中国农业大学学报 2018,23(9):166-170
Journal of China Agricultural University
http://zgnydxxb.ijournals.cn
DOI:10.11841/j.issn.1007-4333.2018.09.20

# 机器学习在近红外光谱法判别鲍鱼品种研究中的应用

高婧娴[1] 黄扬明[1] 雷春丽[1] 闫 红[1] 闵顺耕[1] 熊艳梅[1] 闵志勇[2*]
(1.中国农业大学 理学院,北京 100193;2.莆田学院 环境与生物工程学院,福建 莆田 351100)

**摘 要** 为解决市场上鲍鱼产品缺乏科学分类方法的问题,利用近红外光谱分析技术结合机器学习方法对鲍鱼快速分类进行研究,使用 MicroNIR™1700 便携式近红外光谱仪采集 3 种鲍鱼,即绿盘鲍(25 只)、红壳鲍(31 只)、皱纹盘鲍(35 只)的光谱数据,采用 CART 算法建立鲍鱼分类决策树模型,以模型对测试集样本的预测准确率衡量决策树模型优劣,分裂策略为在每个节点处选择 Gini 不纯度最大的方式进行分裂,通过交叉验证控制决策树深度。结果表明,对训练集 180 条光谱建立模型,采用 5 折交叉验证,模型准确率为 90.00%,对测试集 93 条光谱的预测准确率为 90.32%。本研究方法可以很好地区分绿盘鲍、红壳鲍和皱纹盘鲍,满足鲍鱼现场快速分类的需求。

**关键词** 近红外光谱;鲍鱼;水产品;机器学习;分类回归树
**中图分类号** O657.61 **文章编号** 1007-4333(2018)09-0166-05 **文献标志码** A

# Discrimination of abalone (sub)species basing on near-infrared spectroscopy and machine learning

GAO Jingxian[1], HUANG Yangming[1], LEI Chunli[1], YAN Hong[1],
MIN Shungeng[1], XIONG Yanmei[1], MIN Zhiyong[2*]
(1. College of Science, China Agricultural University, Beijing 100193, China;
2. College of Environmental and Biological Engineering, Putian University, Putian 351100, China)

**Abstract** To study the rapid classification of abalone (sub) species, near-infrared spectroscopy combined with machine learning method was used. The spectra of three (sub) species abalone samples were obtained by portable near-infrared spectrometer MicroNIR™1700. The spectra was divided into training set and testing set, which were 180 and 93 spectra, respectively. The CART method was applied to build a decision tree model and its criterion was Gini impurity. Cross validation was used in the model to control the depth of decision tree model. The accuracy rate of the training set was 90.00%. The final accuracy rate of the testing set reached 90.32%. The combination of NIRS and chemometric method was proposed in this study as a fast and new method for the classification of different abalone (sub)species.

**Keywords** near infrared spectroscopy; abalone; aquatic product; machine learning; classification and regression tree

鲍鱼是中国传统的名贵食材,含有丰富的蛋白质、氨基酸以及多种人体必需的钙、镁、铁、硒等微量元素,对提高人体免疫力以及增强抗病能力等有一定的作用,是一种良好的滋补品。目前,市面上常见的鲍鱼种类繁多,不同品种的鲍鱼在生长习性、养殖周期、品质和价格上都有较大差异,传统的鲍鱼鉴别方法是通过观察鲍鱼的外观特点进行分类,这种方法简单快速,但是对鉴别者的要求高,需要鉴别者有一定的经验积累,主观性较强,不利于市场监管。因此,利用科学手段建立一种鲍鱼快速分类方法对于鲍鱼养殖、贸易以及市场监管都具有非常重要的意义。

收稿日期:2017-10-30
基金项目:国家自然科学基金项目(31301685);莆田市科技计划区域重点项目(2015N1002);福建省科技计划区域重大项目(2009N3002)
第一作者:高婧娴,硕士研究生,E-mail:gaojx@cau.edu.cn
通讯作者:闵志勇,副教授,主要从事海洋养殖技术研究,E-mail:minzhiyong@126.com

近红外光谱具有分析速度快、样品无需预处理、分析效率高、分析成本低，以及分析不具有破坏性等优点，在农产品和食品研究领域中得到广泛应用[1-5]。机器学习作为一门新兴的交叉学科，因其具有卓越的近红外光谱数据分析能力[6-8]，在食品与农产品的溯源分析[9-12]、品质鉴定[13-19]、品种分类[20]以及真伪鉴别[21-28]等方面应用均取得了较好的分类结果。由于活体动物光谱采集存在一定的困难，加上取样重现性不好，因此采用近红外光谱技术对水产品活体进行原位检测的研究较少，目前尚未有采用近红外光谱分析技术进行鲜活鲍鱼品种鉴定的报道。

本研究拟采用 MicroNIR™1700 便携式近红外光谱仪采集来自福建省莆田市的 3 种鲍鱼：绿盘鲍、红壳鲍、皱纹盘鲍的近红外光谱，光谱数据经标准化变换（Standardization）后采用分类回归树（Classification and regression tree，CART）算法建立分类模型，以期为活体鲍鱼快速分类提供新思路。

## 1　材料与方法

### 1.1　仪器及光谱采集参数

试验所用仪器为 MicroNIR™1700 便携式近红外光谱仪（JDSU，美国），测量波长范围为 900～1 700 nm，分辨率 6～10 nm，积分时间 8 ms，每 30 min校正 1 次背景和暗电流。为满足活体测量需要，在仪器测量端加装蓝宝石窗口。皱纹盘鲍、红壳鲍、绿盘鲍样品产地均为莆田，养殖时间为 18 个月，鲍鱼样品经过清水简单清洗后，将仪器贴于活体鲍鱼肉足中心进行光谱扫描，每个样品采集 3 次，最终采集皱纹盘鲍、红壳鲍、绿盘鲍光谱数分别为 105、93 和 75 条。光谱数据分析及数据处理软件为：The Unscrambler 9.7（CAMO，挪威），python 3.6，scikit-learn[29]。

### 1.2　数据分析及模型建立方法

将每一类鲍鱼的光谱分别乱序排列后按 2∶1 比例划分训练集和测试集，划分过程中保证同一个样本的 3 条光谱同属 1 个集合。对光谱数据进行标准化变换，采用 CART 算法建立决策树模型。

决策树算法是机器学习中经典的分类算法，通过学习简单的决策规则建立决策树模型，过程简单直观，有很强的解释性，应用范围十分广泛。一般而言，建立一个完整的决策树模型需要经过特征选择、决策树构建和剪枝 3 个过程。决策树主要有 3 种实现方法：ID3 算法，CART 算法和 C4.5 算法。本研究中选用的 CART 算法[30]是一种二分递归分割技术，即每次决策时只能选择“是”或者“否”，把数据分成 2 部分生成结构简洁的二叉树。

Gini 不纯度（Gini impurity）是 CART 算法中常用的特征选择判据。Gini 不纯度可以用于表示集合中元素的不纯度，即从 1 个集合中随机选取 1 个元素，以该集合中标签的概率分布为基础对元素分配标签，其标签错误的概率。因此，Gini 不纯度的计算可以表示为 1 减去所有分类正确的概率，Gini 不纯度越大，该集合中元素组成越不纯，即集合中不属于该标签的元素数越多；当 Gini 不纯度达到最小值 0 时，可以认为集合中的元素属于同一个类别。Gini 不纯度的计算式为：

$$I_G(f)=\sum_{i=1}^{J} f_i(1-f_i)=1-\sum_{i=1}^{J} f_i^2 \qquad (1)$$

式中：$I_G$ 为 Gini 不纯度；$J$ 为样本类别数；$i\in\{1,2,\cdots,J\}$；$f_i$ 为 $i$ 类元素在集合中的比例。

决策树在建模过程中很容易出现过拟合的现象，过拟合时模型的训练误差很小，但检验误差很大，不利于实际应用。本研究在建模时采用了交叉验证，避免过拟合现象发生。

本研究采用对测试集样本的预测准确率衡量决策树模型优劣，计算公式为：

$$a=\frac{n}{N} \qquad (2)$$

式中：$A$ 为准确率；$n$ 为测试集中被正确预测的样本数；$N$ 为测试集样本总数。

## 2　结果与讨论

### 2.1　鲍鱼近红外光谱

鲍鱼样品 273 条近红外光谱见图 1，训练集光谱共 180 条，测试集光谱共 93 条。样本集划分结果见表 1。

### 2.2　鲍鱼分类决策树模型建立及模型预测结果

鲍鱼分类决策树模型通过调用 python 中 sklearn 机器学习包的决策树模型建立，参数采用默认值。模型判据为 Gini 不纯度，分裂策略选择“最佳”，即在每个节点处选择 Gini 不纯度最大的方式进行分裂。

完全生长的决策树是过拟合的，因此需要经过剪枝才能更好地被用于预测。本研究中采用的策略是在建模阶段通过交叉验证来避免过拟合现象发生。

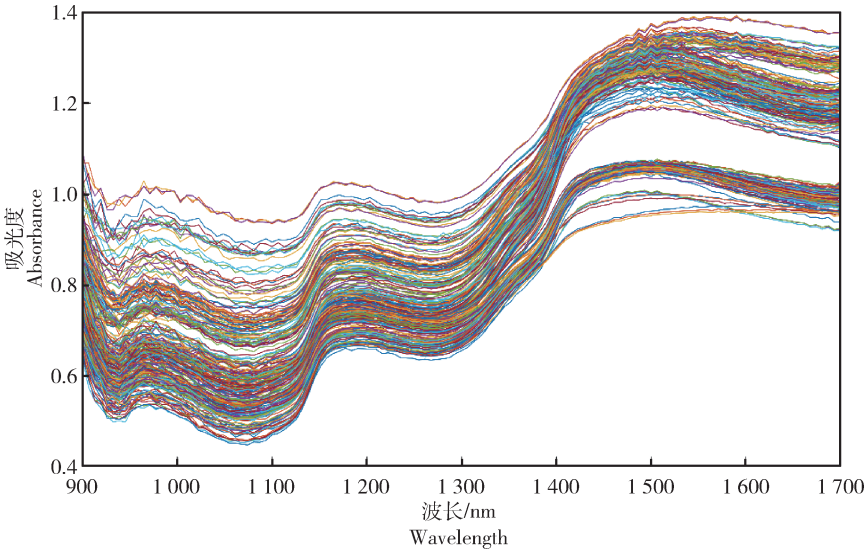



**图 1 鲍鱼样品近红外光谱**

Fig. 1 The near-infrared spectra of abalone samples

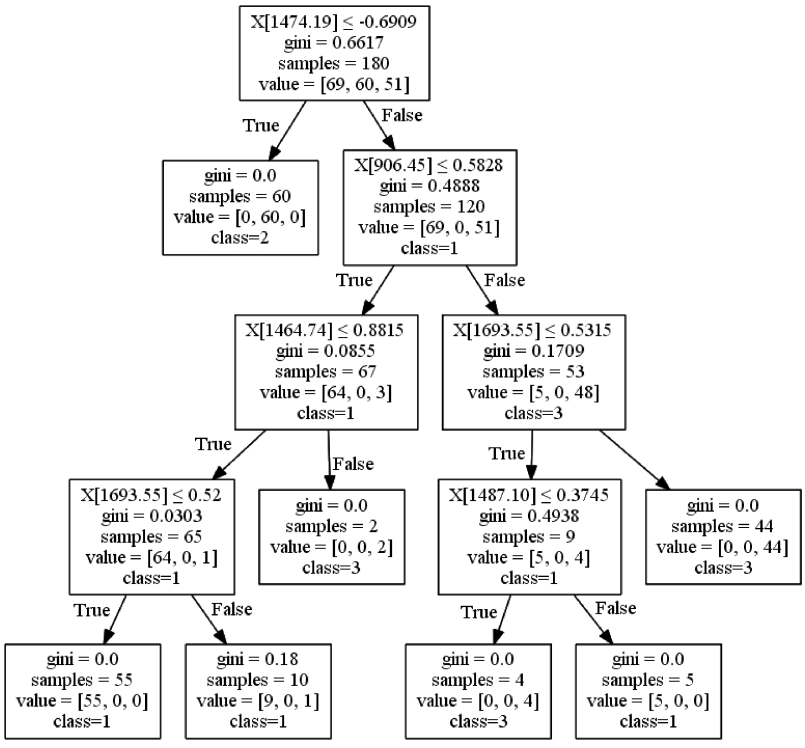
**表 1 样本集划分结果**

Table 1 Sample distribution

| 鲍鱼品种 Abalone (sub)species | 样本分布 Sample distribution | |
|---|---|---|
| | 训练集 Training set | 测试集 Testing set |
| 皱纹盘鲍 Haliotis discus hannai | 69 | 36 |
| 红壳鲍 Abalone with red shell | 60 | 33 |
| 绿盘鲍 Lvpan abalone | 51 | 24 |
| 合计 Total | 180 | 93 |

用训练集中 180 条光谱建立模型，采用 5 折交叉验证，得到的决策树见图 2，模型准确率为 90.00%。可以看出，CART 是以经过标准化变换后的光谱矩阵中对应位置的吸光度值为类别划分依据，从而实现节点的分裂和模型的建立。决策树从上至下，由左至右对应节点的波长分别为 1 474.19、906.45、1 464.74、1 693.55 和 1 487.10 nm，对应的吸光度值大小是鲍鱼种类判别的依据。可以认为，本研究所得的近红外光谱中，906 nm 附近 C-H 基团的三倍频、1 693 nm 附近 C-H 基团的一倍频以及 1 464～1 487 nm 附近蛋白质中 N-H 基团的一倍频或 C-H 基团的组合频对本研究涉及到的 3 种鲍鱼分类起主要作用。其中，通过波长 1 474 nm 附近吸光度值可将红壳鲍与另外 2 种鲍鱼区分开；通过波长 906 nm 附近的吸光度值可大致区分皱纹盘鲍和绿盘鲍，2 种鲍鱼的进一步分类需要通过判断 1 464～1 487 nm 附近以及 1 693 nm 附近吸光度值大小来实现。决策树中有 2 个波长 1 693 nm 的节点，且分别以吸光度值≤0.531 5 和≤0.520 0 为区分皱纹盘鲍和绿盘鲍的标志，可以认为，在波长 1 693 nm 附近，皱纹盘鲍对应近红外光谱的吸光度低于绿盘鲍，是区分皱纹盘鲍和绿盘鲍的重要判据之一。决策树中存在 Gini 不纯度为 0.18 但没有继续分裂的节点，说明决策树并未完全生长，未发生过拟合现象。

利用测试集样本对所得决策树模型的预测能力进行检验，结果表明，该模型对 93 个测试集样本的预测准确率为 90.32%，即有 84 个样本被正确分类，9 个被错误分类。




类别标签 1、2、3 分别代表皱纹盘鲍、红壳鲍和绿盘鲍。

Class 1, 2 and 3 are *Haliotis discus hannai*, abalone with red shell and Lvpan abalone, respectively.

**图 2　鲍鱼分类决策树**

Fig. 2　Classification of abalone by CART

## 3　结束语

本研究利用 MicroNIR™1700 便携式近红外光谱仪获取鲍鱼近红外光谱数据，经标准化变换后采用 CART 算法建立决策树模型，通过在建模时引入交叉验证的方法避免决策树过拟合现象发生。所得决策树模型准确率为 90.00%，对测试集样本预测准确率为 90.32%。本研究中所用仪器轻巧便携，整个过程操作简单，耗时短，对样品没有损害，预测结果令人满意，能够满足鲍鱼现场快速分类的需求。近红外光谱法结合机器学习算法为鲍鱼快速分类提供了一种新思路。

## 参考文献　References


[1] Borjesson T, Stenberg B, Schnurer J. Near-infrared spectroscopy for estimation of ergosterol eontent in barley: A comparison between reflectance and transmittance techniques [J]. *Cereal Chemistry*, 2016, 84(3): 231-236

[2] Fu H Y, Jiang D, Zhou R, Yang T M, Chen F, Li H D, Yin Q B, Fan Y. Predicting mildew contamination and shelf-life of sunflower seeds and soybeans by fourier transform near-infrared spectroscopy and chemometric data analysis[J]. *Food Analytical Methods*, 2017, 10(5): 1597-1608

[3] Lin M, Long M X, Li G L, Chen X, Zheng J, Li C, Kan J Q. Analysis of peanut using near-infrared spectroscopy and gas chromatography-mass spectrometry: Correlation of chemical components and volatile compounds[J]. *International Journal of Food Properties*, 2016, 19(3): 508-520

[4] Buyukcan M B, Kavdir I. Prediction of some internal quality parameters of apricot using FT-NIR spectroscopy[J]. *Journal of Food Measurement and Characterization*, 2017, 11(2): 651-659

[5] Mancini M, Rinnan, Pizzi A, Toscano G. Prediction of gross calorific value and ash content of woodchip samples by means of FT-NIR spectroscopy[J]. *Fuel Processing Technology*, 2018, 169: 77-83

[6] Devos O, Ruckebusch C, Durand A, Duponchel L, Huvenne J P. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation[J]. *Chemometrics and Intelligent Laboratory Systems*, 2009, 96(1): 27-33

[7] Balabin R M, Lomakina E I. Support vector machine regression (SVR/LS-SVM): An alternative to neural networks (ANN) for analytical chemistry: Comparison of nonlinear methods on near infrared (NIR) spectroscopy data[J]. *The*

*Analyst*, 2011, 136(8): 1703

[8] Chauchard F, Cogdill R, Roussel S, Roger J M, Bellon-Maurel V. Application of LS-SVM to non-linear phenomena in NIR spectroscopy: Development of a robust and portable sensor for acidity prediction in grapes[J]. *Chemometrics and Intelligent Laboratory Systems*, 2004, 71(2): 141-150

[9] Shen T T, Zou X B, Shi J Y, Li Z H, Huang X W, Xu Y W, Chen W. Determination geographical origin and flavonoids content of goji berry using near-infrared spectroscopy and chemometrics [J]. *Food Analytical Methods*, 2016, 9(1): 68-79

[10] Woodcock T, Downey G, O'Donnell C P. Confirmation of declared provenance of European extra virgin olive oil samples by NIR spectroscopy[J]. *Journal of Agricultural and Food Chemistry*, 2008, 56(23): 11520-11525

[11] Costa M C A, Morgano M A, Ferreira M M C, Milani R F. Analysis of bee pollen constituents from different Brazilian regions: Quantification by NIR spectroscopy and PLS regression[J]. *LWT-Food Science and Technology*, 2017, 80: 76-83

[12] Teye E, Huang X, Dai H, Chen Q. Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification [J]. *Spectrochimica Acta-Part A: Molecular and Biomolecular Spectroscopy*, 2013, 114: 183-189

[13] Yu J, Zhan J C, Huang W D. Identification of wine according to grape variety using near-infrared spectroscopy based on radial basis function neural networks and least-squares support vector machines [J]. *Food Analytical Methods*, 2017, 10(10): 3306-3311

[14] Siriphollakul P, Nakano K, Kanlayanarat S. Eating quality evaluation of Khao Dawk Mali 105 rice using near-infrared spectroscopy[J]. *LWT-Food Science and Technology*, 2017, 79: 70-77

[15] Reis M M, Martínez E, Saitua E, Rodríguez R, Pérez I, Olabarrieta I. Non-invasive differentiation between fresh and frozen/thawed tuna fillets using near infrared spectroscopy (Vis-NIRS)[J]. *LWT-Food Science and Technology*, 2017, 78: 129-137

[16] Kar S, Tudu B, Bag A K, Bandyopadhyay R. Application of near-infrared spectroscopy for the detection of metanil yellow in turmeric powder[J]. *Food Analytical Methods*, 2017(1): 1-12

[17] Wold J P, Veiseth-Kent E, Høst V, Løvland A. Rapid on-line detection and grading of wooden breast myopathy in chicken fillets by near-infrared spectroscopy[J]. *Plos One*, 2017, 12(3): e0173384

[18] Kutsanedzie F, Chen Q, Sun H, Cheng W. *In-situ* cocoa beans quality grading by near-infrared-chemodyes systems [J]. *Analytical Methods*, 2017,9(37):5455-5463

[19] Huang F R, Li Y P, Wu J, Dong J, Wang Y. Identification of repeatedly frozen meat based on near-infrared spectroscopy combined with self-organizing competitive neural networks [J]. *International Journal of Food Properties*, 2016, 19 (5): 1007-1015

[20] You H, Kim Y, Lee J H, Choi S. Classification of food powders using handheld NIR spectrometer [C]. In: *Ubiquitous and Future Networks*, 2017 *Ninth International Conference*. Piscataway: IEEE,2017: 732-734

[21] Shi J Y, Zhang F, Li Z H, Huang X W, Zou X B, Zhang W, Holmes M, Chen Y. Rapid authentication of Indonesian edible bird's nests by near-infrared spectroscopy and chemometrics[J]. *Analytical Methods*, 2017, 9(8):1297-1306

[22] Karunathilaka S R, Kia A F, Srigley C, Chung J K, Mossoba M M. Nontargeted, rapidscreening of extra virgin olive oil oroducts for authenticity using near-infrared spectroscopy in combination with conformity index and multivariate statistical analyses [J]. *Journal of Food Science*, 2016, 81(10):C2390-C2397

[23] Ma H L, Wang J W, Chen Y J, Cheng J L, Lai Z T. Rapid authentication of starch adulterations in ultrafine granular powder of Shanyao by near-infrared spectroscopy coupled with chemometric methods[J]. *Food Chemistry*, 2017, 215: 108-115

[24] Mendes T O, Rocha R A Da, Porto B L S, Oliveira M A L De, Anjos V D C Dos, Bell M J V. Quantification of extra-virgin olive oil adulteration with soybean oil: A comparative study of NIR, MIR, and raman spectroscopy associated with chemometric approaches [J]. *Food Analytical Methods*, 2015, 8(9): 2339-2346

[25] Li S F, Zhang X, Shan Y, Su D L, Ma Q, Wen R Z, Li J J. Qualitative and quantitative detection of honey adulterated with high-fructose corn syrup and maltose syrup by using near-infrared spectroscopy[J]. *Food Chemistry*, 2017, 218: 231-236

[26] Guelpa A, Marini F, du Plessis A, Slabbert R, Manley M. Verification of authenticity and fraud detection in South African honey using NIR spectroscopy[J]. *Food Control*, 2017, 73: 1388-1396

[27] Grassi S, Casiraghi E, Alamprese C. Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties[J]. *Food Chemistry*, 2018, 243: 382-388

[28] Mossoba M M, Azizian H, Fardin-Kia A R, Karunathilaka S R, Kramer J K G. First application of newly developed FT-NIR spectroscopic methodology to predict authenticity of extra virgin olive oil retail products in the USA[J]. *Lipids*, 2017, 52(5): 443-455

[29] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in Python [J]. *Journal of Machine Learning Research*, 2011,12:2825-2830

[30] Breiman L I, Friedman J H, Olshen R A, Stone C J. Classification and regression trees (CART)[J]. *Biometrics*, 1984, 40(3):358

责任编辑:刘迎春