电子鼻和随机森林算法快速鉴别野生与养殖日本真鲈

孙 永[1,2,3], 刘 楠[1,3], 李智慧[1,3], 马玉洁[1,3], 周德庆[1,3*]

(1. 中国水产科学研究院黄海水产研究所, 青岛 266071; 2. 上海海洋大学食品学院, 上海 201306;
3. 海洋国家实验室海洋药物与生物制品功能实验室, 青岛 266000)

摘 要: 目的 建立电子鼻和随机森林算法快速鉴别野生与养殖日本真鲈的分析方法。**方法** 采用来源确定且规格不同的日本真鲈, 利用电子鼻中 14 个金属氧化物传感器获取 53 份日本真鲈样本(养殖样本 25 份, 野生样本 28 份)的特征信号, 构建得到行 × 列为 53 × 15(含标签列, 野生为 1, 养殖为-1)的初始特征矩阵。构建随机森林(randomforest, RF)模型, 并依据袋外错误率(out-of-bag error rate, OOB)对随机森林模型的估计器(决策树)数量和单一决策树最大特征的 2 个参数进行优化。**结果** 模型最优估计器数为 50, 最大特征数为 14, 模型的鉴别准确率达到 98.2%。通过该模型, 以贡献率为指标, 对电子鼻传感器进行了特征筛选和排序, 其中 S14 和 S4 传感器的贡献率分别为 42.9%和 36.0%。**结论** 该技术可以快速鉴别野生和养殖日本真鲈。

关键词: 电子鼻; 随机森林; 鉴别; 日本真鲈; 特征筛选

Rapid identification of wild and farmed *Lateolabrax japonicus* by electronic-nose technology and random forest algorithm

SUN Yong[1,2,3], LIU Nan[1,3], LI Zhi-Hui[1,3], MA Yu-Jie[1,3], ZHOU De-Qing[1,3*]

(1. *Yellow Sea Fishery Research Institute*, *Chinese Academy of Fishery Sciences*, *Qingdao* 266071, *China*; 2. *College of Food Science and Technology*, *Shanghai Ocean University*, *Shanghai* 201306, *China*; 3. *Laboratory for Marine Drugs and Bioproducts of Qingdao National Laboratory for Marine Science and Technology*, *Qingdao* 266000, *China*)

ABSTRACT: Objective To establish a rapid identification method for wild and farmed *Lateolabrax japonicus* by electronic-nose technology and random forest algorithm. **Methods** Using *Lateolabrax japonicus* of different sizes with confirmed original materials, feature signals of 53 seabass samples (25 farmed samples, 28 wild samples) were obtained by 14 metal-oxides semiconductor sensors of the electronic-nose. An initial feature matrix formed with row × column as 53 × 15(labels column included, 1 for wild, -1 for farmed). A random forest model (RF) was constructed, and 2 parameters (estimator number of the RF model and max features of individual decision tree) were optimized according to out-of-bag error rate (OOB). **Results** The best estimator number was 50, the max feature was 14, and the identification accuracy of the model was 98.2%. According to the model, taking contribution rates as index, the electronic nose senor was selected and ranked, the contribution of S14 and S4 for the identification was 42.9% and 36.0%, respectively. **Conclusion** This method can rapidly identify the wild and farmed *Lateolabrax japonicus*.

基金项目: 中央级公益性科研院所基本科研业务费项目(2016HY-ZD0801)

Fund: Supported by Central Public-interest Scientific Institution Basal Research Fund, CAFS (2016HY-ZD0801)

***通讯作者:** 周德庆, 研究员, 主要研究方向为水产品加工与质量安全。E-mail: zhoudq@ysfri.ac.cn

***Corresponding author:** ZHOU De-Qing, Professor, Yellow Sea Fishery Research Institute, Chinese Academy of Fishery Sciences, No.106 Nanjing Road, Qingdao 266071, China. E-mail: zhoudq@ysfri.ac.cn

KEY WORDS: electronic nose; random forest; identification; *Lateolabrax japonicus*; feature screening

1 引 言

水产品营养丰富、味道鲜美, 深受消费者喜爱。随着渔业科学不断发展, 我国水产品在供应量和种类方面不断增长, 养殖水产品所占的比例也在逐年上升。但野生和养殖水产品在品质上差异明显[1], 一方面, 由于饵料、养殖环境等因素的影响, 养殖水产品存在脂肪含量高、土腥味重等诸多问题, 在风味、口感方面远逊于野生品种[2], 另一方面由于养殖密度和管理方面的问题, 国内养殖水产品经常出现药残超标问题[3], 这些都严重影响了消费者购买养殖水产品的意愿, 加之近年来不断加剧的近海水产资源的枯竭问题, 造成了野生和养殖水产品较大的价格差异。市场上时常有不法商贩以养殖水产品冒充野生水产品, 损害消费者权益的情况发生。鉴别野生和养殖水产品, 专业的分析手段[4,5]往往存在周期长、费用高、污染环境的问题。

电子鼻作为一种新兴仿生技术, 在食品检测中逐渐受到重视, 在饮料[6,7]、粮油[8]、肉类[9]、水产[10]和果蔬类[11,12]等方面的研究多有报道, 表明电子鼻在食品检测上有很好的应用前景。随着新材料、传感器、信息技术的不断涌现, 电子鼻技术的发展也日新月异, 新型气体传感器的应用及其响应特性的研究, 为特征气味的检测与应用提供了重要的基础[13]。随机森林是一种新的机器学习算法, 可以在内部实现交叉验证, 具有分析复杂相互作用分类特征的能力, 对于噪声数据和存在缺失值的数据具有很好的鲁棒性, 并且具有较快的学习速度, 近年来已经被广泛应用于各种分类和特征选择问题中[14-17]。

日本真鲈(*Lateolabrax japonicus*), 又称海鲈鱼、七星鲈, 是我国一种重要的经济鱼类, 据统计, 2017 年我国海水养殖鱼类中, 鲈鱼的产量达 15.66 万吨, 仅次于大黄鱼[18]。本研究建立了电子鼻和随机森林算法快速鉴别的野生与养殖日本真鲈的方法, 以期为快速鉴别野生和养殖水产品提供技术支撑。

2 材料与方法

2.1 材料与试剂

日本真鲈的野生样本分别在不同时间从山东省日照市岚山区小型渔船上取得, 共 28 条, 样本重量范围 1025~1769 g, 养殖样本在不同时间从青岛市南山市场和新贵都市场购得, 共 25 条, 重量范围 1230~1693 g, 样品在购买时均具备较好的新鲜度, 覆冰保鲜运至实验室, -40 ℃条件下冻藏备用。

2.2 仪器与设备

iNose 电子鼻(美国 isenso 公司, 配备 14 个具有不同性质的金属氧化物半导体传感器组合成传感器阵列); BSA124S-CW 电子天平(德国 Sartorius 公司); ZHSY-50N 水浴锅(上海知楚公司); DW-40L92 -40 ℃低温保存箱(中国海尔公司)。

2.3 实验方法

2.3.1 样品制备

样品自-40 ℃冰箱取出后置于 4 ℃冰箱解冻过夜。取鱼背部相同位置的肌肉 2.0 g 置于 20 mL 进样瓶中, 加盖密封, 40 ℃水浴条件下顶空 20 min 后上机测定。每条日本真鲈为一个样本, 每个样本 3 个平行。

2.3.2 电子鼻分析

电子鼻参数: 气体流量 1 L/min, 数据采集时间 120 s, 间隔清洗时间 120 s, 采集各传感器的响应信息, 选择最大值作为特征值。

2.3.3 鉴别模型构建

随机森林(random forest, RF)是由 Breiman[19]将其在 1996 年提出的 Bagging (Bootstrap Aggregating)集成学习理论和 Ho[20]在 1998 年提出的随机子空间方法相结合提出的一种机器学习算法[21]。随机森林是以决策树(decision tree)为基本分类器的一个集成学习模型, 它包含多个由 Bagging 集成学习技术训练得到的决策树, 当输入待分类的样本时, 最终的分类结果由单个决策树的输出结果投票决定。随机森林构建过程如如图 1 所示。

构建每个决策树时, 随机抽取训练样本集和属性子集的过程都是独立的, 且总体都是一样的, 因此图 1 中的 $\{\theta_k,\ k=1,2,...,K\}$是一个独立同分布的随机变量序列。训练随机森林的过程就是训练各个决策树的过程, 由于各个决策树的训练是相互独立的, 因此随机森林的训练可以通过并行处理来实现, 因而可以大大提高生成模型的效率[22]。

本研究中 RF 的构建由 Python3.6 通过调用 numpy、matplotlib 和 scikit-learn 科学计算包实现。为提高模型的泛化能力, 对随机森林模型参数进行调整。随机森林分类器的参数包括框架参数和决策树参数两大部分, 考虑到本文中样本数量和特征变量的规模较小, 因此只对模型中决策树的个数、生成决策树的最大特征数进行调参。

假设某样本的总体分布 X 未知, 有来自此样本集的一个数据样本容量为 N, 每次有放回抽取的 Bootstrap 样本大小也为 N, 那么每个样本未被抽中的概率约为 $\left(1-\frac{1}{N}\right)^N$, 当 N 很大时, 这个概率值趋于 1/e≈0.368, 这表明每次抽样

时原样本集中约有 37%的样本未被抽中，未抽中的这部分样本称为袋外数据(out-of-bag data)[21]。假设袋外数据总数为 P，用这 P 个袋外数据作为已经生成的分类器的输入，分类器会给出 P 个数据相应的分类标签，因为这 P 条数据的分类标签是已知的，则用正确的分类与分类器的结果进行比较，统计分类器分类错误的数目，设为 E，则 E/P 即为袋外错误率(out-of-bag error, OOB error)。本研究主要依据计算袋外错误率对估计器数量进行优化。

另外，随机森林算法一个重要的优势是无需对模型进行交叉验证或者用一个独立的测试集来获得误差的一个无偏估计。它可以在内部进行评估，即在生成的过程中就能够对误差建立一个无偏估计，这是由于未被选中的袋外数据可用于估计模型单棵决策树分类强度、决策树之间的相关性，进而可得到模型泛化误差界的估计。很多研究证明袋外估计是无偏估计，袋外估计与用同训练集相同样本的测试集估计的精度是一致的[23]。

3　结果与分析

3.1　日本真鲈电子鼻测定结果

经电子鼻测定，获得每个日本真鲈样本的信息数据，取 3 个平行样本每个传感器最大响应的平均值，得到电子鼻雷达图，如图 2 所示。设定养殖样品的标签为-1，野生样品的标签为 1，形成样本为行，特征值为列的数据矩阵(参见表 1)，作为输入(不含标题行)进行随机森林模型的构建和优化。

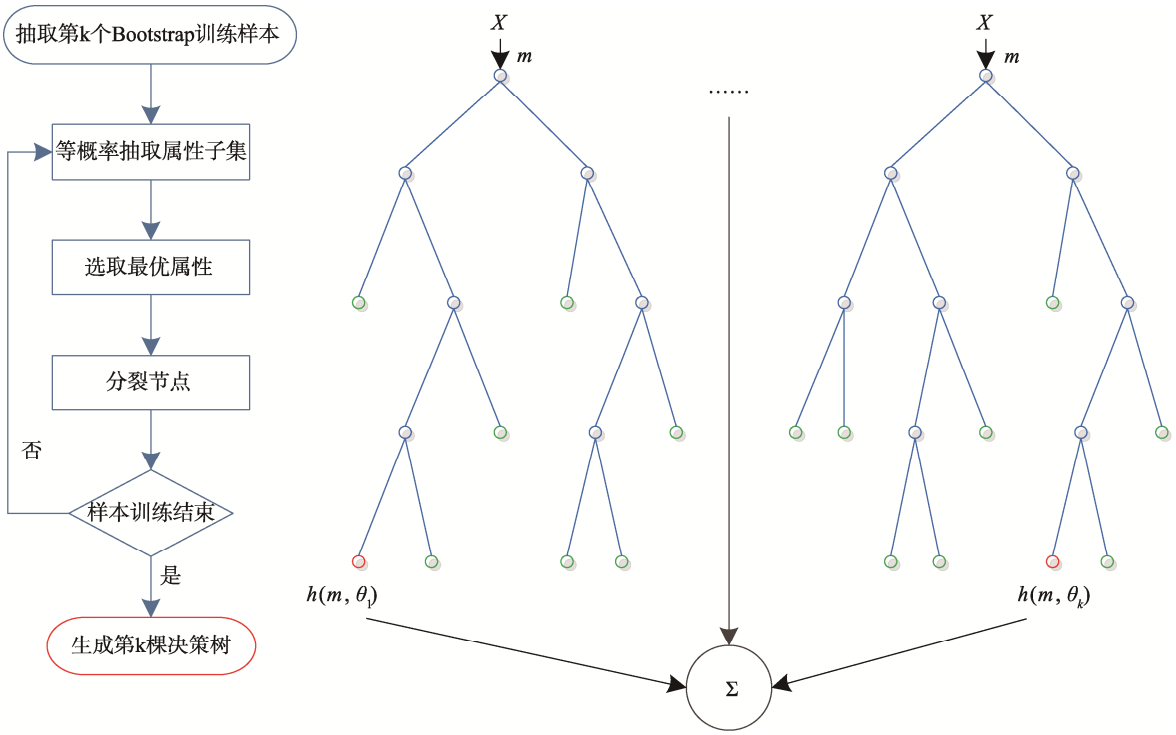

图 1　RF 模型构建示意图

Fig. 1　A schematic illustration of random forest model construction

表 1　电子鼻数据矩阵表

Table 1　Matrix of e-nose data

标签	S1	S2	S3	S4	……	S11	S12	S13	S14
-1	1.196996	1.072978	1.038741	1.08287	……	1.001634	0.943614	0.889187	1.02458
-1	1.215094	1.057456	1.036246	1.073913	……	0.989256	0.929726	0.872234	1.006984
	……								
1	1.272876	1.272046	1.066487	1.194307	……	1.218614	1.080785	1.024777	0.995156
1	1.158023	1.124419	1.032929	1.110488	……	1.055188	0.983574	0.939528	0.991672
	……								

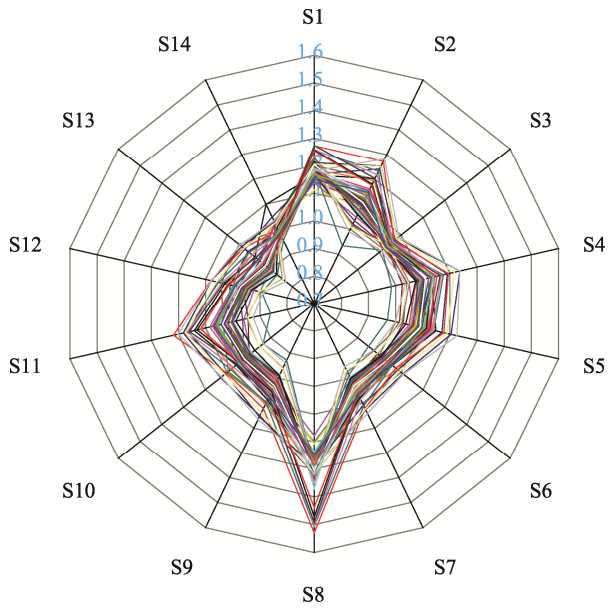

图 2 电子鼻数据雷达图

Fig.2 Radar map of Enose data

3.2 模型参数的优化

每条样本的电子鼻数据由 14 个传感器获得的最大值组成，即样本特征数为 14，因此分别以 2、6、10、14 作为最大特征数，估计器数量范围设定在 10~200 之间，以袋外错误率为指标，对模型进行优化，结果如图 3 所示。

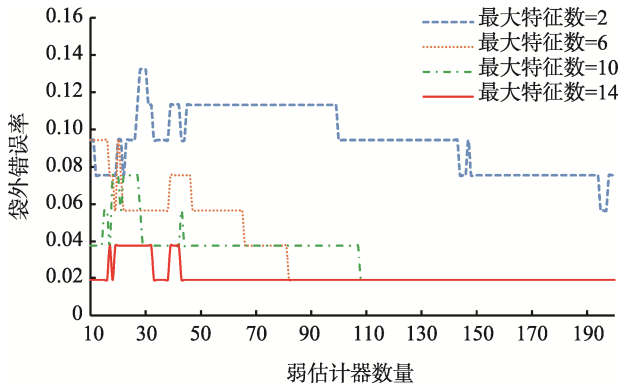

图 3 随机森林模型参数优化曲线

Fig. 3 Optimization curves of random forest models

从图 3 可以看出，模型随着最大特征数和估计器数的增大，其袋外错误率不断下降，除最大特征为 2 的分类器之外，其他 3 个分类器随着估计器个数的增加，袋外错误率不断下降并最终稳定在 0.018，即模型的准确率为 98.2%。对于随机森林模型来说，估计器数量太少，模型容易欠拟合，数量越大，硬件计算量也会越大，且估计器达到一定数量后，再增大对模型的提升很小[21]，所以需要选择一个适中的值。综上，最大特征选择为 14，估计器数为 50，建立模型，并以此对电子鼻 14 个传感器数据进行特征信号筛选。

3.3 特征信号的筛选

由于 OOB 的存在，随机森林建模过程可以直接计算特征变量重要性(即对分类的贡献率)[24,25]。模型中，特征筛选的核心思想是随机检测，其方法是对于某个特征，如果用另外一个随机值代替它，其 OOB 变大，说明该特征比较重要，所占权重应该较大，不能用一个随机值代替；反之，如果随机值替代后，其 OOB 没有太大差别，说明该特征重要性较小。通过随机森林模型计算得到电子鼻 14 个特征对模型分类的重要性，并进行排序，结果如图 4 所示。

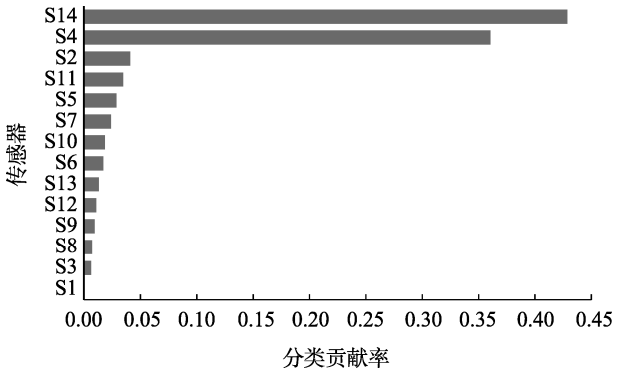

图 4 电子鼻特征对模型分类的贡献率

Fig. 4 Features importance for the classification

从图 4 可以看出，S14 和 S4 传感器对分类的贡献率分别为 42.9%和 36.0%，占所有传感器贡献率的 78.9%，对分类起决定性作用。iNose 电子鼻系统的 S14 传感器对内酯和吡嗪类化合物响应敏感，而 S4 传感器对酯类和萜类化合物响应敏感，说明养殖和野生日本真鲈的气味差异可能主要由上述几类化合物造成。土腥味重是养殖水产品的品质缺陷之一，国内外大量研究指出，2-异丙基-3-甲氧基吡嗪(2-isopropyl-3-methoxy pyrazine, IPMP)和 2-异丁基-3-甲氧基吡嗪(2-isobutyl-3-methoxy pyrazine, IBMP)等吡嗪类化合物是养殖水产品产生土腥味的重要因素[26-28]，说明本研究 RF 模型对电子鼻数据的特征筛选结果与水产品感官特征具有较高的一致性。

4 结　论

采用电子鼻技术结合随机森林算法, 以日本真鲈为研究对象, 建立了野生和养殖水产品的快速鉴别分类模型。研究表明: 1)首先, 利用日本真鲈的电子鼻信号初始特征建立矩阵模型, 其次, 采用袋外错误率对随机森林的估计器数量和单一决策树最大特征对模型进行了训练优化, 得到最优估计器数为 50, 最大特征数为 14, 模型的鉴别准确率达到 98.2%; 2)通过模型筛选得到电子鼻 14 个传感器对分类的贡献率, 其中 S14 和 S4 传感器的贡献分别为 42.9%和 36.0%, 对野生和养殖日本真鲈的鉴别起决定性作用。本研究为快速鉴别野生和养殖水产品提供了技术支撑, 同时也为开发基于气体传感器的低成本快速鉴别装置提供了理论依据。

参考文献

[1] Claret A, Guerrero L, Gartzia I, *et al*. Does information affect consumer liking of farmed and wild fish? [J]. Aquaculture, 2016, (454): 157–162.

[2] 李智慧, 孙永, 史建如, 等. 野生与养殖许氏平鲉品质的比较[J]. 食品工业科技, 2017, 38(8): 87–91.
Li ZH, Sun Y, Shi JR, *et al*. Comparative of quality of wild-captured and farmed *Sebastes schlegeli* [J]. Sci Technol Food Ind, 2017, 38(8): 87–91.

[3] 王玉堂, 吕永辉. 水产养殖用药与水产品质量安全[J]. 农业工程, 2011, 1(3): 44-49.
Wang YT, Lv YH. Aquaculture drugs and safety of aquatic products [J]. Agric Eng, 2011, 1(3): 44–49.

[4] 何杰, 吴旭干, 龙晓文, 等. 池塘养殖和野生长江水系中华绒蟹扣蟹形态学及生化组成的比较研究[J]. 水产学报, 2015, 39(11): 1665–1678.
He J, Wu XG, Long XW, *et al*. Comparative studies of morphology and biochemical composition between wild-caught and pond-reared juvenile Chinese mitten crab for Yangtze population [J]. J Fish China, 2015, 39(11): 1665–1678.

[5] Wang YV, Wan AHL, Lock E, *et al*. Know your fish: A novel compound-specific isotope approach for tracing wild and farmed salmon [J]. Food Chem, 2018, (256): 380–389.

[6] Jin J, Deng S, Ying X, *et al*. Study of herbal tea beverage discrimination method using electronic nose [J]. J Food Meas Charact, 2015, 9(1): 52–60.

[7] Tudu B, Ghosh S, Bag AK, *et al*. Incremental FCM technique for black tea quality evaluation using an electronic nose [J]. Fuzzy Inf Eng, 2015, 7(3): 275–289.

[8] Jonsson A, Winquist F, Schnürer J, *et al*. Electronic nose for microbial quality classification of grains [J]. Int J Food Microbiol, 1997, 35(2): 187–193.

[9] Wojnowski W, Majchrzak T, Dymerski T, *et al*. Electronic noses: Powerful tools in meat quality assessment [J]. Meat Sci, 2017, (131): 119–131.

[10] 赵梦醒, 丁晓敏, 曹荣, 等. 基于电子鼻技术的鲈鱼新鲜度评价[J]. 食品科学, 2013, 34(6): 143–147.
Zhao MX, Ding XM, Cao R, *et al*. Identification of *Lateolabrax japonicas* freshness by electronic nose [J]. Food Sci, 2013, 34(6): 143–147.

[11] Qiu S, Wang J. The prediction of food additives in the fruit juice based on electronic nose with chemometrics [J]. Food Chem, 2017, (230): 208–214.

[12] Chen H, Zhang M, Bhandari B, *et al*. Evaluation of the freshness of fresh-cut green bell pepper (*Capsicum annuum* var. grossum) using electronic nose [J]. LWT-Food Sci Technol, 2018, (87): 77–84.

[13] 王俊, 崔绍庆, 陈新伟, 等. 电子鼻传感技术与应用研究进展[J]. 农业机械学报, 2013, 44(11): 160–167.
Wang J, Cui SQ, Chen XW, *et al*. Advanced technology and new application in electronic nose [J]. Trans Chin Soc Agric Mach, 2013, 44(11): 160–167.

[14] Salles T, Gonçalves M, Rodrigues V, *et al*. Improving random forests by neighborhood projection for effective text classification [J]. Inform Syst, 2018, (77): 1–21.

[15] Vigneau E, Courcoux P, Symoneaux R, *et al*. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception [J]. Food Qual Prefer, 2018, (68): 135–145.

[16] Huang C, Ma YH, Zhao HB, *et al*. Spectral classification of asteroids by random forest [J]. Chin Astronom Astropys, 2017, 41(4): 549–557.

[17] Edla DR, Mangalorekar K, Dhavalikar G, *et al*. Classification of EEG data for human mental state analysis using random forest classifier [J]. Proced Comput Sci, 2018, (132): 1523–1532.

[18] 中华人民共和国农业农村部渔业渔政管理局. 2018 年中国渔业统计年鉴[M]. 北京: 中国农业出版社, 2018.
Bureau of Fisheries, Ministry of agriculture and rurual affairs of the People's Republic of China. China fisheries statistics yearbook 2018 [M]. Beijing: China Agriculture Press, 2018.

[19] Breiman L. Bagging predictors [J]. Mach Learn, 1996, 24(2): 123–140.

[20] Ho T. The random subspace method for constucting decision forests [J]. IEEE Trans Patt Anal Mach Intell, 1998, 20(8): 832–844.

[21] Breiman L. Random forests [J]. Mach Learn, 2001, 45(1): 5–32.

[22] 董师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013, 2(1): 1–7.
Dong SS, Huang ZX. A brief theoretical overview of random forests [J]. J Integr Technol, 2013, 2(1): 1–7.

[23] Wolpert DH, Macready WG. An efficient method to estimate bagging's generalization error [J]. Mach Learn, 1999, 35(1): 41–55.

[24] Genuer R, Poggi J, Tuleau-Malot C. Variable selection using random forests [J]. Patt Recogn Lett, 2010, 31(14): 2225–2236.

[25] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. J Mach Learn Res, 2003, (3): 1157–1182.

[26] Lovell RT. Off-flavors in pond-cultured channel catfish [J]. Water Sci Technol, 1983, 15(6–7): 67–73.

[27] Mahmoud MAA, Buettner A. Characterisation of aroma-active and off-odour compounds in German rainbow trout (*Oncorhynchus mykiss*).

Part II: Case of fish meat and skin from earthen-ponds farming [J]. Food Chem, 2017, (232): 841-849.

[28] Chen X, Luo Q, Yuan S, *et al*. Simultaneous determination of ten taste and odor compounds in drinking water by solid-phase microextraction combined with gas chromatography-mass spectrometry [J]. J Environ Sci, 2013, 25(11): 2313-2323.

(责任编辑：陈雨薇)

作者简介

孙　永，硕士，助理研究员，主要研究方向为水产品加工与质量安全。
E-mail: sunyong@ysfri.ac.cn

周德庆，博士，研究员，博士生导师,主要研究方向为水产品加工与质量安全。
E-mail: zhouqd@ysfri.ac.cn