

基于 CNN-RNN 的小麦抗寒性分类模型

来纯晓¹, 李艳翠^{2*}

(1. 河南科技学院信息工程学院, 河南 新乡 453003; 2. 河南师范大学计算机与信息工程学院, 河南 新乡 453007)

摘要:为提升小麦抗寒性分类的准确度,进而为杂交组合选择提供参考。本研究以 3 049 个国审小麦品种文本为试验数据,采用卷积神经网络结合循环神经网络(CNN-RNN)的方法对小麦抗寒性分类进行预测。结果表明,该方法具有较好的表现,准确率可达 73.28%, Kappa 系数为 0.595 6。为降低试验样本不均衡对本研究准确性的干扰,进而采用 SMOTE 技术,以实现样本均衡。过采样后, CNN-RNN 的准确率和 Kappa 系数分别提升 7.67% 和 0.02。说明上述方法组合能够有效提高小麦抗寒预测的准确性以及一致性检验系数,可应用于小麦抗寒分类预测,以达到缩短育种周期的目的。

关键词:小麦; 抗寒性分类; 育种周期; CNN-RNN; SMOTE

中图分类号: TP391

文献标识码: A

文章编号: 2096-5877(2023)04-0117-05

Classification of Cold Resistance of Wheat Based on CNN-RNN Model

LAI Chunxiao¹, LI Yancui^{2*}

(1. College of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003; 2. Henan Institute of Science and Technology, College of Life Science and Technology, Xinxiang 453007, China)

Abstract: To improve the accuracy of cold resistance classification of wheat, which further provide references for hybrid combination selection. The experimental data of 3 049 national wheat varieties were used in this study, and the cold resistance classification of wheat was predicted by using convolutional neural network combined with circulating neural network (CNN-RNN). The results showed that this method has state-of-the-art performances with an accuracy of 73.28% and a Kappa coefficient of 0.5956. Meanwhile, to reduce the interference of experimental sample imbalance in the accuracy of this study, SMOTE technology was used to achieve the sample balance. After oversampling, the accuracy and kappa coefficient of CNN-RNN were improved by 7.67% and 0.02, respectively. It indicated that the combination of the above methods could effectively improve the accuracy and consistency test coefficient of wheat cold-resistance prediction, which could be applied to wheat cold-resistance classification prediction, thereby shorten the traditional breeding cycle.

Key words: Wheat; Cold resistance; Breeding cycle; CNN-RNN; SMOTE

北方小麦种植区在冬春时节的倒春寒、寒潮等极端低温天气发生频率高、影响面积大、危害程度重^[1-2]。小麦品种抗寒性能日益成为影响小麦高产和稳产的重要因素,对小麦抗寒性分类的相关研究具有重要意义。近年来,研究者从不同角度对小麦的抗寒性进行研究,但多数研究集中于生物细胞^[3-4]、基因工程^[5-6]和模拟低温环境^[7]等方面,受技术影响较大,效率偏低。随着人工智

能与农业的深度融合,深度学习方法成功应用到玉米叶部病害识别^[8]、田间杂草分类识别^[9]和作物器官识别^[10]等农业领域,但是利用深度学习对小麦抗寒性的相关研究较少。因此,采用深度学习方法提高对小麦抗寒性分类识别的准确率和效率,对小麦抗寒性育种研究工作的开展具有较大意义。本研究采用 CNN-RNN 融合模型,将深度学习技术^[11-12]引入到小麦抗寒性分类任务中。使用 CNN^[13-14]提取小麦特征词之间的上下文信息,融合 RNN^[15]进一步提取全部输入小麦特征词的序列信息,结合模型提取的特征信息,采用 Softmax 分类器完成小麦抗寒性分类识别。最后通过样本均衡化处理,有效提升模型的分类识别性能。本研究可为小麦抗寒性分类识别提供参考,从而辅

收稿日期: 2020-10-10

基金项目: 河南省科技计划项目(232102210079、212102210431、182102210048)

作者简介: 来纯晓(1991-),男,讲师,硕士,主要从事自然语言处理、农业工程与信息技术研究。

通讯作者: 李艳翠,女,博士,副教授, E-mail: liyancui@htu.edu.cn

助小麦抗寒性育种研究工作。

1 基于 CNN-RNN 的小麦抗寒性分类模型

1.1 数据来源

本研究爬取 1978~2018 年农业农村部种业管理司的中国种业大数据平台的国审小麦品种信息,获取小麦品种数据包含 3 513 个小麦品种的熟性、生育期、苗性、株高和容重等 37 个特征。

通过查询小麦相关资料并结合小麦育种专家的建议,采用人工结合程序的方式清洗数据,剔

除空值属性较多的无效记录,共保留 3 049 条特征较为完整的小麦数据。为了构建抗寒性分类识别模型,选择小麦抗寒性作为数据标签,将数据标签标记为抗寒性强、抗寒性中、抗寒性弱和抗寒性未显 4 种类型。

1.2 CNN-RNN 模型

本研究结合试验数据和任务要求,采用基于卷积神经网络混合循环神经网络的 CNN-RNN 模型进行小麦抗寒性分类研究,模型结构如图 1 所示,主要包括词向量嵌入层、卷积层、池化层、循环神经网络层、全连接层和分类层 6 层。

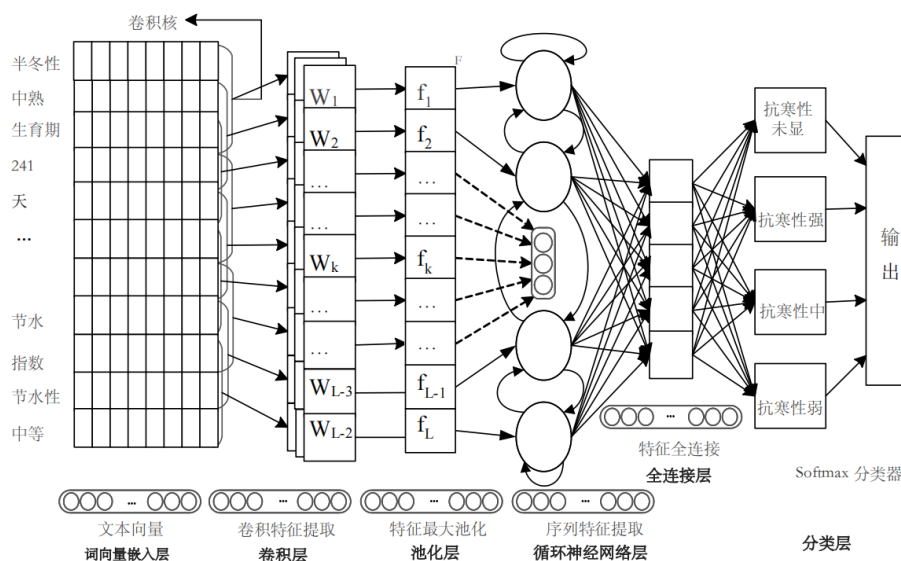


图 1 CNN-RNN 模型结构图

词向量嵌入层,将预处理后小麦品种的特征序列“半冬性/中熟/生育期/241/天/.../节水性/中等”,使用 Word2vec 训练词向量,文本向量表示为 $X \in R^{L \times d}$ 。此时输入某个小麦品种的特征词序列表示为: $X = (x_1, x_2, \dots, x_i, \dots, x_{L-1}, x_L)$, L 表示分词后每个样本的特征词个数, x_i 表示输入样本的第 i 个 ($1 \leq i \leq L$) 特征词的向量表示, d 表示训练的词向量的维度。

卷积层的输入是表示某个小麦样本的矩阵 $X \in R^{L \times d}$, 每一行代表一个特征维度为 d 的样本特征词 x_{ij} 。卷积操作所采用的卷积核向量 $kernel \in R^{m \times d}$, m 表示卷积核的窗口大小。样本 x_i 在第 k 个位置上的特征,可以得到由 m 个连续的小麦特征向量构成的窗口矩阵 W_k ,如式(1)所示:

$$W_k = [x_k, x_{k+1}, \dots, x_{k+m-1}] \dots \dots \dots (1)$$

池化层对卷积提取的特征进行 2 次提取,通过池化下采样,在确保小麦主要特征信息不丢失的基础上,降维减参丢弃非主要的特征。池化层的计算如式(2)所示:

$$X_j^{l+1} = f(\beta_j^{l+1} \text{down}(X_j^l) + b_j^{l+1}) \dots \dots \dots (2)$$

式中 X_j^{l+1} 表示池化后第 j 个小麦样本数据的特征信息, $f(\cdot)$ 为 ReLU 激活函数, $\text{down}(\cdot)$ 为下采样函数, β_j^{l+1} 为乘性偏置, b_j^{l+1} 为加性偏置。

循环神经网络层依据输入长度的变化动态调整自身网络状态,实现对输入特征的序列信息提取。当前隐藏层的状态 s_t 是根据上一个输入 $t-1$ 的隐藏层的状态 s_{t-1} 和当前输入 t 的输入 x_t 计算得到的,隐藏层的状态 s_t 和输出序列信息 h_t 的计算如式(3)和(4)所示:

$$s_t = f(W \otimes s_{t-1}, U \otimes x_t) \dots \dots \dots (3)$$

$$h_t = g(V \otimes s_t) \dots \dots \dots (4)$$

式(3) $f(\cdot)$ 采用 ReLU 为隐藏层的激活函数, $\otimes$ 为矩阵的乘积运算, U 和 W 分别为输入序列信息 X 和隐藏层的状态 S 的权重参数矩阵。式(4) $g(\cdot)$ 为输出的激活函数, V 为输出序列信息的权重参数矩阵。

全连接层中的每个神经元都和前一层所有神经元进行全连接,将前一层提取到的高维小麦特

征信息映射为一维向量。全连接层的计算如式(5)所示:

$$h_{full}(x) = f(W^T \otimes x + b) \quad \dots\dots\dots (5)$$

式中 $h_{full}(x)$ 为神经元的输出值, x 为神经元输入的小麦特征向量, W^T 为权重向量, b 为偏置项, $f(\cdot)$ 使用 ReLU 作为输出层的激活函数。

最后, 使用 Softmax 分类器, 将多个神经元的输出映射到 (0,1) 区间内, 得到所属类别的概率, 输出概率值最大的类别。

1.3 样本均衡化

小麦抗寒性的 4 种标签对应样本个数分别为抗寒性未显 2 030 个、抗寒性好 600 个、抗寒性中 337 个和抗寒性差 82 个。小麦抗寒性较弱的样本数低于其他几种类型, 数据集存在样本不均衡现象。为解决上述问题, 试验采用 SMOTE 随机过采样技术, 合成训练集中样本数目较少的类别, 使各个类别的样本趋于均衡化, 同时尽可能减轻因数据增多而引起模型过拟合的问题。

2 结果与分析

为进行小麦抗寒性分类试验, 将数据按照 4 种抗寒性标签, 以 6:2:2 的比例分为训练集、验证集和测试集。试验使用 Keras 构建 CNN-RNN 模型, 在卷积层中使用 64 个窗口大小为 3 的卷积核, 设定卷积步长为 1。在池化层选择最大池化的方式, 池化窗口大小为 2。在循环神经网络层, 隐藏神经元个数为 64。在全连接层隐藏神经元个数为 32, 激活函数为 Relu。为防止过拟合, 在循环神经网络层和全连接层之间, 添加丢弃率为 0.2 的 Dropout 层。在 Softmax 分类层, 神经元的个数为 4, 激活函数为 Softmax。为防止过拟合, 在试验过程中添加丢弃率 p 为 0.2 的 Dropout 层和 λ 为 0.05 的 L2 正则项惩罚因子。试验选择 Adam 优化器, 学习速率 η 为 0.001。试验的训练批次设定为 32。

2.1 CNN-RNN 试验结果

为了验证模型的有效性, 本研究使用机器学习常用的评估指标: 准确率 (Accuracy)、F1 值 (F1-score) 和 Kappa 系数 (主要用于衡量分类精度)。选取 SVM、CNN、RNN、CNN 融合支持向量机 (CNN-SVM) 模型和本试验所用的卷积神经网络融合循环神经网络 (CNN-RNN) 模型进行比较, 模型训练均采用 Early-stopping 方式监控验证集的 loss 值, 当验证集的 loss 值连续 10 轮没有下降则终止训练。各个模型在测试集上, 依据各项评价指标的性能对比结果如表 1 所示。

表 1 各个模型在测试数据集上的性能对比

模型	准确率 (%)	F1 值 (%)	Kappa 系数
SVM	64.01	41.72	0.389 8
CNN	71.29	49.14	0.526 6
RNN	69.31	47.38	0.499 7
CNN-SVM	68.64	42.21	0.399 7
CNN-RNN	73.28	51.99	0.595 6

从表 1 对比分析可知, 尽管 SVM 模型可以解决非线性、高维数、局部极小点等问题, 但在处理复杂词向量数据时效果欠佳, 正确率和 F1 值仅为 64.01% 和 41.72%, 而 Kappa 系数低至 0.389 8。利用 CNN 提取的上下文特征信息和 RNN 提取的序列特征信息分别训练模型, 结果表明两者性能优于 SVM, 对比 CNN 和 RNN 的预测结果, CNN 的准确率和 Kappa 系数分别达到 71.29% 和 0.526 6, 要比 RNN 在准确率和 Kappa 系数上分别高出 1.98% 和 0.026 9。将卷积提取的特征信息, 使用 SVM 构建分类模型的结果显示, CNN-SVM 模型的准确率、F1 值和 Kappa 系数要比 SVM 高 4.63%、0.49% 和 0.009 9, 说明卷积操作使得特性信息的复杂度降低, 有效改善了 SVM 的性能, 但是也说明 SVM 处理复杂数据的能力不足。本研究的 CNN-RNN 模型在测试集上的正确率、F1 值和 Kappa 系数分别达到 73.28%、51.99% 和 0.595 6, 试验结果表明 CNN 融合 RNN 的模型结构先利用 CNN 提取特性的上下文信息, 之后将提取的信息使用 RNN 提取其隐藏的序列信息可以有效实现小麦抗寒性分类识别。

2.2 样本均衡化后 CNN-RNN 试验结果

训练集样本均衡化前抗寒性未显、较强、适中与较弱的样本数分别为 1 218、360、202 和 49。样本均衡化是对抗寒性适中和较弱的样本进行过采样, 均衡化后 4 种分类的样本数分别为 1 218、360、360 和 360。训练样本均衡化前后的分布如图 2 所示。

从图 2 可以看出, 样本均衡化之后抗寒性适中和较弱的分类样本数增多, 各个分类样本的分布较均衡化前更清晰, 但是也出现多个分类样本重合的现象。这说明在对训练集样本均衡化之后, 模型训练过程中出现较为明显的过拟合现象, 本研究通过增大模型的 L2 正则项因子 $\lambda=0.1$ 和设定 Dropout 层的丢弃率 $p=0.5$ 改善模型的过拟合, 提升模型在测试集上的泛化能力。样本均衡化前后的试验结果对比如图 3 所示。

通过图 3 可以发现, 不论是单一的 SVM、CNN

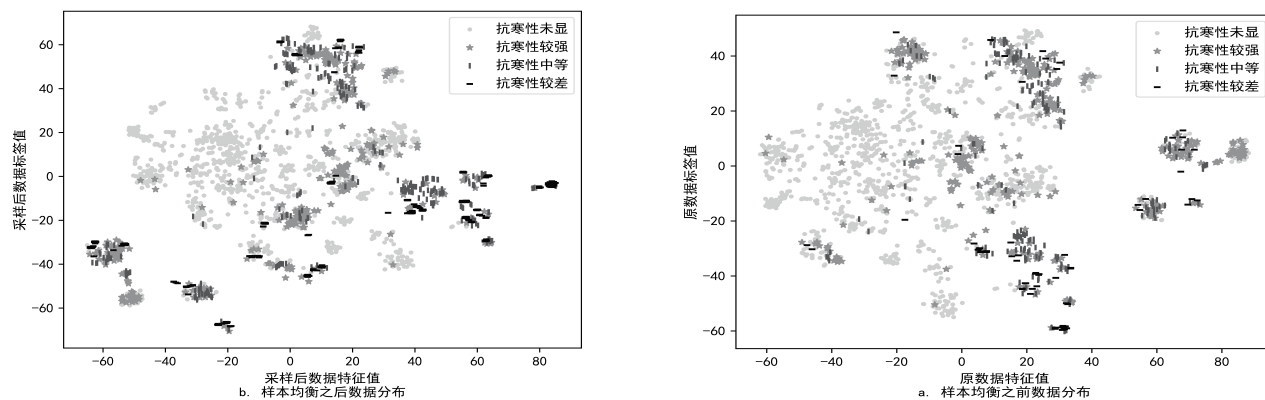


图2 训练样本均衡化前后数据分布

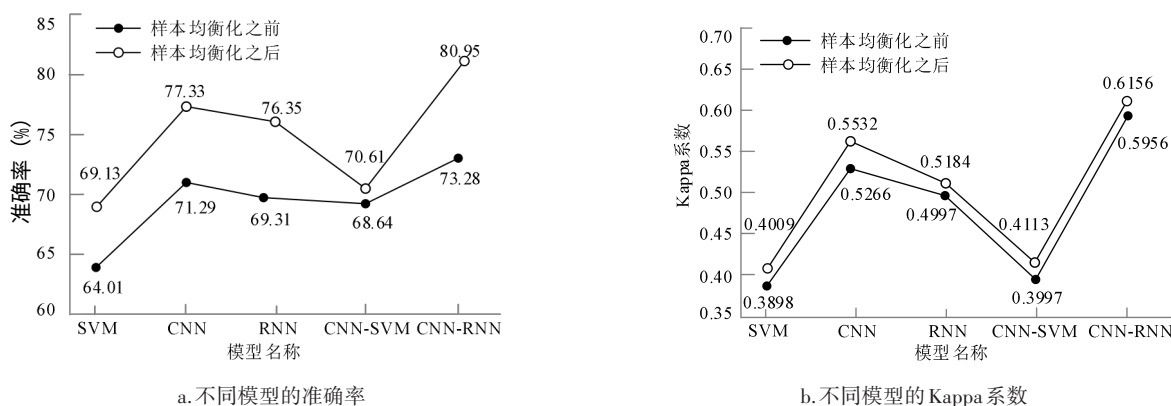


图3 样本均衡化前后实验结果对比

和 RNN, 还是混合的 CNN-SVM 和 CNN-RNN 模型, 其准确率和 Kappa 系数均有提升。样本均衡化之后训练的 CNN 与 RNN 的混合模型在测试集上的性能依然优于其他模型, 模型的准确率为 80.95%, Kappa 系数为 0.615 6, 样本均衡化之后的准确率和 Kappa 系数分别提升 7.67% 和 0.02。这说明对训练集样本均衡化的处理, 有效弥补了因类别标签数量过少而导致该标签被忽略的不足, 同时提高模型预测的准确率和一致性检验程度。

3 讨论与结论

本研究构建 CNN-RNN 模型解决小麦抗寒性分类问题。将样本的特征文本训练为高维的词向量, 充分挖掘各个特征词之间的关系; 对文本向量进行卷积处理, 提取样本特征的局部信息, 并对提取信息进行池化处理, 减少参数个数和网络复杂度; 再使用循环神经网络, 提取卷积池化后特征数据的序列特征信息。经过卷积神经网络联合循环神经网络进行特征提取, 试验结果表明, 本研究模型小麦抗寒性分类的准确度和精确度要高于其他模型。

使用 SMOTE 随机过采样技术可以解决样本

不平衡问题。样本训练数据集均衡化前后的准确率和 Kappa 系数分别提升 7.67% 和 0.02, 此时模型分类的准确度和精确度均有提高。在数据集样本类别分布不平衡时, 通过增加数目较少的标签数量, 可以有效缓解预测标签倒向一边或被忽略的问题, 从而提升模型分类的性能。

深度学习方法在不同的文本分类方面有较好的表现。采用 LSTM 和 CNN 混合模型, 对分类文本的准确率达到 96.45%^[16]。采用 LSTM-TextCNN 联合模型, 对试验短文本的分类准确率为 83.3%^[17]。本研究从自然语言处理角度, 对小麦抗寒性进行分类, 将循环神经网络提取的序列特征信息和卷积神经网络的局部特征信息联合, 进行小麦抗寒性分类识别, 并辅以 SMOTE 技术进行样本均衡化处理, 对少数样本进行扩充。结果表明, 利用 CNN-RNN 方法结合 SMOTE 技术对小麦抗寒性进行分类, 能够有效提升预测准确性, 可应用于小麦育种工作中, 以缩短传统育种周期。

参考文献:

- [1] 赵虹, 王西成, 胡卫国, 等. 黄淮南片麦区小麦倒春寒冻害成因及预防措施[J]. 河南农业科学, 2014, 43(8): 34-38.
- [2] 崔红. 小麦抗寒性研究进展[J]. 吉林农业科学, 2012, 37

- (2):16-19.
- [3] Bohn M, Sabine Luthje, Sperling P, et al. Plasma membrane lipid alterations induced by cold acclimation and abscisic acid treatment of winter wheat seedlings differing in frost resistance[J]. *Journal of Plant Physiology*, 2007, 164(2):146-156.
- [4] 张敏,蔡瑞国,贾秀领,等.小麦抗寒机制的研究进展[J]. *东北农业科学*, 2016, 41(4):37-42.
- [5] 牟永潮,崔红,于晶,等.低温胁迫下东农冬麦1号分蘖节SSH文库的构建及文库中3个基因的表达模式[J]. *作物学报*, 2011, 37(5):918-923.
- [6] Yan L, Fu D M, Li C C, et al. The wheat and barley vernalization gene *Vrn3* is an orthologue of *FT*[J]. *Proceedings of the National Academy of Sciences*, 2006, 103: 19581-19586.
- [7] 赵瑞玲,赵勇,易腾飞,等.小麦种质资源的抗寒性鉴定及品种筛选[J]. *山东农业大学学报(自然科学版)*, 2019, 50(1):25-30.
- [8] 刘丽娟,刘仲鹏,张丽梅.基于图像处理技术的玉米叶部病害识别研究[J]. *吉林农业科学*, 2014, 39(1):61-65.
- [9] 赵静,李志铭,鲁力群,等.基于无人机多光谱遥感图像的玉米田间杂草识别[J]. *中国农业科学*, 2020, 53(8):1545-1555.
- [10] 周云成,许童羽,郑伟,等.基于深度卷积神经网络的番茄主要器官分类识别方法[J]. *农业工程学报*, 2017, 33(15):219-226.
- [11] 吕盛坪,李灯辉,冼荣亨.深度学习在我国农业中的应用研究现状[J]. *计算机工程与应用*, 2019, 55(20):24-33, 51.
- [12] 傅隆生,宋珍珍, Zhang Xin, 等.深度学习方法在农业信息中的研究进展与应用现状[J]. *中国农业大学学报*, 2020, 25(2):105-120.
- [13] 谢为俊,丁冶春,王凤贺,等.基于卷积神经网络的油茶籽完整性识别方法[J]. *农业机械学报*, 2020, 51(7):13-21.
- [14] 赵志衡,宋欢,朱江波,等.基于卷积神经网络的花生籽粒完整性识别算法及应用[J]. *农业工程学报*, 2018, 34(21):195-201.
- [15] 杨祎玥,伏潜,万定生.基于深度循环神经网络的时间序列预测模型[J]. *计算机技术与发展*, 2017, 27(3):35-38, 43.
- [16] 王海涛,宋文,王辉.一种基于LSTM和CNN混合模型的文本分类方法[J]. *小型微型计算机系统*, 2020, 41(6):1163-1168.
- [17] 李志杰,耿朝阳,宋鹏.LSTM-TextCNN联合模型的短文本分类研究[J]. *西安工业大学学报*, 2020, 40(3):299-304.

(责任编辑:王昱)

(上接第101页)

- [13] 廖光联,吕正鑫,朱壹,等.亚热带赣南地区优良软枣猕猴桃“红宝石星”引种栽培研究[J]. *中国南方果树*, 2020, 49(1):120-122.
- [14] 张敏,王贺新,娄鑫,等.世界软枣猕猴桃品种资源特点及育种趋势[J]. *生态学杂志*, 2017(11):3289-3297.
- [15] 胡忠荣.猕猴桃种质资源描述规范和数据标准[M].北京:中国农业出版社,2006:1.
- [16] 张治安,陈展宇.植物生理学实验技术[M].长春:吉林人民出版社,2008:125-130.
- [17] 崔春兰,郑德哲,张力飞,等.苹果疏花剂对果实生长及品质的影响[J]. *吉林农业科学*, 2013, 38(4):75-78.
- [18] Zhishen J, Mengcheng T, Jianming W. The determination of flavonoid contents in mulberry and their scavenging effects on superoxide radicals[J]. *Food Chemistry*, 1999, 64: 555-559.
- [19] Katarzyna S, Rafal S, Krystyna M J. Involvement of melatonin applied to *Vigna Radiata* L. seeds in plant response to chilling stress[J]. *Central European Journal of Biology*, 2014, 9(11):1117-1126.
- [20] 钟金仙,罗英,曾仁杰,等.黄瓜种质资源果实性状的主成分分析与综合评价[J]. *中国农学通报*, 2017, 33(16):46-52.
- [21] 牟书靓,李玉发,牛海龙,等.基于主成分分析的花生种质资源苗期抗旱性鉴定与筛选[J]. *吉林农业科学*, 2015, 40(6):26-30, 69.
- [22] 周米生,肖正东,季琳琳,等.长山核桃优株坚果性状主成分分析及综合评价[J]. *安徽林业科技*, 2016, 42(6):9-12.
- [23] 刘胤,陈涛,张静,等.中国樱桃地方种质资源表型性状遗传多样性分析[J]. *园艺学报*, 2016, 43(11):2119-2132.
- [24] 郁香荷,章秋平,刘威生,等.中国李种质资源形态性状和农艺性状的遗传多样性分析[J]. *植物遗传资源学报*, 2011, 12(3):402-407.
- [25] 陈秀萍,黄爱萍,蒋际谋,等.枇杷种质资源数量分类研究[J]. *园艺学报*, 2011, 38(4):644-656.
- [26] 林蝉蝉,何舟阳,单文龙,等.基于主成分与聚类分析综合评价杨凌地区红色鲜食葡萄果实品质[J]. *果树学报*, 2020, 37(4):520-532.
- [27] 秦红艳,许培磊,艾军,等.软枣猕猴桃种质资源果实品质、表型性状多样性及主成分分析[J]. *中国农学通报*, 2015, 31(1):160-165.
- [28] 李旭,曹万万,姜丹,等.长白山野生软枣猕猴桃资源分布与果实和叶片性状多样性[J]. *北方园艺*, 2015(15):22-27.
- [29] 仇占南,张茹阳,彭明朗,等.北京野生软枣猕猴桃果实品质综合评价体系[J]. *中国农业大学学报*, 2017, 22(2):45-53.
- [30] 刘香苏,黄圆博,刘铭,等.长白山野生软枣猕猴桃果实多样性和倍数性的调查分析[J]. *北方园艺*, 2017(19):56-60.
- [31] 黄春辉,廖光联,谢敏,等.不同猕猴桃品种果实发育过程中总酚和类黄酮含量及抗氧化活性的动态变化[J]. *果树学报*, 2019, 36(2):174-184.
- [32] Zuo L L, Wang Z Y, Fan Z L, et al. Evaluation of antioxidant and antiproliferative properties of three *Actinidia* (*Actinidia kolomikta*, *Actinidia arguta*, *Actinidia chinensis*) extracts in vitro[J]. *International Journal of Molecular Sciences*, 2012, 13(5):5506-5518.
- [33] Wang Y, Li J Y, Zhao C L, et al. Evaluation of biochemical components and antioxidant capacity of different kiwifruit (*Actinidia* spp.) genotypes grown in China[J]. *Biotechnology & Biochemical Equipment*, 2018, 32(12):1-8.

(责任编辑:王昱)