食品工业科技
SCIENCE AND TECHNOLOGY OF FOOD INDUSTRY
• 荷兰《文摘与引文数据库》（Scopus）
• 瑞典开放存取期刊目录 DOAJ
• 美国《化学文摘》CA
• 英国《食品科技文摘》FSTA
• 日本科学技术振兴机构数据库JST
• 世界期刊影响力指数（WJCI）报告
• 食品科学与工程领域高质量科技期刊分级目录第一方阵T1
• 北大核心期刊
• 中国精品科技期刊
• 中国科技核心期刊CSTPCD
• 中国核心学术期刊RCCSE
• 中国农林核心期刊A
• 中国生物医学Sino Med
半月刊 ISSN 1002-0306 CN 11-1759/TS 邮发代号：2-399

基于自然语言处理的山楂果实品种近红外无损鉴别方法

邓志扬，廖 强，邵淑娟，刘 军

Nondestructive Near-infrared Identification of Hawthorn Fruit Cultivars Based on Natural Language Processing

DENG Zhiyang, LIAO Qiang, SHAO Shujuan, and LIU Jun

在线阅读 View online: https://doi.org/10.13386/j.issn1002-0306.2023010132

邓志扬, 廖强, 邵淑娟, 等. 基于自然语言处理的山楂果实品种近红外无损鉴别方法 [J]. 食品工业科技, 2023, 44(22): 249−256. doi: 10.13386/j.issn1002-0306.2023010132
DENG Zhiyang, LIAO Qiang, SHAO Shujuan, et al. Nondestructive Near-infrared Identification of Hawthorn Fruit Cultivars Based on Natural Language Processing[J]. Science and Technology of Food Industry, 2023, 44(22): 249−256. (in Chinese with English abstract). doi: 10.13386/j.issn1002-0306.2023010132

· 分析检测 ·

基于自然语言处理的山楂果实品种近红外无损鉴别方法

邓志扬[1], 廖　强[1], 邵淑娟[2], 刘　军[1,*]
（1.中国农业大学食品科学与营养工程学院, 北京 100083;
2.菏泽市食品药品检验检测研究院, 山东菏泽 274000）

摘　要: 不同品种的山楂果实在营养组成、感官品质等方面存在差异，在工业生产中适用不同的加工方式。传统的检测方法耗时长、具有破坏性以及成本高，为适应规模化生产山楂果实制品的需要，需对山楂果实品种进行无损鉴别。研究共收集了 4 个品种 240 个山楂果实样本的近红外光谱数据，采用不同的预处理算法处理光谱数据后，使用自然语言处理（Natural Language Processing，NLP）模型进行分析，以实现山楂果实品种的无损鉴别。结果表明，长短期记忆网络（Long Short-Term Memory，LSTM）以及门控循环单元（Gated Recurrent Unit，GRU）神经网络模型对主成分分析法（Principal Component Analysis，PCA）预处理后的光谱的鉴别准确率高，验证集的准确率均为 99.46%±0.00%，测试集的准确率均为 100%±0.00%。逻辑回归模型对山楂果实光谱鉴别能力优异，除对二阶差分（Difference Of Second Order，D2）预处理的光谱鉴别能力较差外（验证集准确率 96.65%，测试集准确率 89.58%），其他预处理方式验证集、测试集的准确率均达到或极接近 100%。朴素贝叶斯模型对经 PCA 处理后的光谱的鉴别效果较优，验证集准确率为 95.65%，测试集准确率为 95.83%。本研究证实了 NLP 运用于山楂果实近红外无损鉴别是可行的。

关键词: 自然语言处理, 机器学习, 山楂果实, 近红外, 无损检测
中图分类号: TS207.3　　**文献标识码:** A　　**文章编号:** 1002−0306(2023)22−0249−08
DOI: 10.13386/j.issn1002-0306.2023010132
本文网刊:

Nondestructive Near-infrared Identification of Hawthorn Fruit Cultivars Based on Natural Language Processing

DENG Zhiyang[1], LIAO Qiang[1], SHAO Shujuan[2], LIU Jun[1,*]
（1.College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, China;
2.Heze City of Food and Drug Inspection and Testing Institute, Heze 274000, China）

Abstract: Hawthorn fruits of different varieties have varied nutritional composition, sensory properties etc., thus required for different processing for product development. Due to the limitations of traditional analytical methods of time-consuming, destructive sample preparation, and high cost ect., non-destructive techniques for variety identification are needed which would benefit for large scale production of foods with hawthorn fruits. In this study, a total of 240 hawthorn fruit samples from four different varieties were subjected for near-infrared spectroscopy analysis and the collected spectral data were pre-processed by different algorithms. In order to achieve non-destructive identification of hawthorn varieties, natural language processing (NLP) model was applied for data analysis, including long short-term memory (LSTM), gated recurrent unit (GRU) neural network, logistic regression, native Bayes, decision trees, and k-nearest neighbors. The results showed that the two deep learning models both had the best discrimination effect on the spectral preprocessed by principal component analysis (PCA) with the accuracy of the validation set and test set reached 99.46%±0.00% and 100%±0.00%.

收稿日期: 2023−02−01
作者简介: 邓志扬（1999−）, 男, 硕士, 研究方向: 生物信息学, E-mail: dzycau@163.com。
*** 通信作者:** 刘军（1986−）, 男, 博士, 副教授, 研究方向: 食品生物技术, E-mail: junliu@cau.edu.cn。

While, the logistic regression model showed excellent discrimination ability for hawthorn fruit spectra but poor discrimination ability for the difference of second order (D2) pretreatment spectra (accuracy of 96.65% in the validation set and 89.58% in the test set). The naive Bayes model also showed excellent discrimination effect on the spectra processed by PCA, and the accuracy of the validation set was 95.65%, and the accuracy of the test set was 95.83%. Results gained in this study confirmed the feasibility of applying NLP to the near-infrared non-destructive identification of hawthorn fruits.

Key words: natural language processing; machine learning; hawthorn fruit; near infrared spectroscopy; nondestructive identification

山楂(*Crataegus pinnatifida* Bunge)在我国具有悠久的药用以及食用历史,山楂果实及其制品深受消费者欢迎[1]。我国的山楂品种资源丰富,据不完全统计约有 500 余份,经《中国果树志·山楂卷》核实收载的代表性品种资源有 142 份[2]。不同品种的山楂果实往往在感官品质、营养成分等方面存在差异,适合不同的加工食用方式。例如,昌黎紫肉山楂果实大而整齐、果肉紫红、味酸微甜,适宜鲜食;敞口山楂果实常加工制成山楂片,出片率高且质量好[2]。因此,在加工前有必要对山楂果实品种进行鉴别,以适应不同加工食用方式的要求。传统的农产品鉴别分类主要依赖感官品评或者理化鉴定,感官品评受主观影响较大,而理化鉴定则步骤繁琐且成本高[3]。近红外光谱检测技术具有无损、快速、高效、操作简便等特点[4]。近红外光谱在农产品检测中具有广泛应用,如产地鉴别[5],营养成分定量分析[6−8],霉变鉴定[9]等。

近红外光谱数据包含信息复杂,解析困难。近红外光谱主要采集 C-H、O-H、N-H 等含氢基团的化学键伸缩振动的倍频或合频吸收所反映的光谱信息,该区域谱峰较宽且重叠严重,加之吸收强度低,因此难以得到分子中官能团的特征吸收峰[10]。合适的数据处理方法可有效分析光谱信息,构建准确率较高的预测模型。常用于鉴别农产品品种的模型有偏最小二乘判别分析法(Partial Least Squares Discriminant Analysis, PLS-DA)、支持向量机(Support Vector Machine, SVM)和最小二乘支持向量机(Least Squares-Support Vector Machines, LS-SVM)等[11]。

自然语言处理(Natural Language Processing, NLP)的主要对象具有序列特性,如文本信息是文字按照语法规则的逻辑顺序排列;语音信息是单位时间的音频信号按照时间顺序排列构成的。NLP 模型大多有较强的序列信息处理能力[12]。近红外光谱数据亦是一种序列数据,是按照波长或波数的大小,将吸光度按顺序排列构成,因此,可考虑将 NLP 运用到近红外光谱数据解析。目前已有研究者将 NLP 运用到农产品的无损检测中并取得了良好的效果,如长短期记忆网络(Long Short-Term Memory, LSTM)、门控循环单元(Gated Recurrent Unit, GRU)神经网络、时间卷积网络(Temporal Convolutional Network, TCN)模型可根据草莓酱的中红外光谱数据实现对草莓酱掺假的鉴别[13];卷积神经网络(Convolutional Neural Networks, CNN)、长短期记忆网络(Long Short-Term Memory, LSTM)以及 CNN-LSTM 模型可分析近红外高光谱数据,实现对新鲜茶叶中掺入陈年茶叶的鉴别[14]。

本研究将 NLP 应用于山楂果实的近红外光谱数据解析,实现对山楂果实品种的无损鉴别。共采集了 4 个品种 240 个山楂果实样本的近红外光谱,训练模型,检验模型鉴别的准确率,旨在为基于近红外光谱的农产品无损鉴别分析提供参考。

1 材料与方法

1.1 材料与仪器

山里红大果山楂果实　产自吉林四平;五棱大果山楂果实　产自山东烟台;棉球大果山楂果实　产自山东临沂;甜红子樱桃山楂果实　产自山东临沂。

Antaris II 型傅立叶变换近红外光谱仪　赛默飞世尔(上海)仪器有限公司。

1.2 实验方法

1.2.1 样品预处理与近红外光谱数据采集　对收集所得的不同品种山楂果实进行随机取样(n=60),山楂果实清水洗净后擦去果实表面水分,将果实放置于近红外光谱仪的光源中央,确保光源平行于山楂果实的赤道面照射,采集山楂果实的近红外光谱数据。光谱采集参数为:分辨率 4 cm^{-1};扫描信号次数 32 次;扫描范围 10000~4000 cm^{-1}。每次采集后将山楂果实以果柄为轴线旋转 120°,每个山楂果实样本采集三个不同角度的光谱数据,取对应波数吸光度的平均值作为该样本的近红外光谱。

1.2.2 光谱数据的预处理　为提升模型的准确性,使用主成分分析结合马氏距离法剔除异常光谱[15],使用主成分分析法(Principal Component Analysis, PCA)、SG 滤波法(Savitzky-Golay, SG)、一阶差分(Difference of First Order, D1)、二阶差分(Difference of Second Order, D2)对山楂果实的近红外光谱数据进行预处理,以提升模型的分类效果。

1.2.3 深度学习模型的搭建

1.2.3.1 长短期记忆网络　LSTM 是由循环神经网络(Recurrent Neural Network, RNN)改进而来。RNN 广泛应用于时间序列信息的处理,但其在训练中存在梯度消失的问题,即某一时刻的梯度无法很久地影响结果[16]。而 LSTM 可通过添加的“遗忘机制”使得网络对长序列信息的记忆更好,可有效解决长序列训练过程中的梯度消失和梯度爆炸问题[16]。LSTM 网络的单元结构如图 1 所示。

其中 x_t 为当前时刻的输入数据,上一时刻的存

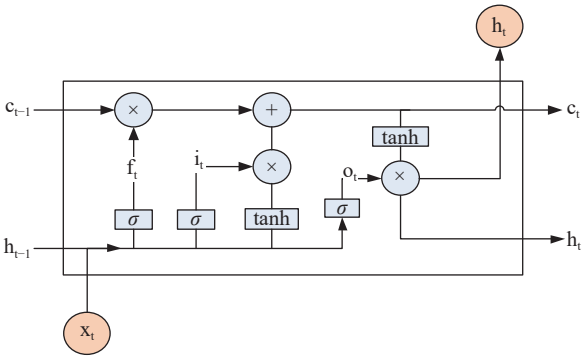

图 1　LSTM 单元结构[17]

Fig.1　Structure of LSTM unit[17]

储单元信息 c_{t-1} 以及上一时刻的隐藏层信息 h_{t-1} 也作为 t 时刻的输入；当前时刻的存储单元信息 c_t 以及隐藏层信息 h_t 为 t 时刻的输出。i_t 为输入门，f_t 为遗忘门，o_t 为输出门，通过遗忘门可以选择性地记忆信息，从而对长序列信息有更好的记忆效果[12]；激活函数 tanh 可将实数输入映射到 [−1,1] 范围内[18]，σ 表示 sigmoid 激活函数，可将实数输入映射到 [0,1] 范围，激活函数的作用为加入非线性因素，提高神经网络解决非线性问题的能力[19]。

本研究的山楂果实样本数为 240 个，采集的山楂果实的近红外光谱数据序列较长，每个样本的近红外光谱数据包含 1556 个波数下的吸光度，即为 1556 维的向量，为序列数据，因此采用 LSTM 网络模型对其进行分析。本研究中搭建的 LSTM 模型结构如图 2 所示。

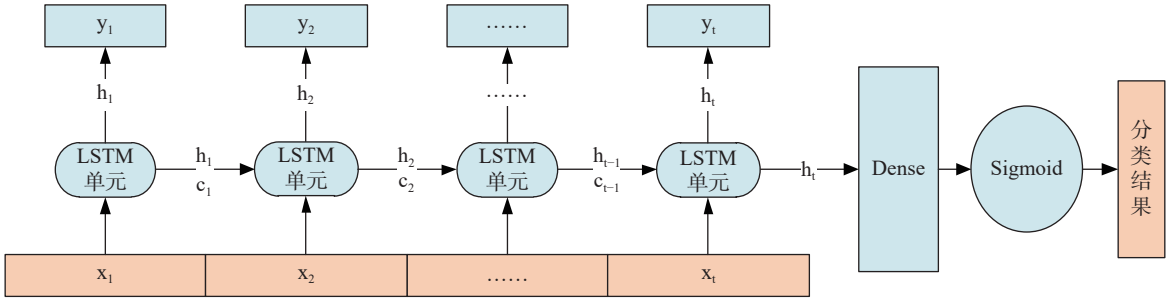

图 2　LSTM 网络模型

Fig.2　LSTM network model

每个样本的近红外光谱数据按照时间步长被分成若干个向量(x_1, x_2, x_3.....x_t)后按顺序输入 LSTM 单元中，最后输入全连接层 Dense，再经过 Sigmoid 函数计算后获得分类结果。

1.2.3.2　门控循环单元网络　GRU 是 LSTM 单元结构的一种变体，是将 LSTM 单元结构的输入门和遗忘门合并为更新门(Z_t)，输出门改为重置门(r_t)(图 3)[20]。因此，GRU 相较于 LSTM 单元结构简单，参数更少，更便于训练。图 3 中 X_t 为本时刻输入的向量，h_{t-1} 为上一时刻的输出，h_t 为本时刻的输出[21]。用 GRU 代替图 2 中的 LSTM 单元，可构成 GRU 神经网络模型。

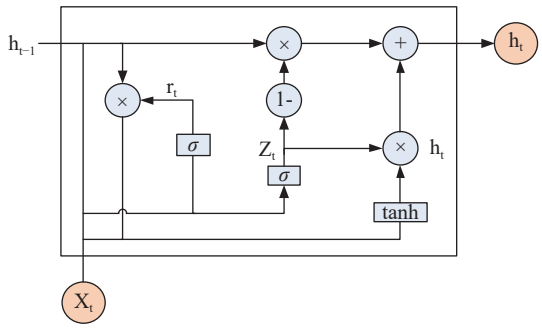

图 3　门控循环单元[21]

Fig.3　Gated recurrent unit[21]

1.2.4　传统机器学习模型建立　本研究采用逻辑回归、朴素贝叶斯、决策树、K 近邻算法四种 NLP 常用的传统机器学习模型对山楂果实的近红外光谱数据进行分析。

1.2.4.1　逻辑回归模型　逻辑回归模型[22] 的数学表达式如下所示：

$$\hat{y} = h(x) = g(w_1x_1 + w_2x_2 + \cdots + w_nx_n + b)$$

其中，x 是自变量，w 是参数，$\hat{y}$是估计值。本研究将山楂果实的近红外光谱特征作为多维自变量 x，品种作为因变量 y，建立逻辑回归模型预测山楂果实的品种。

1.2.4.2　朴素贝叶斯模型　采用朴素贝叶斯模型对山楂果实样本的近红外光谱数据进行分析，以期实现山楂果实品种分类。设 C 为山楂果实品种的集合，n 为品种数，则集合为 $C=\{c_1,c_2,c_3.....c_n\}$。x 为某一待分类山楂果实样本的光谱特征集合 $x=\{a_1,a_2,a_3.....a_m\}$，m 为光谱的特征数，依据贝叶斯定理，计算每个山楂果实品种对于该待分类山楂果实样本的光谱特征集合 x 的条件概率 $P(c_j|a_1,a_2.....a_m)$，其中 j=1,2,.....m，条件概率中最大的一项的类即为待分类山楂果实样本所属的品种[23]。

1.2.4.3　决策树模型　采用决策树模型从根节点出发对待分析山楂果实样本的近红外光谱的一个特征进行判断，根据判断的结果分配到子节点中，进而对山楂果实样本的近红外光谱的下一个特征进行判断分类，如此循环，直到将最后一个特征分配到带有山楂果实品种标签的叶子节点中，实现山楂果实品种的分类[24]。

1.2.4.4 K 近邻算法 K 近邻算法根据距离函数计算待分类的山楂果实样本近红外光谱 X 与训练集中每个山楂果实样本的近红外光谱之间的距离，选择与待分类山楂果实样本距离最小的 K 个样本作为 X 的 K 个最近邻，最后依据 X 的近邻中的大多数样本的类别作为 X 的类别[25]。

1.2.5 数据集划分与模型评价验证 将数据集按照训练集∶验证集∶测试集=6∶2∶2 划分，训练集用于模型的拟合调试，验证集用于模型超参数的调整，测试集不参与模型的调试只用于检验模型的预测能力。采用外部验证法以验证集和测试集预测的准确率来评价模型的预测能力[26]。

1.3 数据处理

在 Jupyter Notebook 6.0.1 开发环境下，利用 Python 3.7.0 对近红外光谱数据进行分析建模，深度学习框架采用 Keras 2.3.1，机器学习库采用 Scikit-learn 0.21.3，异常光谱检验采用 SciPy 1.3.1。

2 结果与分析

2.1 山楂果实的近红外光谱

本研究采用 PCA 结合马氏距离法进行异常光谱的检测与剔除。共从光谱样本中剔除异常样本 9 个，其中甜红子樱桃、山里红大果样本各剔除 3 个，棉球大果样本剔除 1 个，五棱大果样本剔除 2 个。图 4 为剔除异常光谱后的山楂果实的近红外光谱图，由图 4 可知在 5200 cm^{-1} 附近有吸收峰，可能与 C-H 和 C=O 伸缩振动的合频有关[10]；7000 cm^{-1} 附近的吸收峰可能与水中 O-H 的一级倍频有关[27]。该光谱图与 Dong 等[27] 收集的山楂果实的近红外光谱图形状相近。由于山楂果实的近红外光谱吸收峰范围相近，形状相似，难以直接区分品种。

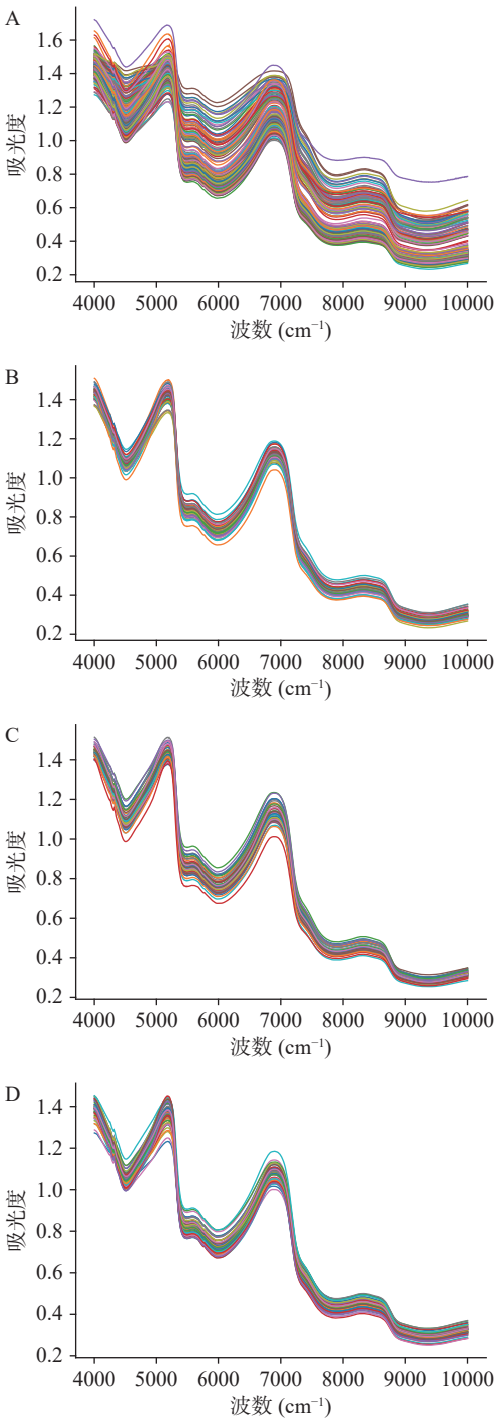

图 4 不同品种山楂果实样本的近红外光谱图

Fig.4 Near-infrared spectra of four hawthorn species

注：A：棉球大果光谱；B：山里红大果光谱；C：五棱大果光谱；D：甜红子樱桃光谱。

2.2 模型构建

2.2.1 深度学习模型训练 利用四个品种山楂果实的近红外光谱数据训练 LSTM 与 GRU 神经网络模型进行品种鉴别，优化后的两种深度学习模型的训练参数与训练结果如表 1 所示。由表 1 可知，LSTM 与 GRU 神经网络模型在训练集中的准确率分别为 98.30%±0.46% 和 97.87%±0.46%，在验证集中的准确率分别为 95.47%±0.83% 与 96.01%±0.63%。

采用混淆矩阵对验证集预测结果进行可视化处理，进一步分析深度学习模型对山楂果实品种鉴别的准确率。如图 5 所示，LSTM 与 GRU 神经网络模型对棉球大果鉴别的准确率均较高，均为 100%。LSTM 对甜红子樱桃、五棱大果品种的鉴别能力较差，准确率低于 90%。GRU 神经网络模型对四种山楂果实品种的鉴别准确率较为稳定，均在 90% 以上。本研究中 GRU 神经网络模型的准确率略优于 LSTM 模型，原因推测为数据集规模较小，GRU 神经网络模型在较小规模的数据集中的性能往往优于 LSTM[28]。

表 1 LSTM 与 GRU 神经网络模型的训练参数与结果

Table 1 Training parameters and results of LSTM and GRU neural network models

模型名称	训练批次	训练轮次	时间步长	优化器	Dropout	训练集准确率(%)	验证集准确率(%)
LSTM	50	1300	4	Adam	0.00	98.30±0.46	95.47±0.83
GRU神经网络	50	1300	4	Adam	0.00	97.87±0.46	96.01±0.63

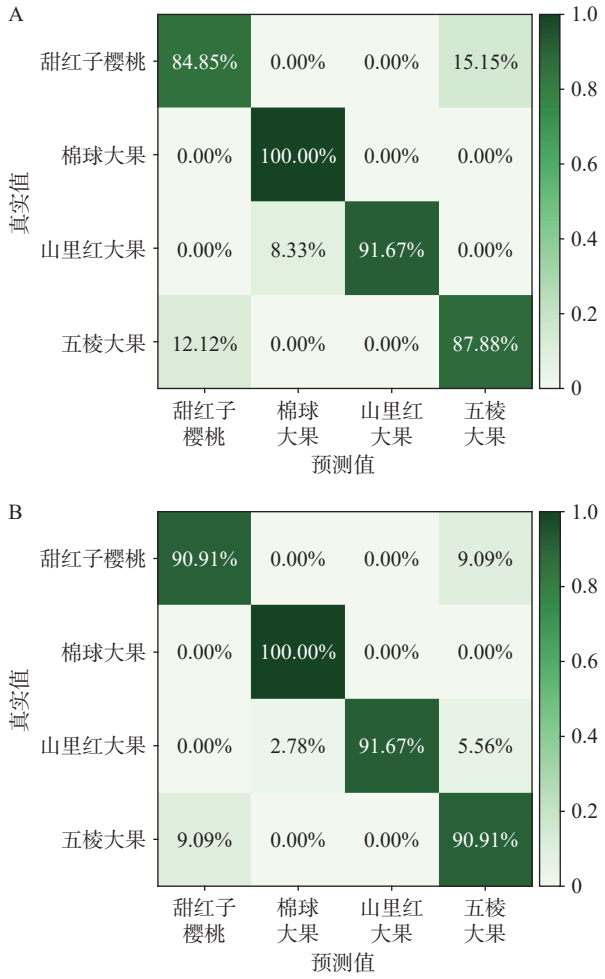

图 5　LSTM 与 GRU 神经网络模型验证集混淆矩阵

Fig.5　Confusion matrix of validation sets for LSTM and GRU neural network models

注：A：LSTM 模型；B：GRU 神经网络模型。

2.2.2　不同预处理方法对深度学习模型训练结果的影响　对光谱进行预处理往往可以提升近红外分析模型的准确性，因此，本研究尝试在训练深度学习模型之前对近红外光谱数据进行预处理。光谱进行预处理后，由于数据发生了变化，深度学习模型的参数需要优化调整才能得到较优的效果。优化后的参数如表 2 所示，深度学习模型在验证集中的准确率如图 6 所示。

表 2　深度学习模型的训练参数

Table 2　Training parameters of the deep learning model

预处理方法与模型	训练批次	训练轮次	时间步长	优化器	Dropout	输出维度
LSTM	50	1300	4	Adam	0	389
GRU神经网络	50	1300	4	Adam	0	389
PCA+LSTM	50	500	2	Adam	0	50
PCA+GRU神经网络	50	500	2	Adam	0	50
D1+LSTM	50	1000	1	Adam	0	311
D1+GRU神经网络	50	1000	1	Adam	0	311
D2+LSTM	50	1000	3	Adam	0.15	311
D2+GRU神经网络	50	1000	3	Adam	0.15	311
SG+LSTM	50	1000	4	Adam	0	389
SG+GRU神经网络	50	1000	4	Adam	0	389

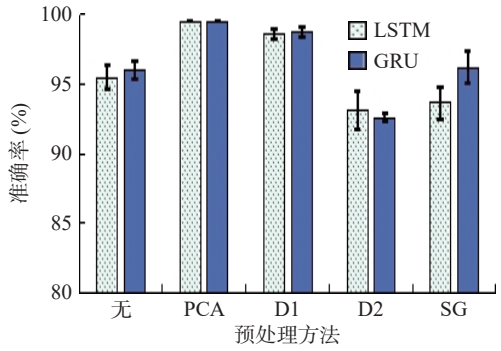

图 6　不同预处理方法对深度学习模型验证集准确率的影响

Fig.6　Influence of different preprocessing methods on the validation set accuracy of deep learning models

在未预处理的条件下，LSTM 与 GRU 神经网络模型的山楂果实品种鉴别的准确率分别为 95.47%±0.83% 与 96.01%±0.63%，准确率高，表明两种 NLP 常用的深度学习模型对序列数据特征提取能力强[29]，即使在无预处理情况下，也可以充分提取分析不同品种山楂果实近红外光谱数据特征。因此，LSTM 与 GRU 神经网络模型不但对文本数据有强的分析能力，对光谱序列数据也有较好的分析效果。经过 PCA 预处理后，两种模型的准确率提升至 99.46%±0.00%，可见 PCA 预处理进一步提升 LSTM 与 GRU 神经网络模型对不同品种山楂果实光谱特征的提取能力。利用 D1 预处理后，两种模型的山楂果实品种鉴别的准确率分别提升至 98.55%±0.31%、98.73%±0.31%，较之于 PCA 略低。利用 D2 预处理后，两种模型的准确率下降，分别降为 93.12%±1.37%、92.57%±0.31%。利用 SG 预处理后，LSTM 模型的准确率下降，GRU 神经网络模型的准确率略有提升，准确率分别为 93.66%±1.13%、96.20%±1.09%。光谱经预处理后，准确率下降的可能原因为，SG、D2 预处理虽降低了噪声，但影响了模型对近红外光谱数据的特征提取能力，导致模型对山楂果实品种鉴别的准确率降低[30]。

2.2.3　传统机器学习模型的构建　传统机器学习模型在验证集中的准确率如图 7 所示，常用于文本分类问题的逻辑回归模型在本研究中的准确率最高，其在无预处理条件下，采用 PCA 或 SG 算法预处理光谱数据后准确率均为 100%。

朴素贝叶斯模型在无预处理的条件下对山楂果实品种鉴别的准确率仅为 76.09%，推测原因为朴素贝叶斯模型的假设条件是特征之间相互独立[31]，而山楂果实近红外光谱各波长的吸光度数据之间存在着多重共线性问题，即具有较强的相关性[32]，不满足朴素贝叶斯模型的假设条件。光谱数据经过 PCA 预处理后，朴素贝叶斯模型在验证集中的准确率提升至 95.65%。PCA 预处理可将光谱特征降维，组成若干相互独立的、新的一组特征，符合朴素贝叶斯模型成立的假设条件[33]。而 D1 预处理近红外光谱数据后，朴素贝叶斯模型的准确率提升至 89.13%。D1 预

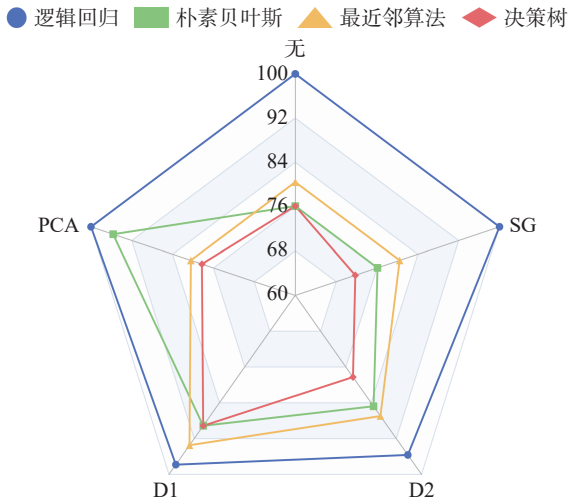

图 7 预处理方法对机器学习模型验证集准确率的影响

Fig.7 Influence of preprocessing method on accuracy of machine learning model validation set

处理可消除近红外光谱基线漂移、平缓背景干扰，提供比原光谱更高的分辨率和更清晰的光谱轮廓变化信息，因而可提升朴素贝叶斯模型的准确率[34]。决策树模型和 K 近邻算法在光谱数据无预处理的情况下在验证集准确率仅为 75% 左右，而经合适的预处理后，其品种鉴别的准确率显著提升。其中，决策树模型对经 D1 处理后的光谱鉴别的准确率可达 89.13%；K 近邻算法对经 D1 处理后的光谱鉴别的准确率可达 93.48%。

综上，NLP 常用的传统机器学习模型可用于解决基于近红外光谱的山楂果实品种鉴别问题，而 PCA 及 D1 预处理方法可有效提升传统机器学习模型的准确率。

2.3 模型预测能力的对比与验证

PLS-DA、SVM 是两种基于近红外光谱的农产品品种鉴别常用模型[11]。分别采用这两种模型对不同品种山楂果实的近红外光谱数据进行建模分析并与本研究中使用的 NLP 模型的品种鉴别准确率进行对比分析，采用外部验证法对构建模型的预测能力进行验证[26]。

如表 3 所示，在验证集中，PLS-DA 模型准确率高（无预处理以及使用不同预处理方法，准确率均为 100%）。SVM 模型准确率低（准确率均低于 80%）。本研究中的 NLP 模型在合适的预处理条件下准确率均高于 85%，优于 SVM 模型。LSTM、GRU 神经网络、逻辑回归、朴素贝叶斯模型在光谱数据经合适的预处理后准确率均可达到或接近 PLS-DA 模型的准确率。

在测试集中，PLS-DA 模型的准确率均为 100%。SVM 模型的准确率均在 75% 以下。PCA 预处理结合 LSTM 与 PCA 预处理结合 GRU 神经网络模型在测试集中的准确率为 100%±0.00%，说明模型具有很强的预测能力，与 PLS-DA 模型的准确率相当，且 LSTM 与 GRU 神经网络模型在光谱经其他方式预处理的条件下准确率也均高于 95%。逻辑回归模型在光谱无预处理、D1、PCA 或 SG 预处理条件下，准确率均为 100%，与 PLS-DA 模型的准确率相当。PCA 预处理结合朴素贝叶斯模型的准确率为 95.83%。决策树和最近邻算法在测试集中的表现相对较差，D1 预处理后模型的准确率均为 91.67%。

表 3 模型预测能力的对比与验证

Table 3 Comparison and verification results of the prediction ability of the model

模型种类	预处理方法	验证集准确率(%)	测试集准确率(%)
PLS-DA	无	100	100
	D1	100	100
	D2	100	100
	SG	100	100
SVM	无	76.08	60.42
	PCA	76.08	72.92
	D1	26.08	25.00
	D2	26.08	25.00
	SG	76.08	60.42
LSTM	无	95.47±0.83	97.74±0.30
	PCA	99.46±0.00	100±0.00
	D1	98.554±0.31	98.61±0.30
GRU神经网络	无	96.01±0.63	97.57±1.08
	PCA	99.46±0.00	100±0.00
	D1	98.73±0.31	98.96±0.00
逻辑回归	无	100	100
	PCA	100	100
	D1	97.83	100
	D2	95.65	89.58
	SG	100	100
朴素贝叶斯	PCA	95.65	95.83
	D1	89.13	91.67
决策树	PCA	78.26	89.58
	D1	89.13	91.67
最近邻算法	D1	93.48	91.67

3 讨论与结论

农产品的近红外光谱无损鉴别研究多将近红外光谱数据视为高维度、具有多重共线性且包含复杂信息的数据[10]，常采用 SVM、PLS-DA 等模型实现对农产品的鉴别[11]。本研究认为近红外光谱数据与自然语言数据均为序列数据，可采用 NLP 模型实现近红外光谱数据解析。以本研究中的光谱数据进行建模分析，逻辑回归、朴素贝叶斯以及 LSTM、GRU 神经网络模型均能实现与 PLS-DA 模型等同的分类准确率（最优预处理条件下，测试集准确率均为 100%）。Hong 等[14] 将近红外高光谱数据视为序列数据，使用 CNN-LSTM 以及 LSTM 等 NLP 方法对茶叶样本的近红外高光谱数据进行分类，可实现对新鲜茶叶中掺入陈年茶叶的鉴别且与 SVM 的效果接近。CNN-LSTM、LSTM 以及 SVM 模型的验证集准确率分别为 83.102%、82.548% 以及 80.332%，表明了 NLP 相关模型适用于序列数据—近红外高光谱的分类，这与本研究的结果相近。

Dong 等[27] 收集了我国三个省份共 96 枚山楂果实的近红外光谱，构建了 PLS-DA、反向传人工神经

网络模型（Backpropagation Artificial Neural Networks, BP-ANN）对山楂果实的产地进行预测。PLS-DA 在测试集中的准确率为 83%，BP-ANN 在测试集中的准确率为 95.8%，本研究使用的深度学习模型以及逻辑回归、朴素贝叶斯模型，在最优的预处理条件下测试集准确率均可达到或接近 100%，与之相较，准确率更高。Peng 等[35] 采用气相色谱飞行时间质谱法对 333 份武夷肉桂岩茶样品的挥发性成分进行了测定，并建立了多层感知机器、SVM、随机森林等多种机器学习模型，发现多层感知机在测试集中的准确率最高（83.2%）。本研究与之相比，不会破坏农产品，无须复杂耗时的检测分析，便可实现极高的分类准确率（最优可达 100%），由此可见近红外光谱无损鉴别农产品的优势。

然而，本研究仅可实现对四种山楂果实品种的鉴别，对于实现更多品种的鉴别以及同时实现山楂果实营养成分的测定，仍有待研究。该目标的实现，依赖于充足、高质量的数据集以及更加可靠的模型，而目前少有公开的农产品的近红外光谱数据集，这一定程度上限制了农产品的近红外光谱无损检测研究。当今 NLP 技术蓬勃发展，一系列功能强大的 NLP 模型如 Transformer、GPT-3 相继诞生[36−37]。可考虑将更复杂、功能更强的 NLP 模型运用于近红外光谱解析中，为农产品的近红外光谱解析提供更多、更有效的方法。

本研究使用 NLP 对山楂果实的近红外光谱数据进行分析，实现对山楂果实品种的无损鉴别。逻辑回归模型在光谱无预处理条件下以及经 PCA 或 SG 预处理后，验证集、测试集准确率均为 100%。LSTM 和 GRU 神经网络模型在光谱无预处理条件下，验证集准确率分别为 95.47%±0.83% 和 96.01%±0.63%，测试集准确率分别为 97.74%±0.30% 和 97.57%±1.08%，光谱经 PCA 预处理后验证集准确率可达 99.46%±0.00%，测试集准确率可达 100%±0.00%。朴素贝叶斯模型，在光谱经 PCA 预处理后，验证集准确率为 95.65%，测试集准确率为 95.83%。深度学习模型（LSTM 和 GRU 神经网络模型）以及传统机器学习模型（逻辑回归模型和朴素贝叶斯模型）依据山楂果实的近红外光谱鉴别山楂果实品种的准确率高，逻辑回归模型与深度学习模型（LSTM 和 GRU神经网络模型）的鉴别准确率最优。本研究表明基于自然语言处理的模型可用于山楂果实品种近红外无损鉴别，为农产品近红外光谱数据分析提供了更多可参考的模型，为更复杂、功能更强的 NLP 模型运用于该领域提供参考。

参考文献

[1] 李丽，袁建琴，王文斌. 山楂果肉中多酚闪式提取工艺的研究[J]. 中国酿造，2020，39(5)：179−182. [LI L, YUAN J Q, WANG W B. Flash extraction process of polyphenols from hawthorn pulp[J]. China Brewing, 2020, 39(5): 179−182.]

[2] 丰宝田，赵焕谆. 中国果树志·山楂卷[M]. 北京：中国林业出版社，1996：16−94. [FENG B T, ZHAO H X. Chinese fruit tree records·Hawthorn part[M]. Beijing: China Forestry Publishing House, 1996: 16−94.]

[3] 李长滨，牛畅炜，苏丽，等. 不同产地山药的近红外鉴别和差异分析[J]. 食品研究与开发，2022，43(15)：175−181. [LI C B, NIU C W, SU L, et al. Identification and variance analysis of Chinese Yam from different origins by nearinfrared spectroscopy[J]. Food Research and Development, 2022, 43(15): 175−181.]

[4] POREP J U, KAMMERER D R, CARLE R. On-line application of near infrared (NIR) spectroscopy in food production[J]. Trends in Food Science & Technology, 2015, 46(2): 211−230.

[5] YANG H L, ZANG H C, HU T, et al. Classigcation and quantigcation analysis of hawthorn from different origins with near-infrared diffuse reection spectroscopy[J]. Chinese Journal of Pharmaceutical Analysis, 2014, 34(3): 396−401.

[6] 张静，徐阳，姜彦武，等. 近红外光谱技术在葡萄及其制品品质检测中的应用研究进展[J]. 光谱学与光谱分析，2021，41(12)：3653−3659. [ZHANG J, XU Y, JIANG Y W, et al. Recent advances in application of near-lnfrared spectroscopy for quality detections of grapes and grape products[J]. Spectroscopy and Spectral Analysis, 2021, 41(12): 3653−3659.]

[7] ZHANG C, WU W Y, ZHOU L, et al. Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (*Lycium ruthenicum* Murr.) using near-infrared hyperspectral imaging[J]. Food Chemistry, 2020, 319: 126536.

[8] SHAO Y N, HE Y, BAO Y D, et al. Near-infrared spectroscopy for classification of oranges and prediction of the sugar content[J]. International Journal of Food Properties, 2009, 12(3): 644−658.

[9] TIAN X, WANG Q Y, HUANG W Q, et al. Online detection of apples with moldy core using the VIS/NIR full-transmittance spectra[J]. Postharvest Biology and Technology, 2020, 168: 111269.

[10] 高荣强，范世福. 现代近红外光谱分析技术的原理及应用[J]. 分析仪器，2002(3)：9−12. [GAO R Q, FAN S F. Principles and applications of modern near infrared spectroscopic techniques[J]. Analytical Instruments, 2002(3): 9−12.]

[11] LI X L, YI S L, HE S L, et al. Identification of pummelo cultivars by using VIS/NIR spectra and pattern recognition methods[J]. Precision Agriculture, 2016, 17(3): 365−374.

[12] 安鹏，曹丹平，赵宝银，等. 基于 LSTM 循环神经网络的储层物性参数预测方法研究[J]. 地球物理学进展，2019，34(5)：1849−1858. [AN P, CAO D P, ZHAO B Y, et al. Reservoir physical parameters prediction based on LSTM recurrent neural network[J]. Progress in Geophysics, 2019, 34(5): 1849−1858.]

[13] ZHONG Z, ZHANG X, YU J X, et al. Deep neural networks for the classification of pure and impure strawberry purees[J]. Sensors, 2020, 20(4): 1223.

[14] HONG Z Q, ZHANG C, KONG D D, et al. Identification of storage years of black tea using near-infrared hyperspectral imaging with deep learning methods[J]. Infrared Physics & Technology, 2021, 114: 10366.

[15] 陈勇，吴彩娥，熊智新. 基于衰减消去蜻蜓算法的小麦粉蛋白质近红外特征波长优选[J]. 食品科学，2022，43(14)：219−225. [CHEN Y, WU C E, XIONG Z X. Selection of near infrared wavelengths using attenuation elimination-binary dragonfly algorithm for wheat flour protein content prediction[J]. Food Science, 2022, 43(14): 219−225.]

[16] 王燕南. 基于深度学习的说话人无关单通道语音分离[D].

合肥: 中国科学技术大学, 2017. [WANG Y N. Speaker independent single-channel speech separation based on deep learning[D]. Hefei: University of Science and Technology of China, 2017.]
[17] 李超凡, 马凯. 基于注意力机制结合 CNN-BiLSTM 模型的电子病历文本分类[J]. 科学技术与工程, 2022, 22(6): 2363−2370. [LI C F, MA K. Electronic medical record text classification based on attention mechanism combined with CNN-BILSTM[J]. Science Technology and Engineering, 2022, 22(6): 2363−2370.]
[18] FAN E. Extended tanh-function method and its applications to nonlinear equations[J]. Physics Letters A, 2000, 277(4): 212−218.
[19] YIN X Y, GOUDRIAAN J, LANTINGA E A, et al. A flexible sigmoid function of determinate growth[J]. Annals of Botany, 2003, 91(3): 361−371.
[20] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using rnn encoder-decoder for statis tical machine translation[C]//Doha: Conference on Empirical Methods in Natural Language Processing, 2014: 1724-1734.
[21] 王鹏新, 王婕, 田惠仁, 等. 基于遥感多参数和门控循环单元网络的冬小麦单产估测[J]. 农业机械学报, 2022, 53(9): 207−216. [WANG P X, WANG J, TIAN H R, et al. Yield estimation of winter wheat based on multiple remotely sensed parametersand gated recurrent unit neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(9): 207−216.]
[22] SPERANDEI S. Understanding logistic regression analysis [J]. Biochemia Medica, 2014, 24(1): 12−18.
[23] 管小艳. 贝叶斯网基础及应用[M]. 武汉: 武汉大学出版社, 2019: 19-20. [JIAN X Y. Foundation and application of bayesian networks[M]. Wuhan: Wuhan University Press, 2019: 19-20.]
[24] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 153−174. [ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 153−174.]
[25] 匡芳君. 大数据挖掘与分析在金融领域中的应用研究[M]. 哈尔滨: 哈尔滨工业大学出版社, 2020: 68−79. [KUANG F J. Research on the application of big data mining and analysis in the financial field[M]. Harbin: Harbin Institute of Technology Press, 2020: 68−79.]
[26] 覃礼堂, 刘树深, 肖乾芬, 等. QSAR 模型内部和外部验证方法综述[J]. 环境化学, 2013, 32(7): 1205−1211. [QIN L T, LIU S S, XIAO Q F, et al. Internal and external validtions of QSAR model: Review[J]. Environmental Chemistry, 2013, 32(7): 1205−1211.]
[27] DONG W J, NI Y N, KOKOT S. A near-infrared reflectance spectroscopy method for direct analysis of several chemical components and properties of fruit, for example, Chinese hawthorn[J]. Journal of Agricultural and Food Chemistry, 2013, 61(3): 540−546.
[28] 杨暑东. Emoji 自然语言处理综述[J]. 计算机应用与软件, 2022, 39(9): 11−20. [YANG S D. Survey on emoji-embedded natural language processing[J]. Computer Applications and Software, 2022, 39(9): 11−20.]
[29] 李华旭. 基于 RNN 和 Transformer 模型的自然语言处理研究综述[J]. 信息记录材料, 2021, 22(12): 7−10. [LI H X. A review of natural language processing based on RNN and Transformer models[J]. Information Recording Materials, 2021, 22(12): 7−10.]
[30] 邵帅斌, 刘美含, 石宇晴, 等. 基于卷积神经网络的乳粉掺杂物拉曼光谱分类方法[J]. 食品科学, 2022, 43(14): 296−301. [SHAO S B, LIU M H, SHI Y Q, et al. Raman spectroscopic classification of adulterants in milk powder samples using convolutional neural network[J]. Food Science, 2022, 43(14): 296−301.]
[31] 李思奇, 吕王勇, 邓柙, 等. 基于改进 PCA 的朴素贝叶斯分类算法[J]. 统计与决策, 2022, 38(1): 34−37. [LI S Q, LÜ W Y, DENG X, et al. Naive Bayes classification algorithm based on improved PCA[J]. Statistics & Decision, 2022, 38(1): 34−37.]
[32] 白文明, 王来兵, 成日青, 等. 近红外高光谱成像技术在药物分析中的研究进展[J]. 药物分析杂志, 2018, 38(10): 1661−1667. [BAI W M, WANG L B, CHENG R Q, et al. Research advance in pharmaceutical analysis based on near-infrared hyperspectral imaging technique[J]. Chinese Journal of Pharmaceutical Analysis, 2018, 38(10): 1661−1667.]
[33] 李楚进, 付泽正. 对朴素贝叶斯分类器的改进[J]. 统计与决策, 2016(21): 9−11. [LI C J, FU Z Z. Improvement of naive Bayes classifier[J]. Statistics & Decision, 2016(21): 9−11.]
[34] 田海清. 西瓜品质可见/近红外光谱无损检测技术研究[D]. 杭州: 浙江大学, 2006. [TIAN H Q. Nondestructive evaluation of watermelon internal quality byvisible and near-infrared spectroscopy [D]. Hangzhou: Zhejiang University, 2006.]
[35] PENG Y F, ZHENG C, GUO S, et al. Metabolomics integrated with machine learning to discriminate the geographic origin of Rougui Wuyi rock tea[J]. NPJ Science of Food, 2023, 7(1): 7−10.
[36] WANG F Y, YANG J, WANG X X, et al.Chat with chatgpt on industry 5.0: Learning and decision-making for intelligent industries[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(4): 831−834.
[37] FLORIDI L, CHIRIATTI M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30(4): 681−694.