

泌乳牛自由采食量预测

王国赛*

July 14, 2017

1 问题描述

泌乳牛的草料采食量会影响其产奶量及其健康状况。在实际生产中草料需要提前 1 天时间制备,并且 1 天后吃剩的草料将不再提供给牛食用,以免变质的草料影响牛的健康情况。因此,饲养员需要提前估计未来 1 天内牛的自由采食量,并由此上报需要制备的草料量。理想情况下,饲养员希望控制报草量(即实际发放至牛棚的草料量)略大于牛的自由采食量,即一天后草料有剩余但剩余不多,以免不能满足牛的草料需求或者造成草料的浪费。具体地,牧场规定草料剩余量最好不超过当天报草量的 5%。

泌乳牛的草料采食量受到多方面因素的影响,因而每天每牛棚的自由采食量会发生波动。传统生产中饲养员会根据经验对相关因素进行估计判断,以估计每天每牛棚的报草量。这种方式可能会由于饲养员的经验差异导致或大或小的预测误差,可能会造成饲料不足或饲料浪费的问题。本工作希望通过对历史采集数据的分析,量化地构建预测泌乳牛采食量的模型,从而给饲养员提供较准确的采食量预测作为参考,辅助饲养员更好地制定报草量。

本文分为以下几个部分:第2节简述关于采食量预测的相关工作情况。第3节对采食量预测问题进行分析并形式化地描述问题。第4节概述建模所使用的数据以及对数据的处理方式,第5节介绍预测模型,第6节展示实验结果并对结果进行分析。第7节讨论下一步的工作方向。

2 相关工作概述

《奶牛营养需要》第7版(NRC 2011)^[6]中,第一章讨论奶牛的干物质采食量(Dry Matter Intake, DMI)。下文引自《奶牛营养需要》中文版原文。

* 邮箱: wgs14@mails.tsinghua.edu.cn

¹该手册有中文版,百度搜索其名称即可找到、下载。

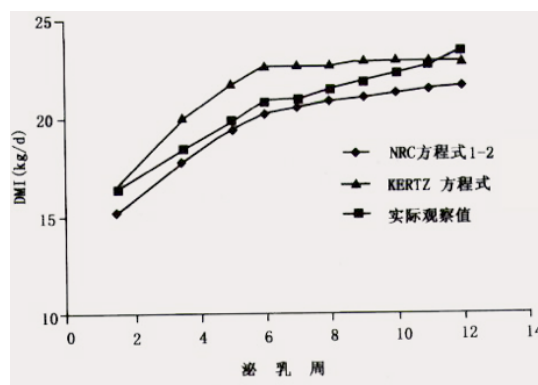


Figure 1: 用方程式1和 KERTZ 等 (1991) 推荐方程式预测奶牛泌乳早期干物质采食量变化。(图中方程式 1-2 对应本文中方程式 1。)

用来预测荷斯坦泌乳牛 DMI 的方程式为:

$$DMI(kg/d) = (0.372 \times FCM + 0.0968 \times BW^{0.75}) \times (1 - e^{-0.192 \times (WOL + 3.67)}) (1)$$

式中 FCM=4% 校正乳产量 (kg/d); BW= 体重 (kg); WOL 为泌乳周龄; $1 - e^{-0.192 \times (WOL + 3.67)}$ 为校正泌乳早期 DMI 下降的校正项。对于泌乳早期的产奶牛来说,方程式 1-2 预测的结果与 Kertz 等 (1991) 所建立方程式的预测结果相一致。最初 14 周龄泌乳牛干物质采食量以不同方程式预测的比较结果列于图1。

方程式1的数据全部来自荷斯坦奶牛。目前还没有公开发表关于 DMI 的数据用于发展或修正目前预测 DMI 的方程式,以便能用在荷斯坦牛以外其他品种牛上。关于娟姗牛 DMI 的预测问题,请参见 Holter 等 (1996) 的文章。

DMI 预测方程式用于经产奶牛可不必进行校正。在热中温区 (5~20°C) 以外,泌乳牛的 DMI 受到环境的影响。Eastridge 等 (1998) 和 Holter 等 (1997) 的研究都表明,当环境温度在 20°C 以上时,

DMI 随温度的升高而下降。由于没有足够的数据来确定热中温区以外环境对 DMI 的影响程度, 本版 NRC 泌乳牛 DMI 预测方程式 (方程式1) 没有考虑温度或湿度校正因子。

3 问题分析和形式化描述

3.1 采食量的影响因素

泌乳牛自由采食量主要受到内部因素和外部因素影响。部分相关因素列举如下:

内部因素:

- 奶牛品种
- 奶牛泌乳期 (初产牛、经产牛)
- 奶牛泌乳周 (泌乳阶段, 如高产、中产、低产、干奶)
- 奶牛体重
- 奶牛产奶量 (产奶净能)
- 奶牛运动量
- 奶牛身心状态 (疾病、情绪等)

外部因素:

- 饲料特性
- 温度湿度 (热应激)
- 其他应激 (如疫苗注射, 受到惊吓等)
- 其他环境因素 (如较宽的槽位可提升采食量)

理想情况下, 在预测牛的采食量时, 模型应将尽可能多的上述因素纳入考虑。但实际情况中常面临两个问题: (1) 数据类别采集不全, (2) 数据量积累较少。问题 (1) 会制约模型预测目标值采食量的能力, 因为未观测/记录到的因素会对采食量带来模型无法预测的波动。问题 (2) 会影响模型的精度, 因为通常基于机器学习或者数据挖掘技术构建的模型在训练集越大, 模型性能会越好, 尤其是一些复杂度高的模型如人工神经网络 (Artificial Neural Network) 深度学习 (Deep Learning) 等。故在本工作中我们会取舍地考虑部分因素。详情请见第4、5节。

3.2 以牛棚为建模单位

生产环境中, 牧场对草料的发放以及对剩草量的统计以牛棚为单位, 故我们不以单头牛的采食量为预测目标, 而是以牛棚中牛群整体的采食量 (或者等价的牛棚中头均采食量) 为预测目标。通常同牛棚内牛的品种相同, 泌乳期相近或相同, 泌乳周相近或相同, 故我们在建模时可忽略牛棚内单头牛之间的差异, 仅考虑牛群整体的特性 (或等价的头均特性)。

3.3 问题形式化

生产环境中, 每牛棚每天上报一次草量, 对应当天中午、晚上和第二天早上三次投喂草料的总草量。

我们记某牛棚第 t 天上午上报的头均草量 (报草量/牛的数量) 为 r_{t+1} ², 实际采食量为 y_{t+1} 。我们用向量 X_t 表示第 t 天的输入变量。具体地, X_t 的每一个元素为一个关于牛的或关于环境的观测数据, 如第 t 天牛的产奶量等等。

我们希望得到一个预测模型 f 使得

$$\hat{y}_{t+1} = f(X_t, X_{t-1}, \dots, X_{t-k}) \approx y_{t+1} \quad (2)$$

其中 \hat{y} 表示对 y 变量的预测值, 非负整数 k 表示我们在第 t 天时, 回顾历史的时间跨度。在最简单的模型中, $k = 0$, 即

$$\hat{y}_{t+1} = f(X_t) \approx y_{t+1} \quad (3)$$

本项工作中我们主要采用平均绝对误差 (Mean Absolute Error, MAE) 来衡量模型 f 的预测精度。对于 n 条样本 y_1, y_2, \dots, y_n 和模型对它们的预测值 $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$, 定义平均绝对误差 ϵ_{MAE} 如下:

$$\epsilon_{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4)$$

4 数据集和特征构造

4.1 数据概览

本项工作使用的主要数据集取自《剩草量分析表》, 包含 2017 年 3 月 5 日至 7 月 4 日约 4 个月的各牛棚采食情况记录。数据集包含的牛棚列举如表1所示。数据集记录了 16 个牛棚每天的牛头数,

²用 $t+1$ 而不是 t 表示是为了避免在时间先后上引起歧义, 可以理解 r_{t+1} 为给牛分配的吃到第 $t+1$ 天的草量。

Table 1: 牛棚概览（备注信息来自《剩草量分析表》，备注状态记录的时间不确定）。

| 牛棚 | 备注 | 牛棚 | 备注 |
|-------|----|-------|----|
| B1-1N | 怀孕 | B1-1S | 头胎 |
| B1-3N | 多胎 | B1-3S | 头胎 |
| B1-5N | 头胎 | B1-5S | 头胎 |
| B1-7N | 头胎 | B1-7S | 多胎 |
| B4-1N | 参配 | B4-1S | 怀孕 |
| B4-3N | 蹄病 | B4-3S | 高产 |
| B4-5N | 多胎 | B4-5S | 头胎 |
| B4-7N | 多胎 | B4-7S | 头胎 |

报草量，减草量³，剩草量和头均产奶量。通过简单计算可以进一步得到计划头均采食量 ($r_t = (\text{报草量} - \text{减草量}) / \text{牛头数}$) 和实际头均采食量 ($y_t = (\text{报草量} - \text{减草量} - \text{剩草量}) / \text{牛头数}$)。

温度湿度是影响头均采食量的重要影响参数。该数据可从记录历史天气的网站如 [7] 上抓取获得。目前由于数据量不大（约为 $16 \times 120 = 1920$ 条），为了避免模型过拟合，初步实验中仅考虑了日最高气温，未考虑日最低气温和湿度。

4.2 关于部分未采用数据说明

- 奶牛品种：各牛棚主要包含两种奶牛：荷斯坦奶牛和娟姗奶牛（同牛棚同日内牛群属同种）。目前暂无每日各牛棚的种类数据。由于当前数据量不大，先考虑不区分牛的品种，对所有数据统一处理。
- 泌乳期和泌乳周：该数据对头均采食量影响较大。当前数据集中无泌乳期泌乳周信息，下一步工作可优先考虑将该信息合并进数据集。
- 奶牛体重：我们将牛棚的牛群作为整体分析其头均采食量，故先假设各牛棚头均体重相似，不是影响头均采食量的主要因素。
- 奶牛运动量：当前先假设各牛棚牛群每日运动量相近。如有计步器数据，可将步数信息和采食量做关联性分析。
- 奶牛身心状况：奶牛身心状况难以量化。如有牛棚专用于养殖疾病奶牛（B4-3N 是否长期属

³每天有一次机会可以对上报草量进行增减。例如观察到中午牛食欲不振，则可以适当减少当天至第二天的总草量。减草量也可以为负值，表示适当增加草量。

于该种情况?)，可将该类牛棚排除在模型输入数据之外。

- 饲料特性：各牛棚除新产牛外（? 待验证），主要采用同种配方的饲料。本项工作主要针对处高产期的牛的采食量建模，故模型未将饲料相关的数据视作输入变量。
- 其他应激：当前数据集未包含疫苗注射记录。据技术人员称疫苗注射会令牛产生应激，影响短期内采食量。下一步工作可将各牛棚疫苗注射记录纳为模型输入变量。

4.3 数据预处理和特征构造

根据《剩草量分析表》中“牛只类型”字段标注，各牛棚在大部分日期内牛属于“高产”，但个别牛棚在部分日期内被标注为“新产”（生产完牛犊后处于泌乳初期的牛?）。在建模前我们将所有标注为“新产”的数据样本剔除掉。同时部分日期的数据存在缺失值。当前我们采取最简易的缺失值处理手段：将不完整的数据样本剔除掉。

在预测第 $t+1$ 天头均采食量 y_{t+1} 时，本工作主要尝试两种构建输入变量 X 的方式：(1) 仅考虑第 t 天的观测情况（温度、头均采食量、头均产奶量），和 (2) 考虑到一段连续日期即 $t-k, t-k+1, \dots, t$ 天的观测情况（温度、头均采食量、头均产奶量、连续 k 天头均采食量的变化梯度⁴）。方式 (2) 主要希望通过引入时域信息、观测数据变化趋势信息来弥补方式 (1) 输入变量较有限的不足。两种方式构建模型的性能请见第6节。

5 采食量预测模型：XGBoost

5.1 XGBoost 模型简介

本工作采用 XGBoost 对采食量进行建模。XGBoost[8] 是一个专注于梯度提升 (gradient boost) 算法的机器学习函数库，主要特点是具有优良的学习效果和高效的训练速度。该函数库诞生于 2014 年 2 月，最初由陈天奇博士设计、开发，论文 [1] 具体介绍了其实现原理和细节，演示文稿 [10] 概述了其主要思想和算法原理。仅在 2015 年，在 Kaggle[4] 竞赛中获胜的 29 个算法中，有 17 个使用了 XGBoost

⁴用线性回归模型拟合 k 天的头均采食量，取拟合直线的斜率，以刻画头均采食量变化趋势。

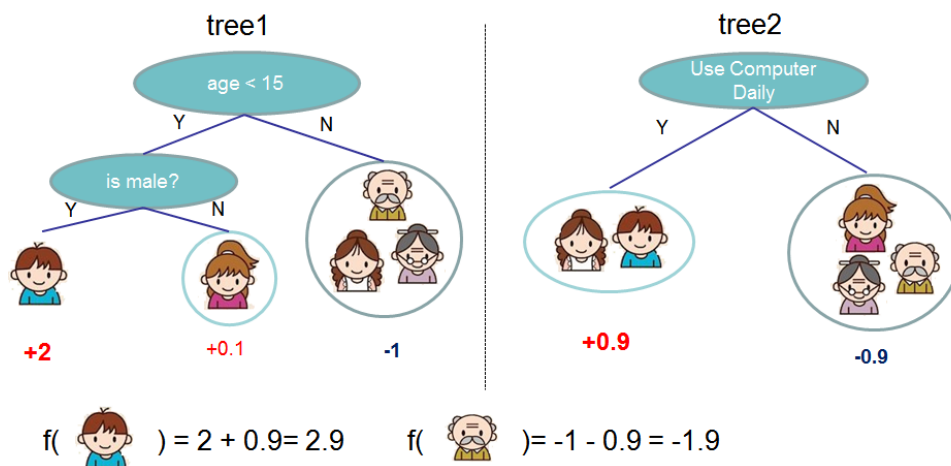


Figure 2: 一个树集成的例子。该包含两棵树的模型解决分类问题，输入变量是每个人的年龄、性别、职业等特征，分类目标是“此人是否喜欢电脑游戏”。树集成将每棵树的输出结果组合起来（本例中为同权重地相加），例如小男孩分类结果为 2.9，老爷爷分类结果为-1.9。如果我们设定模型分类阈值为 0，则模型认为小男孩喜欢电脑游戏，老爷爷不喜欢。

库。在 KDDCup 2015[5] 竞赛中，排名前十的队伍全部使用了 XGBoost 库。

XGBoost 用于解决监督学习 (supervised learning) 问题，在监督学习中模型利用训练集进行学习，以通过给定的量 x_i (多个特征) 来预测 y_i 。XGBoost 是一种树集成 (tree ensemble) 模型⁵，即通过学习得到多棵树，组合起来解决分类问题或回归问题。图2展示了一个树集成的例子。

XGBoost 优化的目标函数如式5所示：

$$obj = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \quad (5)$$

其中 $l(y_i, \hat{y}_i^{(t)})$ 表示模型对第 i 个训练集样本的预测误差（误差指标的定义根据问题的不同可以不同）。在回归问题中，均方误差 MSE 是常用的误差指标）， $\Omega(f_i)$ 表示第 i 棵树的复杂度。树的复杂度作为正则项，目的是为了控制模型的复杂程度，避免过复杂的模型过度拟合 (overfit) 训练集而缺乏泛化能力（对未见过样本的预测能力）。

在学习模型参数时，XGBoost 采用相加策略：它递推地依次学习出各棵树，每次只增加一棵树，在学习第 t 棵树时，固定前 $t-1$ 棵树不变，而用第 t 棵树去矫正前 $t-1$ 棵树对于样本预测结果的误差（残差学习）。具体地，学习第 t 树时，优化目标函

数 $obj^{(t)}$ 为：

$$obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{常数} \quad (6)$$

更多关于 XGBoost 算法细节请参考 XGBoost 网站文档、演示文稿和论文 [3, 10, 1]。下面我们讨论使用 XGBoost 建模时需要重点考虑的模型参数。

5.2 模型参数设置

XGBoost 的参数详解可参考 [9]⁶。

通用参数： booster 指定使用的 booster（提升器），默认值为 gbtrees。使用默认值即可。如想采用线性模型（采食量预测不适合采用线性模型），可以将其设为 gblinear。

Tree Booster 参数： n_estimator 是树的个数，如需改善过拟合问题，可适当调小 n_estimator。max_depth 是最大深度限制（模型构造树时可能尚未到此深度即停止扩展），如需改善过拟合问题，可适当调小 max_depth。由于当前预测采食量的模型使用输入变量很少，通常 2-3 层的树即足够。eta 是学习率，通常使用默认值即可，如需改善过拟合问题，可适当调小 eta。subsample 和相关几个参数 (colsample_bytree、colsample_bylevel) 通过对训练

⁵集成 (ensemble) 模型是一类组合多个性能相对弱的分类器/预测器实现强分类器/预测器的模型。所谓“三个臭皮匠赛过诸葛亮”。

⁶XGBoost 提供多种编程语言 (Python, R, Java, Scala, C++) 接口，该文档介绍 R 语言接口的参数，但其他语言参数与此基本相同。

集欠采样来避免模型欠拟合。当前预测采食量任务不需要欠采样。

其他参数一般不需要特别设置，使用默认值即可。

5.3 最优参数搜索

部分参数取值可以通过由在人为给定的参数空间中网格搜索（grid search，即暴力枚举每一种可能的参数组合）确定。为了避免过拟合，在搜索每组参数组合时，可采用多折交叉验证。本工作中如不特殊说明，我们采用 5 折交叉验证。

用 Python 的 scikit-learn 库中 model_selection 包里的 GridSearchCV 类 [2] 可便捷实现利用网格搜索和交叉验证确定最优参数组合。

6 实验分析

6.1 单日数据预测第二天采食量

本节训练模型使用的数据每条样本格式为

$$\langle (y_t, m_t, T_t^h), y_{t+1} \rangle \quad (7)$$

其中输入变量 y_t 为第 t 天某牛棚的头均采食量， m_t 为第 t 天头均产奶量， T_t^h 为第 t 天的最高气温，输出变量（预测值） y_{t+1} 为第 $t+1$ 天的头均采食量。经过数据预处理，得到数据样本约 1750 条。

在部分实验中我们采用了对照组，对照组作为比较的基准对象用最简易傻瓜的方式建模预测，具体地， $\hat{y}_{t+1} = y_t$ ，即用当天的实际头均采食量直接作为第二天的头均采食量预测值。

6.1.1 模拟对历史数据的拟合

XGBoost 模型拟合历史数据的平均绝对误差请见表2，部分样本的预测值和实际值请见图3。XGBoost 模型的参数 n_estimators 取 200，max_depth 取 2（参见第5.3节）。

表2显示单日数据构建的模型相比于对照组能够有效减小平均绝对误差 MAE， $\epsilon_{MAE} = 1.01$ 指每日预测头均和实际头均的平均绝对误差是 1.01 千克，平均误差率约在 2.5% ~ 3.4% 左右。 R^2 刻画模型对数据的拟合度，0.886 意味着拟合度很高，模型对历史数据拟合能力强。

另外通过 XGBoost 的 plot_importance 函数可以可视化不同特征对于模型的重要性或贡献程

Table 2: 单日数据建模拟合历史数据的误差。

| 指标 | 值 |
|----------------------|-------|
| ϵ_{MAE} | 1.01 |
| 对照组 ϵ_{MAE} | 1.274 |
| R^2 | 0.886 |

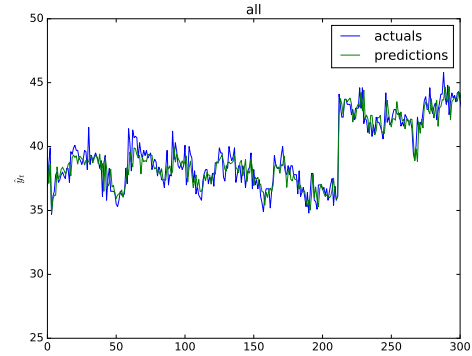


Figure 3: 单日数据建模在部分样本上的预测结果。

度，结果如图4所示。由图可知，当取单日观测数据拟合第二天头均采食量时，各特征重要性排序为：头均采食量 > 头均产奶量 > 最高气温。

模型对历史数据的拟合只能说明模型复杂程度足够高，但不能说明模型对于未见过的数据样本具有较强的预测能力，即对于有预测性质的任务，我们更需要关注模型的泛化能力。我们在下一小节做相关分析。

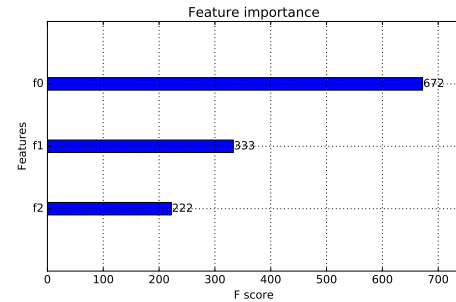


Figure 4: 不同特征对 XGBoost 模型的贡献程度。 f_0, f_1, f_2 分别对应数据样本中的 y_t, m_t, T_t^h 。

Table 3: 单日数据建预测时 5 折交叉验证的误差。

| ϵ_{MAE} | R^2 |
|------------------|-------|
| 1.202 | 0.728 |
| 1.221 | 0.797 |
| 1.243 | 0.744 |
| 1.359 | 0.724 |
| 1.430 | 0.560 |

6.1.2 模拟对未见数据的预测

我们采用交叉验证⁷的方式评估模型的泛化能力，即对历史未见数据的预测能力。

表3显示了对所有数据样本做 5 折交叉验证的结果（交叉验证时不同随机分组会导致不同的结果，表中显示随机挑选某次实验的结果）。比较表3和2，可知预测时的平均 $\epsilon_{MAE} = 1.291$ 略高于对照组的 ϵ_{MAE} ，模型相比于对照组预测失效。虽然在上一节观察到模型拟合历史数据的能力较强，但交叉验证的结果说明模型缺乏泛化能力，这主要由两个可能原因导致：(1) 在部分实验中 ϵ_{MAE} 较高，原因可能是在拆分训练集和测试集时，某些相对难预测的模式样本被分至测试集但未出现在训练集中（导致模型未能学到这些模式）。如数据集规模进一步扩大，则模型的预测性能应当进一步提升。(2) 仅以单日观测数据作为特征，缺乏对采食量变化模式的刻画能力。我们在第6.2节尝试改进措施。

6.1.3 分牛棚的拟合和预测分析

我们进一步分析模型对于不同牛棚的牛群采食量的预测能力。我们先用所有样本数据训练得到模型去拟合各个牛棚的数据，观察模型对于不同牛棚牛群采食量刻画效果的差异，结果如表4所示。

观察可以发现 (1) 各牛棚模型拟合平均绝对误差均低于对照组；(2) 不同牛棚拟合的误差有高低，大致和对照组拟合误差高低趋势一致。观察 (2) 说明模型的预测误差主要来自于日头均采食量的较大波动，而模型难以预测这种突然的升、降。 R^2 值最高的（模型拟合度最高的）几个牛棚为 B1-5S、B1-7N、B1-5N、B4-7S、B4-5S (R^2 均高于 0.75)。这些牛棚除 B4-5S 外，均为“头胎”，其原因可能是“头

Table 4: 不同牛棚头均采食量拟合的误差。

| 牛棚 | 牛只类型 | ϵ_{MAE} | 对照组 ϵ_{MAE} | R^2 |
|-------|------|------------------|----------------------|-------|
| B1-1N | 头胎 | 0.793 | 0.964 | 0.424 |
| B1-1S | 头胎 | 0.657 | 0.865 | 0.73 |
| B1-3N | 多胎 | 0.756 | 0.751 | 0.475 |
| B1-3S | 头胎 | 0.706 | 0.726 | 0.696 |
| B1-5N | 头胎 | 1.173 | 1.824 | 0.846 |
| B1-5S | 头胎 | 1.375 | 2.806 | 0.919 |
| B1-7N | 头胎 | 1.232 | 1.477 | 0.898 |
| B1-7S | 多胎 | 0.995 | 1.105 | 0.458 |
| B4-1N | 参配 | 1.095 | 1.56 | 0.584 |
| B4-1S | 怀孕 | 0.969 | 1.193 | 0.699 |
| B4-3N | 蹄病 | 1.248 | 1.455 | 0.501 |
| B4-3S | 高产 | 1.39 | 1.743 | 0.564 |
| B4-5N | 多胎 | 0.882 | 1.197 | 0.757 |
| B4-5S | 头胎 | 1.008 | 1.088 | 0.556 |
| B4-7N | 多胎 | 1.071 | 1.192 | 0.664 |
| B4-7S | 头胎 | 0.913 | 0.987 | 0.798 |

胎”牛占数据样本中的大多数，使得模型对于该类型牛的采食量模式刻画得较好。而牛只这也说明不同类型牛的采食量模式存在差异。因而在建模时应当考虑如下权衡：

问题：建模时应对不同品种、类型（头胎、多胎、参配等）的牛群分别构建模型，还是用同一模型建模，用某些特征来表征这种差异？

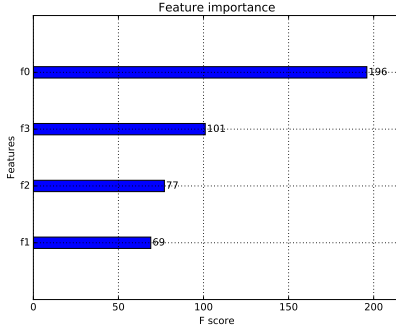
如果数据量充足，且能够提取富有表征能力的特征，则构建统一模型可以达到较好的预测效果（有待未来实验验证）。

但当前训练数据样本有限且特征不够丰富，我们先尝试简化问题进行分牛棚的实验：排除影响模型预测效果的因素，只选取牛只类别为“头胎”的牛棚，单独进行建模。如此筛选得到约 760 条数据样本，在所有“头胎”牛棚牛群上进行 5 折交叉验证，实验结果如表5所示。在只考虑“头胎”牛群时，对照组的平均绝对误差 ϵ_{MAE} 为 1.306。观察表格结果可见 5 折交叉验证的 ϵ_{MAE} 的平均值 1.328 高于对照组，且高于不区分牛群时的平均 ϵ_{MAE} ，说明模型预测失效。原因如第6.1.2节所述，主要有二：(1) 数据量过小，测试集中存在训练集未涵盖的采食量变化模式，导致模型泛化能力差；(2) 模型的输入变量缺乏富有表征力的特征。其中原因 (1) 尤为重要，因为如只选择“头胎”牛群，则数据集样本数仅为约 760，不到上节全部牛群数据量的一半。

⁷具体地， k 折交叉验证将所有数据样本随机平均分为 k 组，重复 k 次测试：每次测试用其中 $k-1$ 组数据样本组合成训练集，训练构建得到模型（预测器），并将模型用于剩下的一组数据样本（作为测试集）进行预测，并在测试集上评估相关误差指标。

Table 5: 单日数据建预测“头胎”牛群时 5 折交叉验证的误差。

| ϵ_{MAE} | R^2 |
|------------------|-------|
| 1.361 | 0.627 |
| 1.404 | 0.560 |
| 1.231 | 0.750 |
| 1.376 | 0.720 |
| 1.267 | 0.653 |

Figure 5: 不同特征对 XGBoost 模型的贡献程度。 f_0, f_1, f_2, f_3 分别对应数据样本中的 $y_t, m_t, T_t^h, g_{t,k}$ 。

6.2 加入时域信息预测第二天采食量

本节训练模型使用的数据每条样本格式为：

$$\langle (y_t, m_t, T_t^h, g_{t,k}), y_{t+1} \rangle \quad (8)$$

相比于式7新增了一项 $g_{t,k}$ ，表示从第 t 天开始向前回溯 k 天的采食量的梯度值。具体地， $g_{t,k}$ 是用线性回归拟合点集 $\{(1, y_{t-k+1}), (2, y_{t-k+2}), \dots, (k, y_t)\}$ 得到直线的斜率。我们希望引入该项来增加刻划时域上头均采食量变化趋势的特征。

我们设定 $k = 3$ （考虑过去 3 天采食量的梯度），经过数据预处理、缺失值剔除，得到全部数据样本约 1700 条。实验中 XGBoost 模型的参数 `n_estimators` 取 150，`max_depth` 取 2（参见第5.3节）。在所有数据样本做拟合分析各特征的重要性，结果如图5所示。由图可知，当引入时域特征拟合第二天头均采食量时，各特征重要性排序为：头均采食量 > 头均产奶量 > 最高气温 \approx 头均采食量时域梯度。

对所有数据样本做 5 折交叉验证的结果如表6所示。对照组（用当天头均采食量直接当做第二天头均采食量预测值）的 ϵ_{MAE} 为 1.274。分析表格可见 ϵ_{MAE} 的平均值 1.209 低于对照组，说明

Table 6: 加入时域信息预测所有牛群时 5 折交叉验证的误差。

| ϵ_{MAE} | R^2 |
|------------------|-------|
| 1.239 | 0.767 |
| 1.193 | 0.772 |
| 1.203 | 0.768 |
| 1.196 | 0.822 |
| 1.215 | 0.706 |

预测有效，虽然改进并不显著。

我们仿照第6.1.3节，剔除其他类别的牛群，仅对“头胎”牛群进行分析，则结果和第6.1.3节类似，预测误差 1.286 高于对照组 1.258，且高于不拆分牛群的预测结果（1.209）。数据集样本量不够大是主要原因。

7 小结与下一步工作方向

7.1 实验结论整理

xgboost 模型对历史数据的拟合能力很强。但是头均采食量预测任务要求模型具有较好的泛化性能（对未见过数据样本的预测能力）。因而不做交叉验证，不拆分训练集、测试集的实验结果意义不大。

在第6.1.3节中，我们提出了建模面临权衡的问题。虽然我们实验发现不同牛棚的头均采食量的波动模式不同，导致模型对不同牛棚头均采食量的刻划能力不同，但是通过实验发现，按照牛棚进行拆分，以“头胎”牛群为例，对单类别牛群单独进行建模并预测的效果并不好。我们认为最重要的原因是数据样本量不够大，导致模型不能充分学到头均采食量变动的各种模式。

通过实验我们发现使用单日观测数据对第二天进行预测的效果较差，当引入时域信息（用最近几日头均采食量的梯度来表征时域上采食量的变化趋势）后，模型的预测性能能够提升，并优于对照组，但提升并不显著。

通过对各特征的重要性进行分析，可知对 xgboost 模型预测第二天头均采食量最重要的特征是当天的头均采食量，其次是当天的头均产奶量，最高气温和时域采食量梯度信息权重相对最低。

7.2 下一步工作方向

接下来工作主要分为可并行开展的三块：数据集扩充，特征扩充，特征工程。

数据集扩充：扩大数据集样本量，积累更多数据。数据积累需要时间。如历史上（2017 年 3 月以前）有各牛棚新增的采食量（或剩草量）数据，也可合并进入当前数据集。且在未来如有可能可在牛棚中增加部署传感器，例如温湿度传感器等等。

特征扩充：如前文所述，当前优先考虑增加的特征包括：泌乳期、泌乳周，其他有记录的应激情况（如牛棚疫苗注射记录）。如有可能，可另外增加牛的运动量数据（如计步器采集数据）。同时我们将增加分析湿度数据，和最高气温一并考虑。

特征工程：实验显示时域信息对于模型预测性能有提升作用。我们将继续在此方面进行实验（尝试各种特征提取、变换方式）。同时对新纳入的输入变量如泌乳期、泌乳周，应激情况，湿度等，我们也尝试进行特征工程处理，实验分析其对模型性能的影响。

参考文献

- [1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 2016.
- [2] Tuning the hyper-parameters of an estimator. http://scikit-learn.org/stable/modules/grid_search.html.
- [3] Introduction to Boosted Trees. <http://xgboost.readthedocs.io/en/latest/model.html>.
- [4] Kaggle. <https://www.kaggle.com/>.
- [5] KDD Cup 2015. <http://kddcup2015.com/submission-rank.html>.
- [6] National Research Council Subcommittee on Dairy Cattle Nutrition USA. Nutrient requirements of dairy cattle. National Academy of Sciences, 2001.
- [7] 天气后报. <http://www.tianqihoubao.com/>.
- [8] XGBoost: eXtreme Gradient Boosting. <https://github.com/dmlc/XGBoost>.
- [9] XGBoost Parameters. <https://github.com/dmlc/xgboost/blob/master/doc/parameter.md>.
- [10] Introduction to Boosted Trees. <http://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf>.