李喜阳,李信颉,赵志超,等. 基于机器学习方法的母猪高低产分类模型研究[J]. 华中农业大学学报,2021,40(3): 221-229.
DOI:10.13300/j.cnki.hnlkxb.2021.03.025

基于机器学习方法的母猪高低产分类模型研究

李喜阳[1],李信颉[1],赵志超[2],李长春[1,2],刘向东[1,2]

1.华中农业大学动物科学技术学院/农业动物遗传育种与繁殖教育部重点实验室,武汉 430070;
2.农业农村部生猪健康养殖重点实验室/广西扬翔股份有限公司,贵港 537100

摘要 为帮助猪场管理者更好地对母猪进行繁殖管理、预测母猪的高低产、及时淘汰低产母猪,收集和整理包含出生场地、分娩栏位、品种和不同胎次、初生窝重信息的 3 个母猪群体的生产数据集,制定母猪高低产的分类标准,使用 R 软件中的 Boruta 包筛选出影响母猪高低产的重要特征,使用 4 种不同的机器学习方法——逻辑回归(logistic regression,LOG)、决策树(decision tree,DT)、随机森林(random forest,RF)和支持向量机(support vector machine,SVM)构建母猪高低产的分类模型,并进行决策树视图分析探究影响母猪最高产的相关因素。结果显示:4 种机器学习方法构建母猪高产分类模型的分类准确率均在 71%左右,最高可达 84%,并且发现 SVM 作为最佳建模方法在所有数据集和不同分类标准下出现的频率最高,其次是 LOG 和 DT。决策树视图显示出生场地、品种和初生窝重是划分最高产母猪的重要叶节点,利用这些特征预测最高产母猪准确率可达 73%～82%。以上结果表明在未来的养猪生产中,利用机器学习方法实现母猪高低产的早期预测将会是一个不错的选择。

关键词 机器学习方法; 精准养猪; 母猪早期选育; 决策树; 随机森林; 支持向量机; 繁殖性能; 产仔数早期预测; 高繁殖力; 分类模型

中图分类号 TP 181; S 828 **文献标识码** A **文章编号** 1000-2421(2021)03-0221-09

母猪的产仔数性状是猪场生产成绩和母猪繁殖力的重要评定指标,据统计许多国家商业母猪群体的年淘汰率在 20%～50%,其中产仔数性状差是母猪淘汰的主要原因之一[1]。母猪的高繁殖力直接决定了规模猪场的经济效益。因此,早在 1980 年,为选育出高繁殖力的母猪群体,欧洲畜产协会统一了母猪产仔数性状的记录方法并将其标准化,最早选育母猪的指标包括总产仔数(total number born,TNB)、产活仔数(number born alive,NBA)和健仔数(number healthy piglets,NHP)等[2]。此外,Nielsen 等[3]和 Su 等[4]研究发现 5 日龄活仔数(number 5 day,N5D)与仔猪成活率之间存在中等遗传相关,对该指标的遗传改良将有利于提高仔猪成活率。因此,构建以上产仔数性状的分类模型,将有利于挖掘影响母猪生产水平的相关因素。

机器学习顾名思义就是让计算机学习,专门研究计算机怎样模拟或实现人类的学习行为,其不仅包含统计学知识,还是多学科知识交互应用的代表,例如其包含有大量的算法理论、概率论以及逼近理论等[5]。随着畜牧业的快速发展,所要处理和分析的数据量愈发庞大、数据结构愈发复杂,使得机器学习方法在畜牧领域得到了广泛应用。Bakoev 等[6]以猪的生长和肉质特征为指标,利用 9 种不同的机器学习分类算法来评估猪的四肢状态。Messad 等[7]利用梯度提升方法鉴定到的重要特征可作为猪饲料效率的可靠预测因子。Shahinfar 等[8-9]利用绵羊的生产管理数据,通过不同机器学习方法构建了绵羊早期胴体性状和绵羊羊毛质量的预测模型,取得了不错的预测效果。Tusell 等[10]基于猪的表型数据和基因组数据利用支持向量机预测猪的饲料效率和生长速度。李信颉等[11]使用几种不同的机器学习方法对母猪的产仔数性状进行预测。然而,之前的研究更多是对动物表型或经济性状的回归分析,涉及分类的研究较少。

收稿日期: 2020-09-11
基金项目:国家自然科学基金项目(31572375);分子育种与繁殖新技术研发与推广合作协议(70711818605)
李喜阳,E-mail:673683740@qq.com
通信作者:刘向东,E-mail:liuxiangdong@mail.hzau.edu.cn

因此，为探究影响母猪生产性能的相关因素（特征），筛选最佳的建模方法，本研究收集整理了包含以上产仔数性状的母猪群体数据集，针对不同产仔数指标制定母猪高低产的分类标准，利用4种不同的机器学习算法（逻辑回归、决策树、随机森林和支持向量机）构建母猪高低产的分类模型，并进行决策树视图分析，以期为实现高产母猪的早期选育提供参考。

1 材料与方法

1.1 数据的预处理

本研究收集整理了广西某猪场2016－2018年3个母猪群体的生产数据（以A、B、C数据集表示）。A数据集包含出生场地、分娩栏位、品种、第1胎初生窝重、第2胎初生窝重和第3胎的产仔数性状，B数据集包含出生场地、分娩栏位、品种、第1胎初生窝重、第2胎初生窝重、第3胎初生窝重和第4胎的产仔数性状，C数据集包含出生场地、分娩栏位、品种、第1胎初生窝重、第2胎初生窝重、第3胎初生窝重、第4胎初生窝重和第5胎的产仔数性状。正态性检验表明各胎次产仔数性状均近似符合正态分布。使用SPSS 19.0和Excel 2019对数据集进行预处理，剔除缺失值，并使用R软件对不同数据集的母猪产仔数性状进行描述性统计（表1）。

表1 不同数据集产仔数性状的描述性统计

Table 1 Descriptive statistics of litter size traits in different data set

数据集 Data set	产仔数性状 Litter size trait	母猪数量 Number of sow	产仔数范围 Range of litter size	均值±标准差 Mean±SD
A	第3胎总产仔数 TNB3	3 658	1～32	15.11±4.79
	第3胎产活仔数 NBA3	3 658	0～27	14.14±4.60
	第3胎健仔数 NHP3	3 658	0～25	13.03±4.27
	第3胎5日龄活仔猪数 N5D3	3 658	0～24	12.65±4.20
B	第4胎总产仔数 TNB4	2 272	1～30	14.99±4.71
	第4胎产活仔数 NBA4	2 272	0～26	13.86±4.38
	第4胎健仔数 NHP4	2 272	0～25	12.74±4.19
	第4胎5日龄活仔猪数 N5D4	2 272	1～25	12.35±4.09
C	第5胎总产仔数 TNB5	1 487	1～28	14.46±4.54
	第5胎产活仔数 NBA5	1 487	0～25	13.22±4.26
	第5胎健仔数 NHP5	1 487	0～24	12.20±3.99
	第5胎5日龄活仔猪数 N5D5	1 487	1～24	11.73±3.87

1.2 母猪高低产分类标准的制定

结合近年来我国核心母猪的生产水平[12]制定母猪高低产的分类标准。如表2所示，以A数据集为例，总产仔数大于等于18头、产活仔数大于等于17头、健仔数大于等于16头、5日龄产仔数大于等于15头的母猪为高产母猪，其余为低产母猪，以此类推，最后将产活仔数和5日龄仔猪数归纳为一个综合指标对所有数据集中的母猪进行再分类，形成最高产母猪。

1.3 筛选构建分类模型的重要特征

使用R软件中的Boruta包对A、B、C 3个数据集中所包含的变量进行特征筛选[13]，特征筛选结果如图1～3所示：除C数据集中的产活仔数模型中的第1胎初生窝重外（图3），3个数据集中所包含的其他变量对母猪产仔数性状分类模型的构建均重要，其中出生场地的重要程度均最高。

1.4 机器学习方法

1）逻辑回归（logistic regression，LOG）。逻辑回归是一种应用非常广泛的机器学习分类算法，它将数据拟合到一个logit函数中，从而完成对事件发生概率的预测。相比传统回归方法，逻辑回归弥补了线性回归无法处理分类问题的缺陷，其判别性能主要基于Sigmoid函数来实现，函数表达式如下：

$$f(x)=\frac{1}{1+e^{-x}}$$

通过Sigmoid函数计算特征得出相应的概率值，大于某一概率阈值的划分为一类，小于某一概率阈值的划分为另一类，以此来判断样本类别[5]。

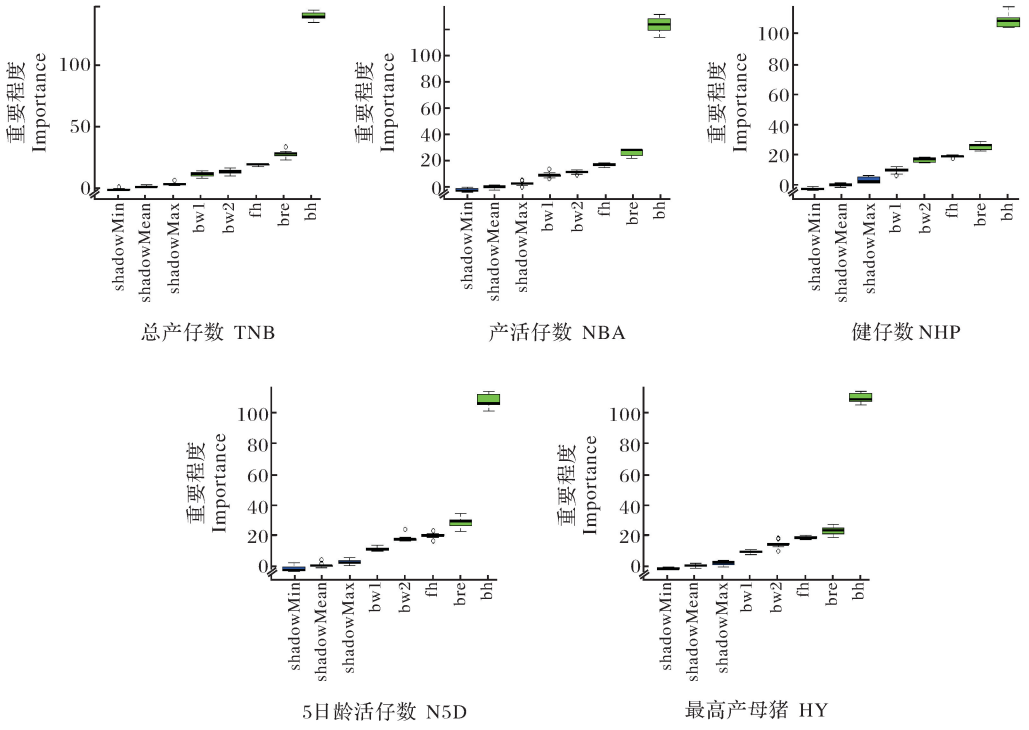
2）决策树（decision tree，DT）。决策树作为最基础、最常见的有监督学习模型，常被用于分类问题和回归问题，它是一种以树结构形式表达的预测分析模型，其独特的树型分类图中从根节点到叶节点每一处都代表了一种特征。决策树算法的重要理论基础是“基尼指数”和“信息熵”，其为量化信息的分析工具。熵代表元素的随机性程度，在数学上，它可以借助于变量的概率来计算：$H=-\sum p(x)\log(x)$，

其中 x 表示离散随机变量，$p(x)$表示变量 x 发生的概率[14]，概率越大，熵值越小，反之熵值越大。基尼系数和熵值的定义类似，基尼系数越大，熵值也越大，说明元素的随机化程度越高。

表 2　高低产母猪的分类标准

Table 2　Classification standard for high and low production sows

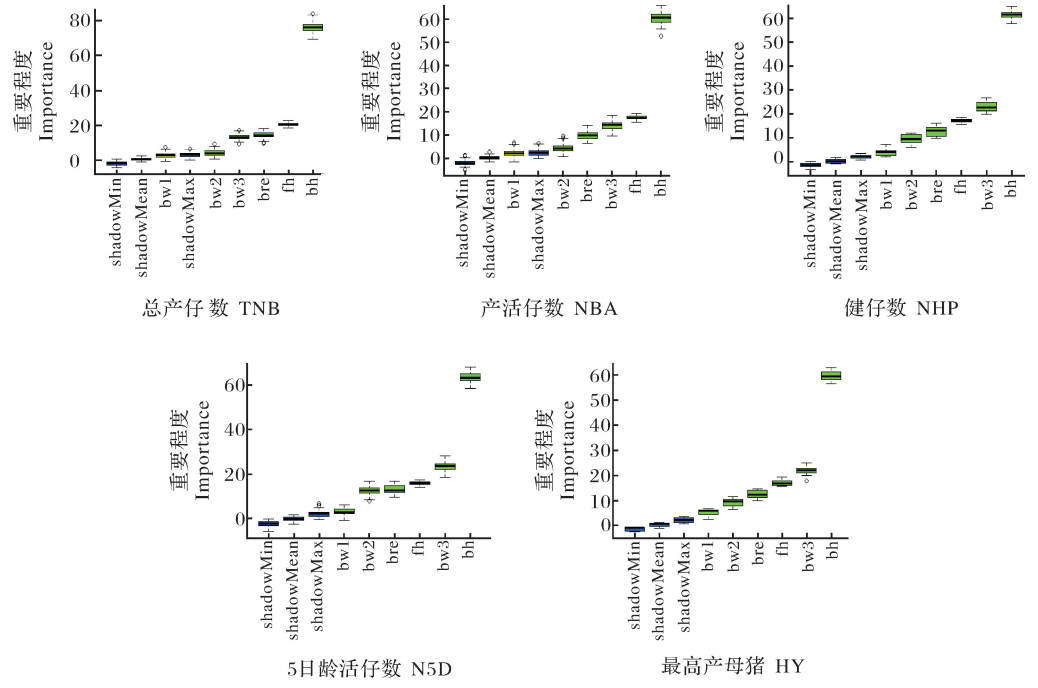
数据集 Data set	总产仔数 TNB	产活仔数 NBA	健仔数 NHP	5 日龄活仔数 N5D	产活仔数+5 日龄活仔数 NBA+N5D
A	≥18：高产 High yield	≥17：高产 High yield	≥16：高产 High yield	≥15：高产 High yield	≥17+15：最高产 Highest yield
	<18：低产 Low yield	<17：低产 Low yield	<16：低产 Low yield	<15：低产 Low yield	<17+15：最低产 Lowest yield
B	≥19：高产 High yield	≥17：高产 High yield	≥16：高产 High yield	≥15：高产 High yield	≥17+15：最高产 Highest yield
	<19：低产 Low yield	<17：低产 Low yield	<16：低产 Low yield	<15：低产 Low yield	<17+15：最低产 Lowest yield
C	≥18：高产 High yield	≥17：高产 High yield	≥16：高产 High yield	≥16：高产 High yield	≥17+16：最高产 Highest yield
	<18：低产 Low yield	<17：低产 Low yield	<16：低产 Low yield	<16：低产 Low yield	<17+16：最低产 Lowest yield



shadowMin：阴影属性的最小值；shadowMean：阴影属性的均值；shadowMax：阴影属性的最大值；阴影属性的最小、平均和最大值为数据集中的阈值，高于阈值水平的特征为重要特征，红色、黄色和绿色方框代表拒绝、暂定和确认的特征；bw1：第 1 胎初生窝重；bw2：第 2 胎初生窝重；fh：分娩栏位；bre：品种；bh：出生场地；TNB：总产仔数；NBA：产活仔数；NHP：健仔数；N5D：5 日龄活仔数；HY：最高产母猪。下图同。shadowMin：Minimum value of the shadow attribute；shadowMean：Average value of the shadow attribute；shadowMax：Maximum value of the shadow attribute；the minimum，average and maximum Z values of the shadow attributes are thresholds in the data set. Features above the threshold level are important features. Red，yellow and green boxes represent rejected tentative and confirmed features；bw1：Birth weight of first litter；bw2：Birth weight of second litter；fh：Farrow herd；bre：Breed；bh：Birth herd；TNB：Total number born；NBA：Number born alive；NHP：Number healthy piglets；N5D：Number 5 day；HY：Highest yield. The same as below.

图 1　分类模型的特征筛选图(A 数据集)

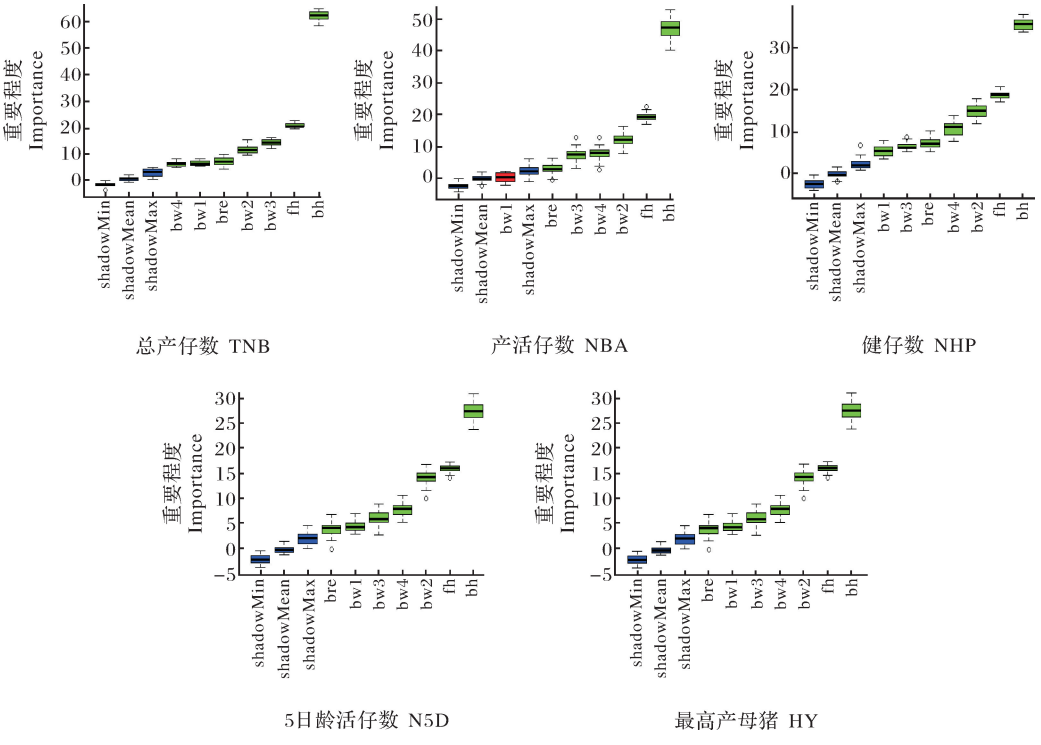
Fig.1　Feature screening diagram of classification model(A data set)



bw3:第三胎初生窝重;bw3:Birth weight of third litter.

图2 分类模型的特征筛选图(B数据集)

Fig.2 Feature screening diagram of classification model(B data set)



bw3:第三胎初生窝重 Birth weight of third litter; bw4:第四胎初生窝重 Birth weight of fourth litter.

图3 分类模型的特征筛选图(C数据集)

Fig.3 Feature screening diagram of classification model(C data set)

3)随机森林(random forest,RF)。随机森林是包含多棵决策树分类器的集合学习算法,在处理决策问题时,会根据集合思想构建多个分类决策树,同时进行决策,最后"遵循少数服从多数的原则"来确

定最终结果，充分避免了单一决策树所产生的决策偶然性，提高了分类的可信度及准确率。

4）支持向量机（support vector machine，SVM）。支持向量机是一种二分类模型，基本思想是求解能够正确划分训练数据集且几何间隔最大的分离超平面。其在线性和非线性的数据结构中都具有很高的应用价值，在处理存在线性关系的数据集时，最佳线性划分方程为：$w^Tx+b=0$。空间中点 x 到最佳分类超平面的距离公式为 $d=\frac{|-w^Tx+b|}{||w||}$，只要使得距离最大，就可以求出最佳的划分线和最佳分类超平面[15]。对于输入空间中的非线性分类问题，通过非线性变换将它转化为某维特征空间中的线性分类问题，在高维特征空间中学习线性支持向量机。

1.5　分类模型性能的评估

分类模型的评估是在已知特征和类别的训练集上构建，再利用从已知的原始数据集中拆分出一部分作为测试集对模型的分类性能进行评估，常使用混淆矩阵来计算其评估指标。本研究首先依据分类标准对 A、B、C 3 个数据集的产仔数性状进行二元处理，然后对数据集随机拆分，其中 70% 的数据集作为训练集来训练模型，30%的数据集作为测试集来评估模型的性能。使用准确率指标对模型进行评价，准确率是指预测正确的结果占总样本的百分比，是分类问题中最简单最直观的评价指标。本研究对分类准确率最高的模型比较其 ROC 曲线的 AUC 值（ROC 曲线下方的面积大小）来评估模型的性能，AUC 值越高则其分类模型的性能越好。

1.6　决策树视图分析

决策树算法具有可视化的分析效果，使用 R 软件中的 rpart 包对经过二元处理后的 A、B、C 3 个数据集进行视图分析，找出重要的叶节点，从而分析影响母猪最高产的相关因素。

1.7　数据处理

本研究使用 Microsoft Excel 2019 和 R 3.5.3 软件进行数据处理，其中用到的 R 包有 Boruta（特征选择）、rpart（决策树）、randomForest（随机森林）、e1071（支持向量机）及 glm( )函数。

2　结果与分析

2.1　基于重要特征构建母猪高低产的最佳分类模型

按照不同的分类标准将母猪产仔数性状进行二元处理，基于筛选出的重要特征，利用 4 种不同的机器学习方法构建母猪高低产分类模型，比较最佳的分类模型。如表 3 所示，在 A 数据集中所有分类标准下，机器学习方法构建分类模型的分类准确率均在 71%～74%；在 B 数据集中所有分类标准下，机器学习方法构建分类模型的分类准确率均在73%～77%；在 C 数据集的所有分类标准下，机器学习方法分类模型的分类准确率均在 76%～84%。

表 3　不同分类模型的准确率比较

Table 3　Comparison of accuracy of different classification models　%

产仔数性状 Litter size trait	模型 Model	A 数据集 A data set	B 数据集 B data set	C 数据集 C data set
总产仔数 TNB	LOG	73	76	76
	DT	73	75	77
	RF	71	74	76
	SVM	74	75	77
产活仔数 NBA	LOG	72	75	76
	DT	72	74	77
	RF	72	73	79
	SVM	73	75	78
健仔数 NHP	LOG	73	75	80
	DT	72	76	80
	RF	72	75	79
	SVM	73	76	81
5 日龄活仔数 N5D	LOG	71	74	83
	DT	71	74	82
	RF	71	73	83
	SVM	72	74	84
最高产母猪 HY	LOG	72	76	83
	DT	73	77	82
	RF	73	75	83
	SVM	73	77	84

以分类准确率为评价指标，筛选出分类准确率最高的模型，对于最高分类准确率相同的模型，通过比较其ROC曲线的AUC值来确定最佳的分类模型(表4)。在不同数据集和不同分类标准下，最佳的分类模型也不同。结果如表5所示，在不同数据集的不同产仔数性状的最佳模型中，SVM(出现6次)、DT(出现4次)、LOG(出现4次)出现的次数较多，而RF只出现1次。

表4 不同数据集中最高准确性模型的AUC值比较

Table 4 Comparison of AUC values of the highest accuracy models in different data sets

数据集 Data set	产仔数性状 Litter size trait	模型 Model	AUC
A	健仔数 NHP	LOG	0.77
		SVM	0.73
	最高产母猪 HY	DT	0.76
		RF	0.73
		SVM	0.72
B	产活仔数 NBA	LOG	0.79
		SVM	0.74
	健仔数 NHP	DT	0.78
		SVM	0.71
	5日龄活仔数 N5D	LOG	0.77
		DT	0.76
		SVM	0.74
	最高产母猪 HY	DT	0.77
		SVM	0.69
C	总产仔数 TNB	DT	0.82
		SVM	0.79

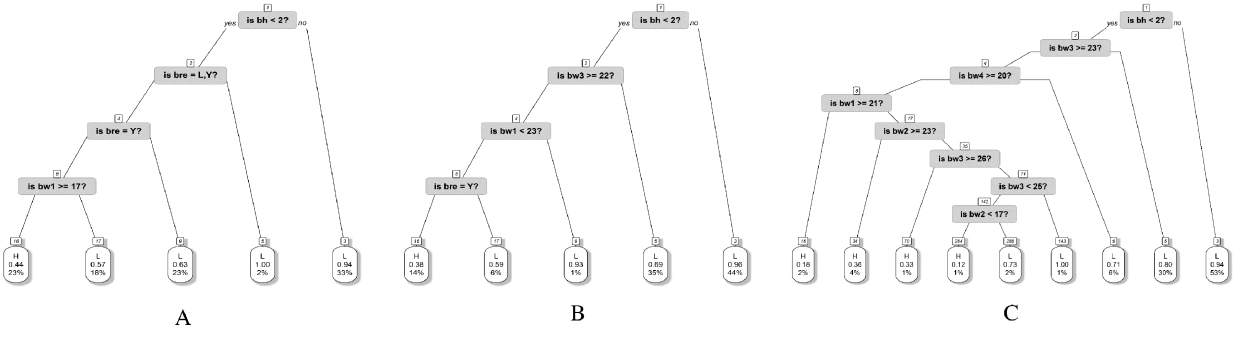
表5 不同分类标准的最佳建模方法

Table 5 The best modeling method of different classification standards

数据集 Data set	总产仔数 TNB	产活仔数 NBA	健仔数 NHP	5日龄活仔数 N5D	最高产母猪 HY
A	SVM	SVM	LOG	SVM	DT
B	LOG	LOG	DT	LOG	DT
C	DT	RF	SVM	SVM	SVM

2.2 决策树视图分析

对A、B、C 3个数据集中的最高产母猪进行决策树视图分析，结果如图4所示。对于A数据集，核心母猪的品种为大白，在1号场生产，第1胎初生窝重大于等于17 kg时其第3胎的产仔数性状较好，结合表3可知，利用决策树模型可推测母猪第3胎有73%的概率产活仔数在17头以上，5日龄产仔数在15头以上(图4A)；对于B数据集，核心母猪在1号场生产，品种为大白，第1胎初生窝重小于23 kg，第3胎初生窝重大于等于22 kg时其第4胎产仔数性状较好，结合表3可知，利用决策树模型可推测母猪第4胎有77%的概率产活仔数在17头及以上，5日龄产仔数在15头以上(图4B)；对于C数据集，核心母猪在1号场生产，第1胎初生窝重大于等于21 kg或第2胎初生窝重大于等于23 kg、第3胎初生窝重大于等于23 kg、第4胎初生窝重大于等于20 kg时其第5胎的产仔数性状较好，结合表3可知，利用决策树模型可推测母猪第5胎有82%的概率产活仔数在17头以上，5日龄产仔数在16头以上(图4C)。



A B C

A:A数据集；B:B数据集；C:C数据集；bre:品种(L:长白猪，Y:大白猪)；bh:出生场地；bw1:第1胎初生窝重；bw2:第2胎初生窝重；bw3:第3胎初生窝重；bw4:第4胎初生窝重；H:最高产母猪；L:低产母猪。A: A data set; B: B data set; C: C data set; bre: Breed (L: Landrace, Y: Yorkshire); bh: Born herd; bw1: Birth weight of first litter; bw2: Birth weight of second litter; bw3: Birth weight of third litter; bw4: Birth weight of fourth litter; H: Highest yield; L: Lowest yield.

图4 最高产母猪的视图分析

Fig.4 View analysis of best sow

3 讨 论

本研究使用 R 软件中的 Boruta 包筛选的重要特征包括出生场地、分娩栏位、品种和不同胎次的初生窝重，如表 3 所示，基于这些特征构建的母猪总产仔数、产活仔数、健仔数和 5 日龄仔猪数的分类模型的准确率均在 71%以上，最高可达到 84%，表明利用机器学习方法构建的母猪高低产分类预测模型具有一定的可靠性。李信颉等[11]比较了 3 种不同的机器学习方法预测生产母猪产仔数性状的性能，发现 SVM 的预测性能要显著优于 KNN 和 DT，这与本研究结果类似。Kirchner 等[16]以母猪总产仔数、产活仔数、健仔数等为预测变量，利用决策树（DT）算法对母猪繁殖力的高低进行预测，其分类准确率在 85%以上，其预测准确性与本研究较为接近。Bakoev 等[6]基于猪生长和肉质特征使用多种不同的机器学习方法对猪的四肢状态进行分类预测，发现随机森林和 K 近邻拥有更好的预测性能，这与本研究的结果有出入，可能是本研究所用到的数据结构和特征不同所致。

决策树是近年来被广泛应用的一种数据挖掘方法，最早被用来挖掘人类社会经济数据中具有价值的数理模型[17]。决策树视图分析方法在畜牧业中的应用研究也较多，如 Monteils 等[18]利用决策树视图分析出了小母牛在生长期间有利于胴体品质的最佳饲养途径，从而更好地指导生产，提高母牛的饲养效率。本研究首次尝试使用决策树视图来分析影响母猪最高产的相关因素，结果发现在 A、B 数据集的最高产母猪的决策树视图中均显示品种是母猪高低产划分的重要叶节点，其中最高产母猪多为大白母猪。这与刘庆伟等[19]研究发现大白猪的产仔数性状要显著高于长白和杜洛克（$P<0.05$）、郭建凤等[20]研究表明大约克猪和长白猪的繁殖性能要显著高于皮特兰和杜洛克（$P<0.05$）的分析结果相吻合。

此外，在不同分类标准下的最优分类模型中，SVM 出现的频次最高且均表现出较高的预测准确性，DT 和 LOG 次之，RF 出现的频次最低（只有 1 次）。Fernandez-delgado 等[21]通过在 121 个 UCI 数据集上进行 179 种分类算法的分类性能比较，发现 RF 的预测性能更好，这与本研究结果有出入。有研究表明随机森林自身不能很好地处理非平衡数据且对于连续性变量处理还需要进行离散化[22-23]，而本研究的 A、B、C 数据集中存在的不同胎次的初生窝重特征恰为连续性变量，这可能是造成此差异的原因。虽然 SVM 模型在不同分类标准及特征下均有较高的分类准确率，但部分 SVM 分类模型的 AUC 值要低于其他的分类模型，且对不同的产仔数性状其最优的机器学习算法也不尽相同。事实上没有哪种单一的分类方法是“最优的”，每种分类算法都有其特定的应用环境，要根据数据结构特点来选择合适的模型[24]。

本研究对已有的生产母猪数据集进行特征筛选，尝试运用 4 种不同的机器学习方法构建母猪高低产分类模型来对下一胎次的高低产进行预测，其预测准确率在 71%以上，最高可达 84%，并利用决策树视图探究了影响母猪高产的相关因素。然而，本研究也存在一定的局限性，如样本量较小、分类模型的预测准确性不高、模型的泛化能力还有待验证、所收集数据包含的变量较少等。在后续的研究中我们将进一步扩充用于构建模型的数据样本量，收集整理更多的变量，例如母猪的发情间隔、公猪的精液品质、母猪的体况和环境数据等，尝试用更科学的算法来构建模型以提高模型分类准确率，使得机器学习方法能够更好地应用于养猪生产，实现高繁殖力母猪的早期选育。

参考文献 References

[1] GADD JOHN，周绪斌，张佳，等.现代养猪生产技术：告诉你猪场盈利的秘诀[M].北京：中国农业出版社，2015. GADD J，ZHOU X B，ZHANG J，et al.Modern pig production technology：tell you the secret to profitability of pig farms [M].Beijing：China Agriculture Press，2015(in Chinese).

[2] SUNDGREN P E，VAN MALE J P，AUMAITRE A，et al.Sow and litter recording procedures.Report of a working party of the E.A.A.P.commission on pig production[J].Livestock production science，1980，7(4)：393-401.

[3] NIELSEN B，SU G，LUND M S，et al.Selection for increased number of piglets at d 5 after farrowing has increased litter size and reduced piglet mortality[J].Journal of animal science，2013，91(6)：2575-2582.

[4] SU G，LUND M S，SORENSEN D.Selection for litter size at day five to improve litter size at weaning and piglet survival rate[J].Journal of animal science，2007，85(6)：1385-1392.

[5] 李航.统计学习方法[M].北京：清华大学出版社，2012. LI H.Statistical learning method [M].Beijing：Tsinghua University

Press,2012(in Chinese).

[6] BAKOEV S,GETMANTSEVA L,KOLOSOVA M,et al.PigLeg:prediction of swine phenotype using machine learning[J/OL].Peer J,2020,8:e8764[2020-09-11].https://doi.org/10.7717/peerj.8764.

[7] MESSAD F,LOUVEAU I,KOFFI B,et al.Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs[J/OL].BMC genomics,2019,20(1):659[2020-09-11].https://doi.org/10.1186/s12864-019-6010-9.

[8] SHAHINFAR S,KAHN L.Machine learning approaches for early prediction of adult wool growth and quality in Australian Merino sheep[J].Computers and electronics in agriculture,2018,148:72-81.

[9] SHAHINFAR S,KELMAN K,KAHN L.Prediction of sheep carcass traits from early-life records using machine learning[J].Computers and electronics in agriculture,2019,156:159-177.

[10] TUSELL L,BERGSMA R,GILBERT H,et al.Machine learning prediction of crossbred pig feed efficiency and growth rate from single nucleotide polymorphisms[J/OL].Frontiers in genetics,2020,11:567818[2020-09-11].https://doi.org/10.3389/fgene.2020.567818.

[11] 李信颉,王海燕,蒋贝加,等.基于机器学习方法预测母猪产仔数性状[J].华中农业大学学报,2020,39(4):63-68.LI X J,WANG H Y,JIANG B J,et al.Prediction of sow litter size trait based on machine learning approaches[J].Journal of Huazhong Agricultural University,2020,39(4):63-68(in Chinese with English abstract).

[12] 高开国,王丽,胡胜兰,等.我国规模化猪场母猪繁殖性能的调查分析[J].中国畜牧杂志,2019,55(9):155-157.GAO K G,WANG L,HU S L,et al.Investigation and analysis of reproductive performance of sows in large-scale pig farms in my country[J].Chinese journal of animal science,2019,55(9):155-157(in Chinese).

[13] KURSA M B,RUDNICKI W R.Feature selection with Boruta package[J].Journal of statistical software,2010,36(11):1-13.

[14] 周志华.机器学习[M].北京:清华大学出版社,2016.ZHOU Z H.Machine learning[M].Beijing:Tsinghua University Press,2016(in Chinese).

[15] WESTREICH D,LESSLER J,FUNK M J.Propensity score estimation:neural networks,support vector machines,decision trees(CART),and meta-classifiers as alternatives to logistic regression[J].Journal of clinical epidemiology,2010,63(8):826-833.

[16] KIRCHNER K,TOLLE K H,KRIETER J.The analysis of simulated sow herd datasets using decision tree technique[J].Computers and electronics in agriculture,2004,42(2):111-127.

[17] FRAWLEY W J,PIATETSKY-SHAPIRO G,MATHEUS C J.Knowledge discovery in databases:anoverview[J].AI magazine,1992,13(3):57-70.

[18] MONTEILS V,SIBRA C.Identification of combinations of influential rearing practices applied during the heifers' whole life on the carcass quality by the decision tree method[J/OL].Livestock science,2019,230:103823[2020-09-11].https://doi.org/10.1021/acs.jafc.7b03239.

[19] 刘庆伟,王春强,刘兴辉,等.不同品种母猪产仔性能的比较研究[J].现代畜牧兽医,2018(10):29-31.LIU Q W,WANG C Q,LIU X H,et al.Comparative study on the litter performance of different breed sows[J].Modern journal of animal husbandry and veterinary medicine,2018(10):29-31(in Chinese with English abstract).

[20] 郭建凤,牛月波,王继英,等.不同品种及产仔季节对母猪繁殖性能影响[J].养猪,2017(1):41-46.GUO J F,NIU Y B,WANG J Y,et al.The effects of different breeds and farrowing seasons on the reproductive performance of sows[J].Swine production,2017(1):41-46(in Chinese).

[21] FERNÁNDEZ-DELGADO M,CERNADAS E,BARRO S,et al.Do we need hundreds of classifiers to solve real world classification problems?[J].Journal of machine learning research,2014,15:3133-3181.

[22] 吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(3):37-41.LÜ H Y,FENG Q.A review of random forests algorithm[J].Journal of the Hebei academy of sciences,2019,36(3):37-41(in Chinese with English abstract).

[23] 黄衍,查伟雄.随机森林与支持向量机分类性能比较[J].软件,2012,33(6):107-110.HUANG Y,ZHA W X.Comparison on classification performance between random forests and support vector machine[J].Software,2012,33(6):107-110(in Chinese).

[24] 王治.基于决策树的多分类器的集成及应用[D].长沙:中南大学,2009.WANG Z.Integration and application of multiple classifiers based on decision tree[D].Changsha:Central South University,2009(in Chinese with English abstract).

Research on sow high and low yield classification model based on machine learning method

LI Xiyang[1], LI Xinjie[1], ZHAO Zhichao[2], LI Changchun[1,2], LIU Xiangdong[1,2]

1.*Key Laboratory of Animal Genetics, Breeding and Reproduction of Ministry of Education/ College of Animal Sciences & Technology, Huazhong Agricultural University, Wuhan* 430070, *China*;

2.*Key Lab of Swine Healthy Breeding of Ministry of Agriculture and Rural Affairs/ Guangxi Yangxiang Co., Ltd., Guigang* 537100, *China*

Abstract In order to help the pig farm managers better carry out the reproductive management of sows, predict the high and low yield sows, and timely eliminate the low yield sows, in this study, we collected and sorted out the dataset of three sow populations, including birth herd, farrow herd, breed and birth weight of different parities, formulated the classification standard of sow high and low yield, and used boruta package in R software to screen out the important characteristics affecting high and low yield of sows. Four different machine learning methods, logistic regression (LOG), decision tree (DT), random forest were (RF) and support vector machine (SVM) were used to construct the classification model of high and low yield sows, and the decision tree view analysis was carried out to explore the related factors affecting the highest yield of sows. The results showed that the classification accuracy of the four machine learning methods for sow high yield classification model was about 71%, and the highest was 84%. It was also found that SVM as the best modeling method appears most frequently across all data sets and different classification criteria, followed by LOG and DT. The decision tree view showed that birth herd, breed and birth weight of different parties were important leaf nodes for dividing the highest yield sows, and these characteristics can be used to predict the most productive sows, with an accuracy of 73%-82%. These results indicated that it will be a good choice to use machine learning method to predict the high and low yields of sows in the future.

Keywords machine learning methods; precision pig raising; early breeding of sow; decision tree; random forest; support vector machine; reproductive performance; early prediction of litter size; high fecundity; classification model

（责任编辑：边书京）