缺失值，其余特征均没有缺失，含有缺失值的观测数量仅占样本总数的 1.6%。可以看到数据集的质量是相当高的。

对于 Waistline 中的 15 个缺失值，Waistline 的数据分布如图1所示：

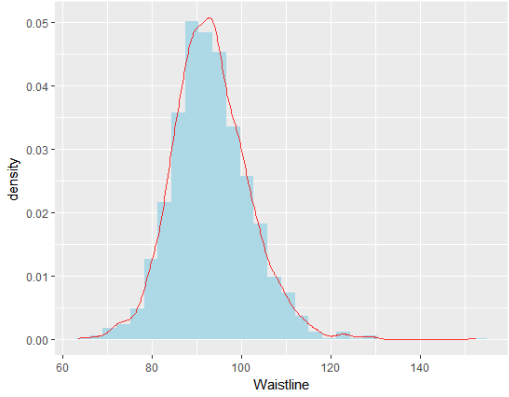

图 1: 腰围数据分布

可以发现该腰围数据的分布比较集中，如果从简化计算的角度考虑，利用数据的均值作为缺失值的填补是一个合理的选择。但由于数据集的其余特征几乎没有缺失，因此本文采用了多重插补的方法，利用 R 语言中的 mice 包对该特征的缺失值进行填补。

对于特征 Height 以及 Weight 包含的缺失值，本文不做任何处理。尽管这两个特征也可以作为因变量来进行研究，但是特征 BMI 并不含有缺失值，可以知道他们之间有如下关系：

$$BMI = \frac{Weight}{Height^2} \tag{1}$$

可以发现 Weight 和 BMI 之间是正比的关系，而 Height 和 BMI 之间是反比的关系，若将 BMI 作为因变量，并且模型能够准确发现肠道细菌与 BMI 之间的关系，那么我们有信心模型在 Weight 和 Height 上也可以有同样的表现。因此 Weight 和 Height 在本案例中不是必须的变量，故不对其缺失值进行处理。

1.2 可视化

要研究背景信息与肠道细菌之间的联系，首先需要对背景信息的数据分布有所了解。图2展示了 4 个较为典型的背景变量的分布情况：

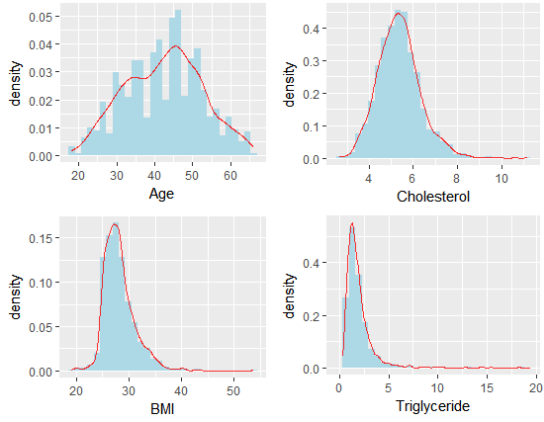

图 2: 连续背景变量的分布

可以看到列举的 4 个变量 (年龄，胆固醇，BMI，甘油三酯) 均为连续型数据，可以看到年龄数据的分布较为分散，而甘油三酯的数据密度函数峰度高，数据较为集中。在对其进行预测时，若将均方误差考虑为衡量指标，则对年龄进行预测的均方误差势必会大于对甘油三酯的预测 (由于年龄数据自身有更大的方差)，而这却并不能说明对年龄的预测结果更糟糕。此时将会把原始数据的方差作为系数来修正衡量指标，以便对预测结果可以有更公正的判断。

图3展示了背景信息中两个离散变量的分布情况：

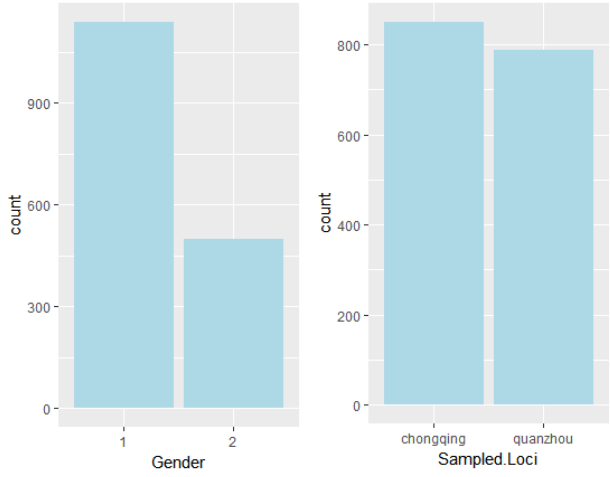

图 3: 离散背景变量的分布

从图3发现来自重庆地区的样本和泉州地区的样本数量相当，若将地域作为因变量，则不需要进行过多的处理。而性别变量，性别 1 是性别 2 数量的一倍之多，存在严重的类别不平衡问题，将性别作为因变量时，应该对性别 1 进行适当的降采样而对性别 2 利用 SMOTE 算法进行适当的过采样 (尽可能防止过拟合问题)。

细菌数据总共包含了两千多个特征，数据是极为稀疏，下一节将通过特征工程对数据进行筛选。

2 特征工程

首先利用最大信息系数和距离相关系数这两个统计量对特征进行初步的筛选，再对得到的特征集合利用递归的 XGBOOST 模型，得到最优的特征子集。

2.1 最大信息系数

最大信息系数对于一般的变量之间关系具有普适效应，不仅可以发现变量之间的线性关系，也可以发现变量之间的非线性关系。

最大信息系数的计算方式完全基于互信息的计算方式。对于两个连续型随机变量，首先将其所在的二维空间使用 m 乘以 n 的网格划分，则可以将落在第 (x, y) 格子中的数据点的频率作为 $p(x, y)$ 的估计：

$$p(x, y) = \frac{n_{x,y}}{N} \tag{2}$$

其中 $n_{x,y}$ 为格子 (x,y) 中的数据个数，N 总样本数。

则根据互信息量的计算公式，可以得到随机变量 X, Y 的互信息 I 为：

$$I = \sum_{x \in X, y \in Y} p(x, y) log \frac{p(x, y)}{\sum_{x \in X} p(x, y) \sum_{y \in Y} p(x, y)} \tag{3}$$

计算互信息在不同网格中的最大值并为其添加归一化因子，则可以得

到最大信息系数 MIC 为：

$$MIC = max_{mn \le B}\left(\frac{max_{x \in X, y \in Y} I}{log(min(m,n))}\right) \tag{4}$$

其中 B 用于限制网格的个数，一般取 $B = N^{0.6}$。

将 Age 作为因变量，对 genus 数据集和 otu 数据集中的所有变量计算与年龄的 MIC，并绘制成散点图：

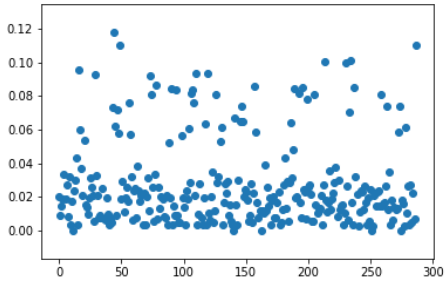

图 4: 数据集 genus 的 MIC

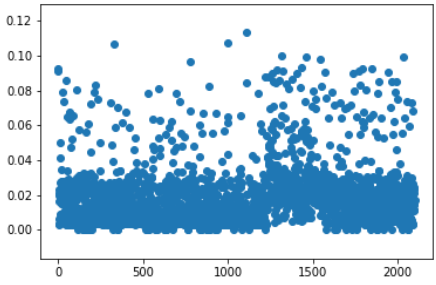

图 5: 数据集 otu 的 MIC

从图4和图5可以看到，大部分特征的 MIC 值都十分集中，接近于 0(即与 Age 没有明显的联系)。若将 MIC 值等于 0.04 作为阈值，可以发现大于该值的特征数量较少，MIC 值均较大，与 MIC 值小于 0.04 的特征有明显的分化趋势。因此将 $MIC = 0.04$ 作为筛选特征的一个衡量标准。

2.2 距离相关系数

与最大信息系数类似，距离相关系数也是用于衡量特征之间相关性的一个统计量，且其具有较好的鲁棒性。特征 a, b 的距离相关系数 $dcor$ 的计

算方式如下:

$$dcor(a,b) = \frac{dcov(a,b)}{\sqrt{dcov(a,a)dcov(b,b)}} \tag{5}$$

其中:

$$dcov^2(a,b) = S_1 + S_2 - 2S_3 \tag{6}$$

其中 S_1, S_2, S_3 分别表示为：

$$S_1 = \frac{1}{n^2}\sum_i\sum_j ||a_i - a_j||_1||b_i - b_j||_1 \tag{7}$$

$$S_2 = \frac{1}{n^2}\sum_i\sum_j ||a_i - a_j||_1\frac{1}{n^2}\sum_i\sum_j ||b_i - b_j||_1 \tag{8}$$

$$S_3 = \frac{1}{n^3}\sum_i\sum_j\sum_l ||a_i - a_l||_1||b_j - b_l||_1 \tag{9}$$

距离相关系数也是介于 0 与 1 之间的统计量。距离相关系数为 0，说明两个特征之间相互独立，相反其值越大，说明两个特征之间的相关程度越大。

依旧是将 Age 作为因变量，对 genus 数据集和 otu 数据集中的所有变量计算与年龄之间的最大相关系数，并将最大相关系数绘制成散点图如图6与图7所示。

从图6和图7可以看到，大部分特征的最大相关系数值都十分集中，接近于 0(即与 Age 没有明显的联系)。若将 dcor 值等于 0.06 作为阈值，可以发现大于该值的特征数量较少，dcor 值均较大，与 dcor 值小于 0.06 的特征有明显的分化趋势。因此将 $dcor = 0.06$ 也作为筛选特征的一个衡量标准。

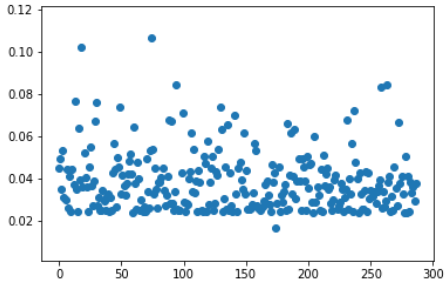

图 6: 数据集 genus 的 dcor

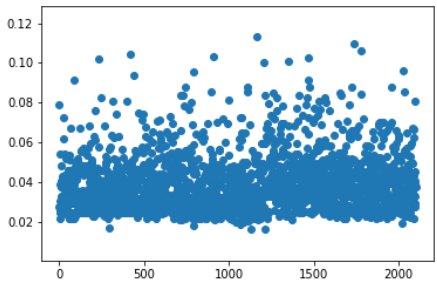

图 7: 数据集 otu 的 dcor

2.3 递归 XGBOOST

2.3.1 XGBOOST

前面两个小节将最大信息系数值大于 0.04 以及最大相关系数值大于 0.06 的特征初步筛选出来，作为与因变量年龄有一定相关关系的特征来进一步研究。在数据集 genus 中总共选出了 18 个特征，在数据集 otu 中总共选出了 70 个特征。

为了进一步研究肠道细菌特征和年龄之间的关系，找到对最终预测有更大影响的变量，降低预测噪声，本文将 XGBOOST 模型作为基模型，采用递归的方式对特征进行进一步筛选。

XGBOOST 是一种集成学习算法，每一个子树是一棵 CART 树。每一棵子树利用基尼纯度来衡量每一个特征的重要性，并且以此来选择分裂特征，而某个特征在所有子树中作为分裂结点的出现次数就是其对应的特征重要性。

本文将年龄作为因变量，将之前筛选出的 88 个特征作为输入，得到的特征重要性如图8所示 (为了清楚的展示得分概况，下图只随机选取了 20 个

特征来作为样例)：

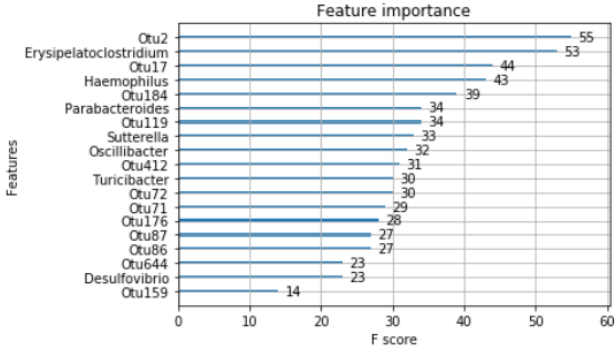

图 8: XGBOOST 下特征的重要性

可以看到变量 Otu2 得分最高，对于预测而言就最为重要，变量 Otu159 得分最低，对于预测而言就相对最不重要。

2.3.2 递归筛选

在选择 XGBOOST 为基模型后，利用递归筛选的方式来选择最优的特征子集。即初始子集是原始特征集合 (包含 88 个特征)，每次删去当前集合中一个最不重要的特征，利用 3 折交叉验证计算模型的预测得分。最终选取得分最高的子集作为最终的特征集合，具体算法流程如下：

算法一：递归筛选特征
step1：初始特征集合为 F
step2: 选出其中最不重要的特征 F_i，并且记录此时的模型得分 s
step3: 将特征集合更新为 $F - F_i$，跳转回 step2 直至 $F = \varnothing$
step4: 根据记录的模型得分选出最优得分对应的集合作为最终特征集合 S

将 88 个特征输入算法，最终筛选得到 39 个特征，genus 数据集包含 7 个特征，otu 数据集包含 32 个特征。这些特征如下表所示 (列举 15 个)：

表 1: 重要特征

Alistipes	Anaerostipes	Blautia	Otu67	Otu68
Erysipelatoclostridium	Granulicatella	Otu10	Otu8	Otu80
Salmonella	Turicibacter	Otu119	Otu102	Otu99

3 神经网络

3.1 网络的架构

本文选择只包含全连接层的神经网络并且结合 dropout 层来进行年龄预测。网络示例如图9所示：

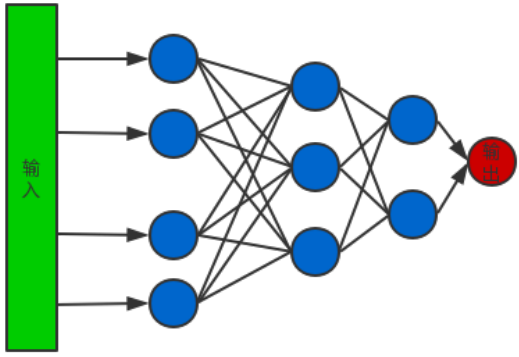

图 9: 神经网络示意图

为了使样本各个特征的比重相同，且使得神经网络的反向传播算法得以很好的收敛，所有输入数据均先进行标准化处理。

3.2 超参数

在确定完网络架构后，模型依旧有几个重要的超参数需要确定。

学习率的确定，学习率需要根据模型的输出结果手动调节，一般初始设置为 0.01，若模型损失函数变化很小，则需要适当调大学习率，若模型损失函数反复变化，则需要降低学习率，并采用自适应学习率。

学习批次的确定，由于样本量较大，若完整训练的数据集再更新权重会耗费大量运算时间，选择训练批次是合理的解决方法，一般批次大小取 2 的指数次方倍，本文的批次设置为 32,64,128 或 256。从中选取损失最小的值作为最终批次。

模型层数，层数设定为 3 层，对于卷积神经网络更多的隐藏层意味着更强的解释能力，但对于全连接层网络，更深的层对网络解释能力帮助并不大，但是 3 层网络普遍由于 2 层网络。

隐藏层结点数确定，一般将隐藏层结点数设置为一个大于输入值的数，本文将三个隐藏层结点数均设置为 64。

网络权重的初始值均采用标准正态分布随机数来初始化，并且确保值在 2 倍方差范围内 (确保网络的收敛性)。[2]

[2]网络的训练以及最终的结果讨论会在最终报告中展示