

引用: 张浩, 刘振, 王玲, 等. 基于近红外光谱结合机器学习算法检测食用明胶品种溯源的研究 [J]. 河南农业大学学报, 2021, 55(3): 460-467. DOI: 10.16445/j.cnki.1000-2340.20210409.003

基于近红外光谱结合机器学习算法检测 食用明胶品种溯源的研究

张浩^{1,2}, 刘振¹, 王玲^{1,2}, 胡建东^{1,2}

(1. 河南农业大学机电工程学院, 河南 郑州 450002;

2. 河南省农业激光技术国际联合实验室, 河南 郑州 450002)

摘要: 采用近红外光谱结合机器学习方法, 对 5 种不同来源的食用明胶进行鉴别。利用 Savitzky-Golay 平滑去噪、多元散射校正和最大最小归一化等方法对原始光谱数据进行预处理。将预处理的光谱数据划分为训练集和验证集, 分别采用支持向量机 (support vector machine, SVM)、随机森林 (random forest, RF) 和反向传播神经网络 (back propagation neural network, BPNN) 建立识别模型。结果表明, SVM 模型、RF 模型和 BPNN 模型的总体准确率均高达 97% 以上, 其中 BPNN 模型的准确率为 100%, 明显优于其他 2 种模型, 能够实现 5 种不同来源食用明胶的完全识别, 而且其运算速度最短, 更适用于明胶品种的溯源。

关键词: 食用明胶; 品种溯源; 近红外光谱; 机器学习

中图分类号: TS202

文献标志码: A

文章编号: 1000-2340(2021)03-0460-08

Determination of edible gelatin origins based on near-infrared spectroscopy coupled with machine learning methods

ZHANG Hao^{1,2}, LIU Zhen¹, WANG Ling^{1,2}, HU Jiandong^{1,2}

(1. College of Mechanical and Electrical Engineering, Henan Agricultural University, Zhengzhou 450002, China; 2. Henan International Joint Laboratory of Laser Technology in Agricultural Sciences, Zhengzhou 450002, China)

Abstract: Near-infrared spectroscopy (NIRS) combined with machine learning methods was used to identify five edible gelatins of different origins. The methods of Savitzky-Golay smoothing, multivariate scattering correction, and maximum and minimum normalization were used to preprocess the original spectral data. The preprocessed spectral data were divided into a training set and a validation set, and then support vector machine (SVM), random forest (RF), and back propagation neural network (BPNN) were utilized to establish the classification model, respectively. The results show that the overall accuracy of the SVM model, RF model, and BPNN model are higher up to 97%, where the accuracy of the BPNN model is 100%, which is significantly better than the other two models. Based on the BPNN model, 5 edible gelatins of different origins can be completely identified. Moreover, the operation speed of BPNN model is the shortest, thus BPNN is regarded as the most suitable method for the identification of gelatin origins.

收稿日期: 2020-10-30

基金项目: 国家自然科学基金项目 (32071890); 河南省自然科学基金项目 (202300410197); 中国博士后科学基金面上项目 (2017M612399); 河南省高等学校青年骨干教师项目 (2020GGJS046); 河南农业大学科技创新基金项目 (KJCX2018A09)

作者简介: 张浩 (1988—), 男, 河南驻马店人, 讲师, 博士, 从事农产品品质和安全的无损检测。

通信作者: 胡建东 (1965—), 男, 江西新余人, 教授, 博士, 博士生导师。

Key words: edible gelatin; origins; NIRS; machine learning

食用明胶是由动物的新鲜原皮和骨骼经过复杂工序提取出胶原蛋白并水解加工而成的。由于其蛋白质含量高,无脂肪和无胆固醇,目前主要作为食品增稠剂、胶凝剂和粘合剂被广泛应用于食品工业中。食用明胶通常来源于猪皮、猪骨、牛皮、牛骨、鱼皮、鱼鳞等,不同来源食用明胶的二级结构和重金属含量是不同的,导致不同来源食用明胶的用途也有所差异^[1-2]。因此,对于食品生产者来说,需要知道确切的食用明胶来源,以便于在食品制作过程中合理使用不同来源的食用明胶,制作出符合国家标准的明胶类食品,同时也能满足不同消费者的口味需求,有利于保护消费者的利益。目前,仅从色泽、气味等方面来考虑,很难区分出不同来源的食用明胶,因此需要一种快速有效的方法实现食用明胶品种的溯源研究。

食用明胶的传统检测方法包括液相色谱法、质谱法、电泳法、酶联免疫法等,尽管这些方法具有较高的检测灵敏度和精度,但是需要昂贵的仪器和复杂的样品前处理、比较耗时^[3-5]。近红外光谱技术(Near Infrared Reflectance Spectroscopy, NIRS)作为一种快速无损的方法,目前被广泛地应用到食品领域中^[6-8]。通过研究明胶的理化特性,NIRS已经被用于食用明胶的检测中。SEGTNAN等^[9]采用近红外光谱对猪明胶和牛明胶的理化参数如冻力(Bloom值)、黏度、pH值和水分含量进行预测;DUCONSEILLE等^[10]利用近红外光谱技术和荧光光谱技术,实现了对明胶在老化过程中分子变化的监测;张浩等^[11]利用近红外光谱对6种不同掺杂比例的食用明胶进行检测,通过采用不同模式识别方法实现了掺假食用明胶的精准识别。CEBI等^[12]采用傅里叶变换红外光谱技术结合主成分-聚类分析方法实现猪明胶、牛明胶和鱼明胶的较好区分。目前的研究中主要对动物皮胶进行检测,缺

少对骨胶的检测。另外,大部分研究主要采用无监督学习方法如主成分分析法对食用明胶进行分析,分类准确率有待进一步提高。在此基础上,本研究拟采用近红外光谱技术结合3种机器学习方法对5种不同来源的食用明胶(猪皮、猪骨、牛皮、牛骨、鱼皮)进行鉴别研究。

1 材料与方法

1.1 材料

实验所用5种不同来源的食用明胶(猪皮、猪骨、牛皮、牛骨、鱼皮)购置于厦门康宝生物科技有限公司,其冻力(Bloom值)分别为220,200,220,250,250。明胶样品用实验室粉碎机(拜杰BJ-800A)磨成粉末,然后用100目筛进行过滤。随后,将每种明胶源称重5g,放入压片机(YP-30T,天津市金孚伦科技有限公司)的模具中,将压力保持在30MPa,3min后压成明胶片。在此过程中,为了避免污染和干扰,每次样品制备后都要清洗模具。每种明胶源制备120个明胶片,共得到600个。明胶片的厚度和直径分别为4.5mm和35mm。

1.2 光谱采集

实验所采用的近红外光谱测量装置由卤钨灯光源(HL-2000,海洋光学)、光纤可调衰减器、分叉式光纤反射探头和近红外光谱仪(NIRQuest 256,海洋光学)等构成,如图1所示。其中光源的波长范围为360~2400nm,近红外光谱仪的光谱范围为900~2500nm。光纤反射探头由7根芯径为600μm的光纤组成,其中一根连接光源,用于传输入射光,另外的6根连接近红外光谱仪,用于收集反射光。实验过程中,光谱每隔3nm采集一个点,每次平均3次,积分时间设置为100ms,平滑度设置为10。

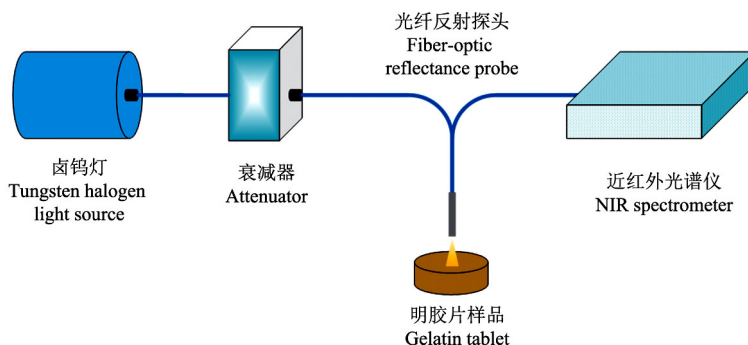


图1 明胶近红外光谱测量装置

Fig. 1 NIR spectral measurement system of gelatin tablets

1.3 识别模型建立和性能评估

本研究分别采用支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RF)、和反向传播神经网络(Back Propagation Neural Network, BPNN) 3 种机器学习算法建立明胶识别模型。3 种模型的建立过程均采用 Matlab 程序,其中 SVM 建模基于林智仁等^[13]开发的 LIBSVM 工具箱编写的 Matlab 程序,RF 建模基于 Jaiantilal 开发的 RF_MexStandalone-v0.02 工具箱编写的 Matlab 程序,BPNN 建模采用 Matlab 软件人工神经网络工具箱编写的 Matlab 程序,所有 Matlab 程序均在 MATLAB R2017a 软件下运行。

为了评估所建模型的识别性能,本文采用 4 种机器学习分类模型效果评估指标,即准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1-score。准确率 $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$,表示所有样本被分类正确的比例;精确率 $Precision = \frac{TP}{TP+FP}$ 表示正样本的分类准确率;召回率 $Recall = \frac{TP}{TP+FN}$,表示分类正确的正样本占总的分类正确样本的比例; $F1\text{-score} = \frac{2 * Precision * Recall}{Precision + Recall}$,表示精确率和召回率的调和平均。其中,TP 表示原本

属于正类的样本分为正类,TN 表示原本属于负类的样本分为负类,FP 表示原本属于负类的样本错分为正类,FN 表示原本属于正类的样本错分为负类。

2 结果与分析

2.1 明胶胶体的近红外吸收光谱特性

为了降低由于仪器、样本和环境因素对光谱造成的噪声干扰,提高光谱信噪比,近红外光谱数据依次通过 Savitzky-Golay(SG)平滑去噪、多元散射校正和最大最小归一化方法进行预处理,其中 SG 平滑方法采用 12 个点的窗口宽度和三次多项式拟合。图 2 为测量所得 5 种不同来源明胶样品的预处理平均光谱,选取光谱波段范围为 1 100~2 200 nm,可以看出其光谱形状基本上是相同的,只是个别波长处的吸收存在差异。图中明胶近红外光谱具有一些比较明显的特征吸收峰,主要由明胶中一些特定的化学官能团的振动所造成的,如 CH、OH、NH 等。水分子的特征吸收峰位于 1 490 nm 和 1 930 nm 左右,主要是由 O—H 基团的泛频和组频吸收带造成的^[9,14]。此外,1 200 nm 和 1 730 nm 的峰值是由明胶中 C—H 基团的泛频和组频吸收带产生的,源自于明胶中芳香族氨基酸^[14]。而 2 000~2 100 nm 的峰值主要来自于 N—H 的组频吸收带,源自于明胶中蛋白质的侧链^[9,14,15]。

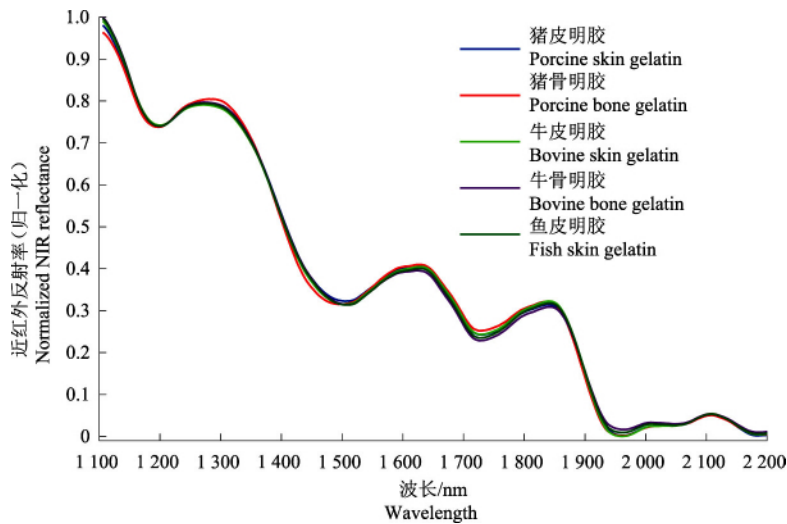


图 2 5 种不同来源明胶样品的归一化近红外光谱

Fig. 2 Normalized NIR spectra from 5 different kinds of gelatin samples

2.2 明胶来源识别模型的构建

为了建立合适的识别模型并评价模型的有效性,光谱数据被分为训练集和验证集,其中训练集用来建立识别模型,验证集用来检验模型的有效性。本研究采用十折交叉验证(10-fold cross vali-

dation)方法从 600 个明胶样品中选择 531 个样本作为训练集,剩余的 69 个样本作为验证集。

2.2.1 SVM 模型 由于每条近红外光谱共 432 个数据点,建模时不仅计算量大,而且会影响分类的精准性。这里采用竞争性自适应重加权算法

(Competitive Adaptive Reweighted Sampling, CARS) 方法从 432 个波长变量选取特征波长变量。本文中, CARS 的初始化参数设置为: 最大主成分数 $A = 10$, 交叉验证组数 $K = 10$, 最大蒙特卡洛 (Monte Carlo, MC) 采样运行次数 $N = 50$ 。如图 2 所示, 随着 MC 采样运行次数的增加 (图 3a), 被筛选出来

的变量数逐渐减小, 而交叉验证均方根误差 (RMSECV) 出现先降低会升高的趋势 (图 3b)。在第 24 次采样时 (图 3c), RMSECV 值最小, 随后开始回升, 表明此时筛选出的波长变量数最优, 共 31 个波长变量。如图 3d 所示, 优选的 31 个波长变量包含了大部分图 1 所示的特征吸收峰值。

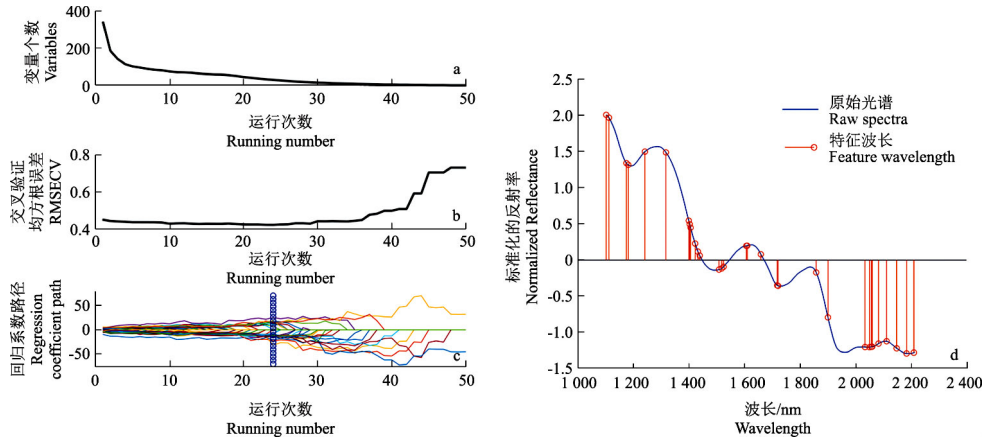
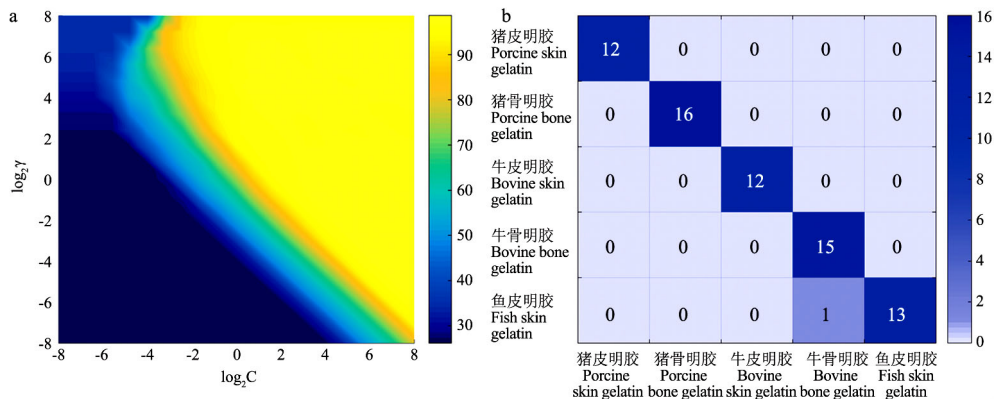


图 3 基于 CARS 的近红外光谱特征变量筛选

Fig. 3 Feature variable selection of NIR spectra based on CARS

由于 SVM 模型参数将极大影响模型的预测准确率, 因此需要寻找最优的惩罚因子 C 与核函数参数 γ 。本次研究中, 网格寻优算法 (Grid Search, GS) 被用来获得最优的 C 和 γ , 并建立相应的 SVM 模型。最优的 C 和 γ 通过交叉验证准确率的大小来确定, 如图 4a 所示, 当 $C = 256$, $\gamma = 3.03$ 时, 训练模型可以得到最优的交叉验证准确率为 98.68%。在优化参数情况下, 将所建 SVM 模型作用于验证

集得到的混淆矩阵如图 4b 所示, 可以看出鱼皮明胶有一个样品被错误分成牛骨明胶, 使得验证集的总体准确率为 98.55% (68/69)。根据验证集混淆矩阵得到的 4 个模型评价指标值如表 1 所示, 可以看出验证集的平均准确率、精确率、召回率和 F1-score 均高于 98%, 表明 SVM 模型对于 5 种不同来源的明胶具有较好的识别能力。



a. 基于 GS 方法的 SVM 模型参数的优化; b. 基于 SVM 模型的验证集混淆矩阵。

a. Parameter optimization of SVM model based on GS method; b. Confusion matrix of validation set based on SVM model.

图 4 SVM 模型的优化及预测结果

Fig. 4 The optimization of SVM model and the prediction result

2.2.2 RF 模型 RF 模型建立过程中, 由于随机森林所包含的决策树个数 n_{tree} 和构建决策树分支时随机抽样的变量个数 m_{try} 这两个参数影响 RF 模

型的准确性, 需要选择合适的 n_{tree} 和 m_{try} 以有效降低模型的预测错误率。 n_{tree} 和 m_{try} 值对 RF 模型性能的影响可以由袋外 (Out of Bag, OOB) 错误率和预

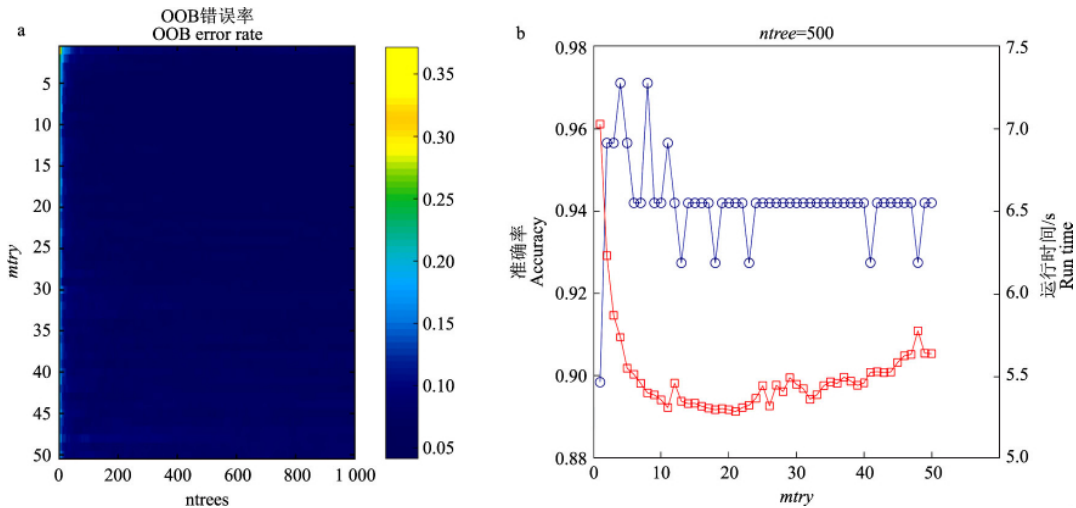
测准确率确定,通常情况下 n_{tree} 值越大,OOB 错误率越稳定,但运行时间越长; m_{try} 越大,OOB 错误率越低。默认情况下 n_{tree} 的值为 500, m_{try} 的取值范围在 1 到 m 之间,其中 m 为数据集变量个数的平方根。如图 5a 所示,随着决策树个数增加,OOB 错误率逐渐减小,当 n_{tree} 超过 200 以后,OOB 误差率基

本保持稳定。综合考虑 n_{tree} 值对 OOB 错误率稳定性和运行时间的影响,本文中选择默认值 500 作为 n_{tree} 的最优值。当 $n_{tree} = 500$ 时,随着 m_{try} 值的增加,RF 模型的准确率和运行时间变化如图 5b 所示,可以看出当 $m_{try} = 8$ 时,RF 模型获得最高的准确率和较少的运行时间。

表 1 基于验证集混淆矩阵的 SVM 模型评价指标

Table 1 The evaluation indicators of SVM model based on confusion matrix of validation set

指标 Indicator	猪皮明胶 Porcine skin gelatin	猪骨明胶 Porcine bone gelatin	牛皮明胶 Bovine skin gelatin	牛骨明胶 Bovine bone gelatin	鱼皮明胶 Fish skin gelatin	平均值 Average
准确率 Accuracy	1	1	1	0.985 5	0.985 5	0.994 2
精度 Precision	1	1	1	0.937 5	1	0.987 5
召回率 Recall	1	1	1	1	0.928 6	0.985 7
F1-score	1	1	1	0.967 7	0.963 0	0.986 1



a. RF 模型中决策树个数 n_{tree} 的优化; b. 最优决策树个数 $n_{tree} = 500$ 时,RF 模型变量个数 m_{try} 的优化。
a. The optimization of n_{tree} in RF model; b. The optimization of m_{try} in RF model when n_{tree} is 500.

图 5 RF 模型参数的优化

Fig. 5 Parameter optimization of RF model

本研究采用平均精确度下降 (Mean decrease accuracy) 和平均基尼指数下降 (Mean decrease of Gini index) 来判断波长变量中的特征重要性,某一变量的这 2 个参数值越大,表明该变量的重要性越大。如图 6 所示,可以判断出 RF 模型中重要性较大的波长基本包含了图 2 所示的特征吸收峰值。

以 $n_{tree} = 500$, $m_{try} = 8$ 作为最优参数建立随机森林模型,获得的验证集预测结果如图 7 所示,从混淆矩阵中可以看出,有两个明胶样品被错误分类,其中一个猪皮明胶样品被错误识别为牛骨明胶,一个鱼皮明胶样品被错误识别为牛皮明胶,使得 RF 模型的总体准确率为 97.1% (67/69)。根据验证集混淆矩阵得到的 4 个模型评价指标值如表 2 所示,其平均准确率、精确率、召回率和 F1-score

均高于 96%,表明 RF 模型的识别能力稍微弱于 SVM 模型,尽管如此,RF 模型也展示了较好的明胶识别能力。

2.2.3 BPNN 模型 为了减小建模时的计算量,提高分类的精准性,本文选择 CARS 方法进行特征选择筛选,选择出 31 个特征波长变量作为 BPNN 的输入节点数,输出节点数为 5。在 BPNN 模型中,隐含层节点数对模型的性能影响很大,因此需要确定一个合适的隐含层节点数。本文首先通过采用以下 3 个参考公式^[16] 确定最佳隐含层节点数所在的一个大致范围,即 $l < n - 1$, $l < \sqrt{m+n} + a$, $l = \log_2 n$ 其中 l 为隐含层节点数, m 为输出层节点数, n 为输入层节点数, a 为 0~10 之间的常数。然后通过依次调整隐含层节点数来优化建模,如图 8a 所示,随

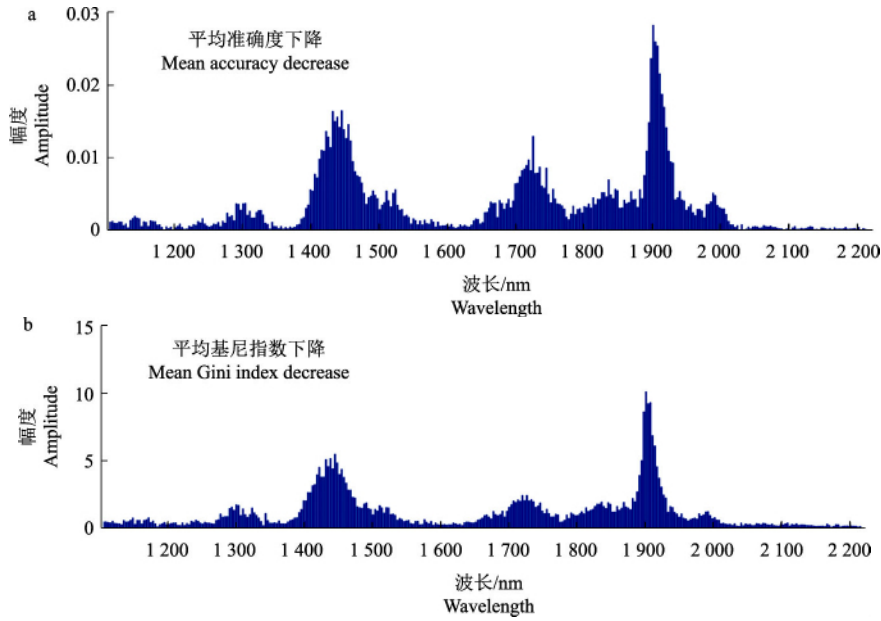


图 6 RF 模型中的波长变量重要性
Fig. 6 Wavelength variable importance of RF model

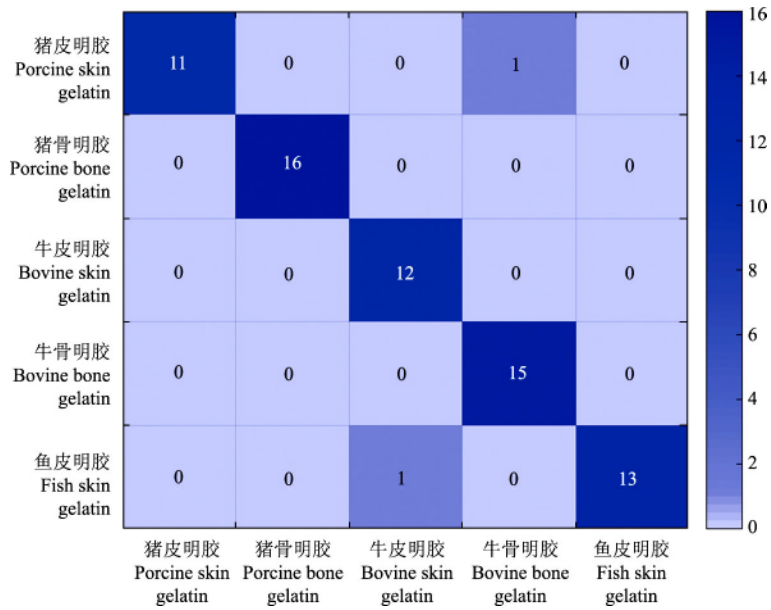


图 7 基于 RF 模型的验证集混淆矩阵
Fig. 7 Confusion matrix of validation set based on RF model

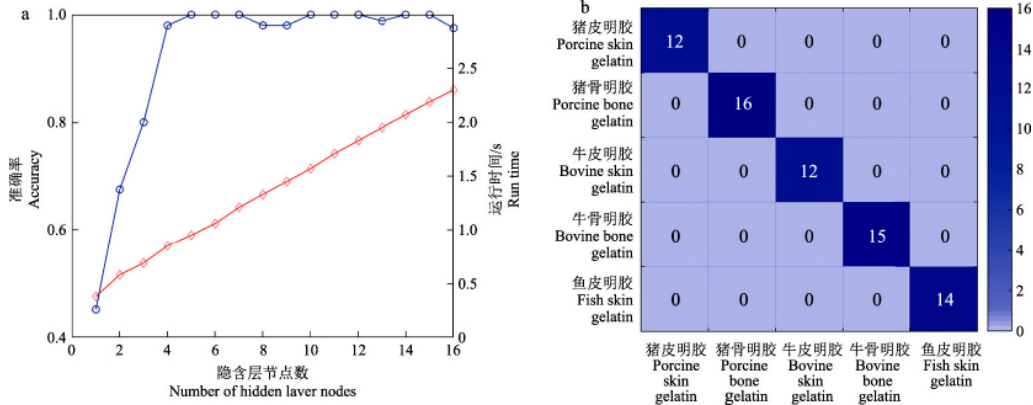
表 2 基于验证集混淆矩阵的 RF 模型评价指标

Table 2 The evaluation indicators of RF model based on confusion matrix of validation set

指标 Indicator	猪皮明胶 Porcine skin gelatin	猪骨明胶 Porcine bone gelatin	牛皮明胶 Bovine skin gelatin	牛骨明胶 Bovine bone gelatin	鱼皮明胶 Fish skin gelatin	平均值 Average
准确率 Accuracy	0.985 5	1	0.985 5	0.985 5	0.985 5	0.988 4
精度 Precision	1	1	0.923 1	0.937 5	1	0.972 1
召回率 Recall	0.916 7	1	1	1	0.928 6	0.969 1
F1-score	0.956 5	1	0.960 0	0.967 7	0.963 0	0.969 4

着隐含层节点数增加,模型分类准确率增加,最终趋于相对稳定状态,而运行时间随着隐含层节点数的增加而逐渐增大。当隐含层节点数为 5 时,模型的准确率为 100%,而且运行时间较短。因此,通过优化获得一个 31(输入层节点)-5(隐含层节点)-5(输出层节点)的最佳三层 BPNN 模型。将

所建 BPNN 模型应用于验证集,得到的预测结果如图 8b 所示,验证集所有样品均被正确识别,得到总体准确率为 100%,进而使得其平均准确率、精确率、召回率和 F1-score 均为 100%。结果表明 BPNN 模型具有最优的识别能力。



a. BPNN 模型隐含层节点数的优化; b. 基于 BPNN 模型的验证集混淆矩阵。

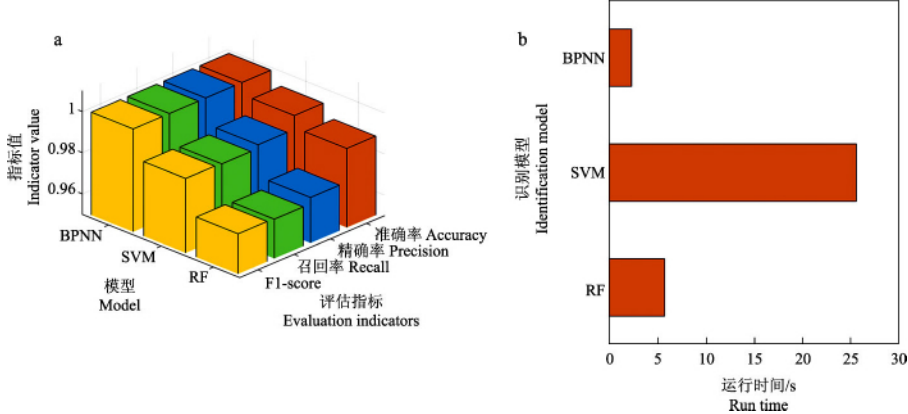
a. Optimization of the number of hidden layer nodes in BPNN model; b. Confusion matrix of validation set based on BPNN model.

图 8 BPNN 模型的优化及预测结果

Fig. 8 The optimization of BPNN model and the prediction result

2.2.4 3 种识别模型的性能比较 研究表明,利用 SVM、RF 和 BPNN 这 3 种识别模型均能较好的进行明胶品种溯源分析,其验证集的总体正确识别率分别为 98.55%、97.1%和 100%。图 9 所示为 3 种识别模型的性能比较,从图 9a 所示的分类模型效果评估指标可以判断出 BPNN 模型要明显优于其他两种模型,能够实现对 5 种不同来源明胶样品的完

全识别,而 SVM 模型要优于 RF 模型。图 9b 所示为 3 种模型的运行时间(或运算速度),可以看出 BPNN 模型的运算速度(2.4 s)同样优于其他 2 种模型,而 RF 模型的运算速度要由于 SVM 模型。因此,从分类模型的准确率、精确率、召回率、F1-score 和运算速度等方面来分析,BPNN 模型均展现出了最优的识别性能。



a. 评估指标的比较; b. 运行时间的比较

a. Comparison of evaluation indicators; b. Comparison of run time

图 9 3 种识别模型的性能比较。

Fig. 9 Comparison of the performance between 3 kinds of identification models.

3 结论

针对目前食品工业中存在的不同来源食用明

胶难以区分问题,本研究采用近红外光谱测量装置获得 5 种不同来源明胶样品的近红外吸收光谱,通过 SG 平滑去噪和最大最小归一化方法进行预处理

理,采用3种机器学习方法建立分类模型研究明胶品种溯源的可行性。研究结果表明,这3种识别模型对不同来源食用明胶均具有较好的识别能力,验证集的总体正确识别率均达到97%以上,相比之下,BPNN模型具有最优的识别效果,其验证集总体正确识别率为100%,能够实现对5种不同来源明胶样品的完全识别,而且其运算速度最快。与传统方法相比,近红外光谱技术简单、便携、成本低、无需复杂的样品处理,可以实现无损快速检测。此外,本研究采用3种有监督学习方法进行明胶识别,可以获得更好的识别精度。因此,本研究中近红外光谱技术结合机器学习方法用于快速准确地实现食用明胶品种溯源是完全可行且有效的。该方法也可以用于食用明胶的质量检测和掺假研究中,同时也为其他食品或添加剂的安全检测提供了研究思路。

参考文献:

- [1] 王卫东,李超,孙月娥.鱼皮明胶的制备、特性及应用[J].食品科学,2009,30(23):484-488.
- [2] 史洲铭,赵宇威,杨晓慧,等.几种不同来源食用明胶的蛋白二级结构分析及安全性能比较[J].应用化工,2014,43(9):1719-1722.
- [3] AZIRA T N,MAN Y B,HAFIDZ R,et al. Use of principal component analysis for differentiation of gelatine sources based on polypeptide molecular weights [J]. Food Chemistry,2014,151:286-292.
- [4] AZILAWATI M I,HASHIM D M,JAMILAH B,et al. RP-HPLC method using 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate incorporated with normalization technique in principal component analysis to differentiate the bovine, porcine and fish gelatins [J]. Food Chemistry,2015,172:368-376.
- [5] TUKIRAN N A,ISMAIL A,MUSTAFA S,et al. Determination of porcine gelatin in edible bird's nest by competitive indirect ELISA based on anti-peptide polyclonal antibody [J]. Food Control,2016,59:561-566.
- [6] GRASSI S,ALAMPRESE C. Advances in NIR spectroscopy applied to process analytical technology in food industries [J]. Current Opinion in Food Science,2018,22:17-21.
- [7] 何鸿举,朱亚东,王慧,等.近红外光谱技术在生鲜禽肉质量检测中应用的研究进展[J].食品科学,2019,40(21):317-323.
- [8] ZAREEF M,CHEN Q,HASSAN M M,et al. An overview on the applications of typical non-linear algorithms coupled with NIR spectroscopy in food analysis [J]. Food Engineering Reviews,2020,12:173-190.
- [9] SEGTMAN V H,KVAAL K,RUKKE E O,et al. Rapid assessment of physico-chemical properties of gelatine using near infrared spectroscopy [J]. Food Hydrocolloid,2003,17(5):585-592.
- [10] DUCONSEILLE A,ANDUEZA D,PICARD F,et al. Molecular changes in gelatin aging observed by NIR and fluorescence spectroscopy [J]. Food Hydrocolloid,2016,61:496-503.
- [11] ZHANG H,SUN H,WANG L,et al. Near infrared spectroscopy based on supervised pattern recognition methods for rapid identification of adulterated edible gelatin [J]. Journal of Spectroscopy,2018,218:1-9.
- [12] CEBI N,DURAK M Z,TOKER O S,et al. An evaluation of fourier transforms infrared spectroscopy method for the classification and discrimination of bovine, porcine and fish gelatins [J]. Food Chemistry,2016,190:1109-1115.
- [13] CHANG C C,LIN C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology,2011,2:1-27.
- [14] CSEFALVAYOVA L,PELIKAN M,CIGIC I K,et al. Use of genetic algorithms with multivariate regression for determination of gelatine in historic papers based on FT-IR and NIR spectral data [J]. Talanta,2010,82(5):1784-1790.
- [15] DUCONSEILLE A,ANDUEZA D,PICARD F,et al. Variability in pig skin gelatin properties related to production site: a near infrared and fluorescence spectroscopy study [J]. Food Hydrocolloids,2017,63:108-119.
- [16] 王小川. MATLAB神经网络43个案例分析[M].北京:北京航空航天大学出版社,2013.

(责任编辑:蒋国良)