第40卷，第9期
2020年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 40，No. 9，pp2809-2814
September，2020

基于氨基酸的仿刺参产地信息认证方法研究

吴 鹏[1,2]，李 颖[1,2*]，刘 瑀[2,3]，陈 晨[1,2]，冉明衢[1,2]，李亚芳[1,2]，赵新达[3]

1. 大连海事大学航海学院，辽宁 大连 116026
2. 大连海事大学环境信息研究所，辽宁 大连 116026
3. 大连海事大学环境科学与工程学院，辽宁 大连 116026

摘 要 仿刺参富含多种活性物质，具有极高的药用价值和经济价值，是水产行业不可或缺的养殖资源。不同产地的地理环境与营养结构存在显著差异，所以仿刺参的生长周期与养殖成本相差巨大。消费者在购买仿刺参时，会将产地信息作为选择的首要因素，因为仿刺参的产地直接反映了食品具有的营养价值。不同产地仿刺参的价格差距悬殊，面对利益的诱惑，产地欺诈事件屡禁不止。因此，研究一种准确率高、稳定性好具有优秀泛化能力的仿刺参产地信息认证方法，能够有效维护品牌产地从业者与消费者的切身利益。氨基酸是仿刺参营养富集的主要物质，通过氨基酸特征能够分析出摄食初级生产者的组成，可以作为产地信息认证的有效工具。气相色谱-质谱分析(GC-MS)技术能够产生独特的化学指纹图谱用于产地信息鉴别。对9个产地的156个仿刺参样品，进行酸水解、衍生化和酯化等操作，通过GC-MS测定出氨基酸含量与氨基酸碳稳定同位素数据。进行置信水平为95%的图基检验，并利用箱型图检查数据分布，筛选出13种氨基酸含量和10种氨基酸碳稳定同位素数据。主成分分析能够在降低数据维度的同时，挖掘出有价值的信息，聚集产地识别特性，提高运算速度与认证精度。通过交叉验证，选取前5个主成分作为氨基酸含量和氨基酸碳稳定同位素模型的输入，累计贡献率分别为98.727%与95.982%。为了充分挖掘出隐藏在氨基酸数据背后的价值，选取了8个家族的12个机器学习方法，共构建出24个单体分类器，根据数据自身特征找到最优的认证方法。应用基于遗传交叉因子改进的粒子群优化算法进行模型参数的优化，得到性能最佳的单体分类器。结果表明，氨基酸碳稳定同位素数据具有更优的产地认证特性，高斯径向基为核的支持向量机与K邻近算法为最佳的两个分类方法。最后利用集成学习汇集单体分类器的优势，构建了一种融合多源数据处理方法的仿刺参产地信息认证方法，模型的平均准确率为99.67%。建立了仿刺参产地信息认证系统，为主管机关监管与消费者维权提供了简单可行的手段，能够有效监管与防止仿刺参产地欺诈事件的发生，保障了水产养殖行业的稳健发展。

关键词 仿刺参；产地认证；氨基酸；主成分分析；机器学习；集成学习

中图分类号：O657.7 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)09-2809-06

引 言

仿刺参(Apostichopus japonicus)是海参纲(Holothuroidea)中最具营养价值与经济价值的一类[1]。仿刺参体内富含皂苷等高活性物质，具有抗肿瘤，降低血脂，改善非酒精性脂肪肝，抑制脂肪堆积，抗高尿酸血症，促进骨髓造血，抗高血压等医学功效[2]。2017年中国共计养殖仿刺参538亿头，年产量219 907 t，行业总产值超过40亿美金。食品欺诈是一种极其有利可图的行为，不法商贩通过不正当手段误导，甚至直接欺骗消费者，从而获取不法暴利[3]。通过地理标志产品保护规定的设立，可以有效保护质量、特色和声誉取决于其产地地理特征的食品，提升优质产地食品的经济价值[4]。尽管法规的设立能够预防食品产地欺诈事件的发生，但面对高额的利益诱惑，食品欺诈事件屡禁不止[5]。

氨基酸是蛋白质的基本组成单位，细胞的一切新生、修复与更新都与氨基酸息息相关。生物体中的氨基酸含量直接体现了其富含的营养价值，不同种类氨基酸的含量反映了其

收稿日期：2019-07-25，修订日期：2019-12-02
基金项目：国家科技支撑计划项目(2015BAD17B05)，国家海洋公益类专项(201305002)资助
作者简介：吴 鹏，1994年生，大连海事大学航海学院硕士研究生 e-mail：18840866641@163.com
*通讯联系人 e-mail：yldmu@126.com

摄食初级生产者的种类与比例[6]。与脂肪酸相比，仿刺参体内含有更多的氨基酸，氨基酸中的碳元素约占到仿刺参总碳量的一半，是仿刺参新陈代谢活动的最主要参与者[7]。特定化合物同位素分析技术(CSIA)结合了稳定同位素分析技术和特定化合物组成分析技术的双重优势，可以更精确地阐述海洋食物网中营养物质的流动路径[8]。特定化合物的碳稳定同位素特征提供了一种更加深入理解营养物质富集的手段，在食品产地信息认证领域取得了良好的效果[9]。

本研究提出了一种融合多源数据处理方法认证仿刺参产地信息的新方法。通过充分发挥不同描述角度数据的价值，使其挖掘出数据背后隐含的规律，建立了准确性更高、稳定性更好、体系架构更完善的产地信息认证模型。构建了仿刺参产地信息认证系统，有效地监管与防止食品产地欺诈事件的发生，维护品牌产地从业者与消费者的切身利益。

1 实验部分

1.1 样品

仿刺参样品采集于 2015 年 11 月，共采集到有效样品 156 个：其中氨基酸含量样品 78 个，氨基酸碳稳定同位素样品 78 个。共包括长海县(CH)、獐子岛(ZZD)、霞浦(XP)、普兰店(PLD)、瓦房店(WFD)、威海(WH)、担子岛(DZD)、莱州(LZ)和牟平(MP)9 个产地的样品。仿刺参的体长范围 15～19 cm，体重范围 100～130 g，霞浦样品的参龄为 1 年，其他 8 个产地的样品参龄均为 2 年。样品捕捞后立即存储在无菌塑料袋中，采用 4 ℃恒温冷藏，防止其因高温产生自溶酶而水解。在实验室内解剖去除沙石、内脏和石灰环，留取体壁并用超纯水洗净，冷冻干燥 48 h 后用玻璃研钵磨制粉末状，过 80 目网筛并干燥保存。

1.2 数据测定

取仿刺参样本 20 mg 放入 pyrex(耐高温)试管中，加入 2 mL 的 6 mol·L^{-1} HCl 溶液，向试管中充 N_2 1 min 去除空气，在 110 ℃恒温密闭条件下酸水解 24 h[10]。水解液在 430 g 条件下离心 10 min，取上清液注入强阳离子交换柱，提取出纯化氨基酸。由于氨基酸为两性离子不易挥发，而气相色谱分析需要对象具有良好的挥发性，因此采用 Metges 改进的方法将氨基酸衍生化成对应的 N-新戊酰基-O-异丙醇酯(NPP)[11]。向冷却后的 NPP 中加入 2 mL CH_2Cl_2，将混合物逐滴通过 6 cm 硅胶(200～400 目)层析柱(内径 4 mm)，去除多余的酰化剂等杂质。在室温下用 N_2 将滤液吹干，得到纯化的 NPP，最后将其溶于 0.2 mL 乙酸乙酯中。

取 1 μL 氨基酸酯化溶液通过气相色谱仪，色谱分离(GC)条件为：采用无分流方式进样，进样口温度 280 ℃；初始加热至 70 ℃并保持 1 min，以 3 ℃·min^{-1} 的速度加热至 220 ℃，再以 10 ℃·min^{-1} 的速度加热至 300 ℃并保持 8 min，最后以 1.2 mL·min^{-1} 的恒定流速充入纯度≥99.999%的 He 作为载气。气相色谱分离后酯化氨基酸再经过气质联用仪进行质谱分析，质谱分析(MS)条件为：传输线温度 250 ℃；离子源温度 230 ℃；通过能量为 70 eV 的 EI 电子进行电离。最终由 GC-MS 实验得到 GC 保留时间和 MS 谱图，与标准谱库(NIST2008)进行比较，确定出氨基酸的种类，并计算得到每种氨基酸的含量数据。测定氨基酸碳稳定同位素数据时，酯化氨基酸色谱分离后，其中 1/10 通过气质联用仪，得到 GC 保留时间和 MS 谱图，确定出氨基酸的种类；剩余 9/10 进入稳定同位素比质谱仪，测定出相应氨基酸的碳稳定同位素数据。

1.3 处理方法

现有食品产地认证方法的研究多侧重于化学计量工具方面，在数据处理方法上只停留在简单运用已有方法进行产地分类的层面，受制于样本数量与食品实际数量的巨大差距，将产地认证方法推广到尚未测量的数据时会存在明显偏差。当研究人员选择处理方法时，会选择一种他所期望的“最佳”分类方法，而不是从数据自身特征的角度进行最优方法的选取[12]。受制于对可用方法上的知识限制与数据特征的不确定性，选取单一认证方法无法充分发挥出数据的价值。以深度神经网络为例，需要通过对大量数据的不断训练，才能展现出神奇的分类效果，而食品产地信息认证领域往往解决的是小样本问题，采用结构简单的机器学习方法，得到的认证结果会更加准确。

在进行数据处理方法选择时，遵从没有免费午餐理论(NFLT)，即针对食品产地认证领域的所有问题，所有方法的期望是相等的，没有任何一种方法可以表现得比其他方法更好[13]。为了充分的挖掘出隐藏在数据背后的价值，采用来自 8 个家族的 12 个机器学习方法进行数据处理，利用数据自身特征主动计算出最佳方法，消除人为选择的干扰。在经过不断训练与优化得到最佳分类方法之后，没有直接采用奥卡姆剃刀原则，选择性能最佳且最简单的分类算法进行产地信息的认证；而是将不同分类方法建立出的模型，采用集成学习构建出一个泛化能力更强的产地信息认证整体。

1.4 认证模型

认证模型由样品预处理、数据测定、主成分分析、分类方法建立、模型优化、认证方法集成和在线系统构建 7 部分组成，整体结构如图 1 所示。

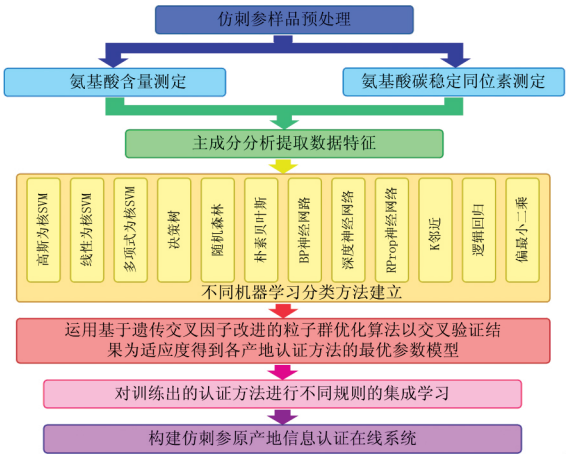

图 1 认证模型的整体结构

Fig. 1 The structure of the authentication model

1.4.1　主成分分析提取特征

主成分分析(PCA)是数据发掘领域常用的一种统计与降维算法，利用彼此不相关的综合变量代替数量较多的原始变量，在降低维度的同时保留数据自身有价值的信息。通过总方差不变的线性变换，提取出最具产地识别特性的氨基酸类别，去除随机误差影响，聚集产地特征，提高模型的运算速度与计算精度。

1.4.2　机器学习分类方法选择

为了充分发挥数据自身的价值，选择一定数量且具有足够广泛代表性的分类方法。利用 Manuel 等在 UCI 数据库中 121 个数据集上对 17 个分类器家族的评估结果，选取了最优的 8 个家族 12 个分类方法进行认证模型的训练[12]。选择的 12 个分类方法为：高斯径向基为核的支持向量机、线性为核的支持向量机、多项式为核的支持向量机、决策树、随机森林、朴素贝叶斯、BP 神经网路、深度神经网络、RProp 神经网络、K 邻近、逻辑回归与偏最小二乘。

1.4.3　交叉验证与粒子群优化算法

交叉验证是一种预测在未知数据上表现的模型评价方法。通过交叉验证可以有效了解模型的准确率、稳定性和对新样本的泛化能力，挑选出性能最优的分类器与模型参数，预防与限制过拟合与欠拟合的发生，挖掘出有限数据背后隐藏的价值。

采用马尔科夫蒙特卡洛(MCMC)方法进行训练数据的采样，在进行不同 K 值 100 次交叉验证前生成一条马尔科夫链使其收敛至平稳分布，保证待采样的数据符合后验分布，消除数据划分的干扰，保证对于不同分类器与不同粒子的评价标准一致。

对于已经确定好的数据集合，通过调整分类器的参数可以使其达到最佳的工作表现。因此，采用基于遗传交叉因子改进的粒子群优化算法(GPSO)优化模型参数，得到最为稳健的单体分类器[14]。

1.4.4　认证方法的集成学习

集成学习是将一系列训练好的分类器，利用集成规则组合起来，构成一个比单体分类器更加强大的认证整体。经过训练并优化好的分类器就像是一位专家，采用的方法是其擅长的理论，想要一位专家解决所有问题是不现实的。幸运的是，利用集成学习将所有专家的智慧汇聚在一起，能够针对食品产地认证领域的所有问题提供一个接近最优的方法[16]。

2　结果与讨论

2.1　仿刺参数据测定结果

氨基酸含量样品共测定出 16 种特征氨基酸，氨基酸碳稳定同位素样品共测定出 14 种特征氨基酸。通过置信水平为 95%的单总体图基检验，剔除无法有效认证的氨基酸种类，选取出氨基酸含量数据 13 种，氨基酸碳稳定同位素数据 10 种。对不同产地的氨基酸数据，采用箱型图方法分析数据的分布，检测异常值的干扰，最终建立出仿刺参氨基酸数据库。仿刺参氨基酸样品的气相色谱图如图 2 所示，产地为长海县的氨基酸碳稳定同位素数据箱型图如图 3 所示。

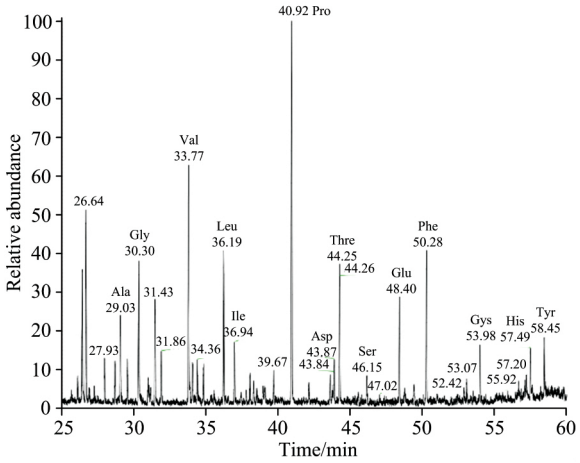

图 2　氨基酸样品气相色谱图

Fig. 2　Chromatogram of amino acids samples

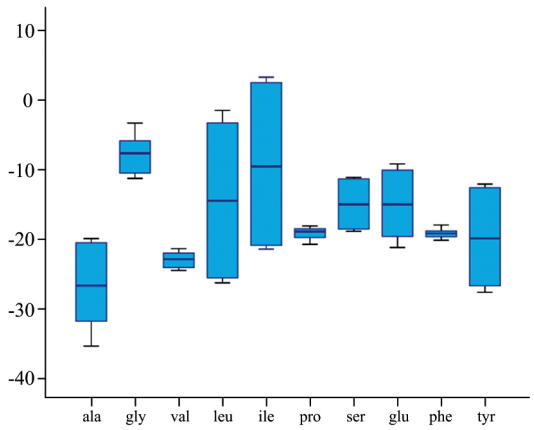

图 3　长海县氨基酸碳稳定同位素数据箱型图

Fig. 3　CH amino acids carbon stable isotope data box-plot

2.2　主成分分析提取结果

经过主成分分析舍弃掉贡献率小于 1 的主成分，保留下氨基酸含量数据的前 5 个主成分；氨基酸碳稳定同位素数据的前 7 个主成分。在保证每一类都有训练样本的条件下，依次对前 N 个主成分进行初始种群规模为 50，遗传进化代数为 40 的模型运算，计算得到最优前 100 项不同 K 值交叉验证的平均准确率，结果如表 1 和表 2 所示。

表 1　氨基酸含量模型的平均准确率

Table 1　Average accuracy of amino acidscontent model

主成分	$K=3$ /%	$K=7$ /%	$K=11$ /%	$K=26$ /%	$K=78$ /%	AVG /%
$N=2$	50.962	54.675	55.327	57.200	61.080	55.849
$N=3$	62.580	67.460	68.949	70.367	75.310	68.933
$N=4$	61.757	66.454	67.911	70.540	74.260	68.184
$N=5$	67.485	75.167	77.301	80.247	83.090	76.658

选取前 5 个主成分作为氨基酸含量模型的输入，累计贡献率为 98.727%；选取前 5 个主成分作为氨基酸碳稳定同位

素模型的输入，累计贡献率为 95.982%。图 4 和图 5 为氨基酸含量与氨基酸碳稳定同位素数据前 3 个主成分的空间分布，氨基酸碳稳定同位素数据具有更加显著的产地聚集特性。

表 2 氨基酸碳稳定同位素模型的平均准确率

Table 2 Average accuracy of amino acids carbon stable isotope model

主成分	$K=3$ /%	$K=7$ /%	$K=11$ /%	$K=26$ /%	$K=78$ /%	AVG /%
$N=2$	93.293	96.653	97.260	98.080	98.740	96.805
$N=3$	95.456	97.305	97.627	98.110	99.120	97.524
$N=4$	98.512	99.651	99.800	99.963	100.000	99.585
$N=5$	99.761	99.998	100.000	100.000	100.000	99.952
$N=6$	99.697	99.998	100.000	100.000	100.000	99.939
$N=7$	99.602	99.987	99.996	100.000	100.000	99.917

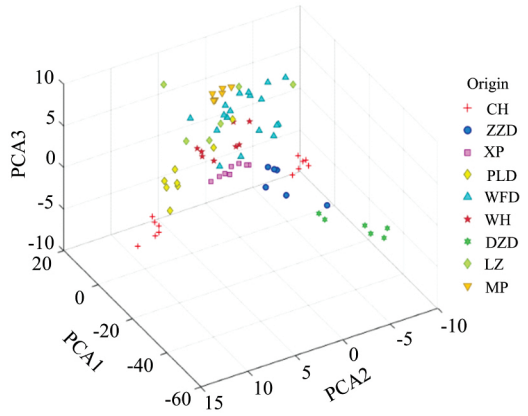

图 4 氨基酸含量数据主成分分析结果

Fig. 4 PCA results of amino acids content data

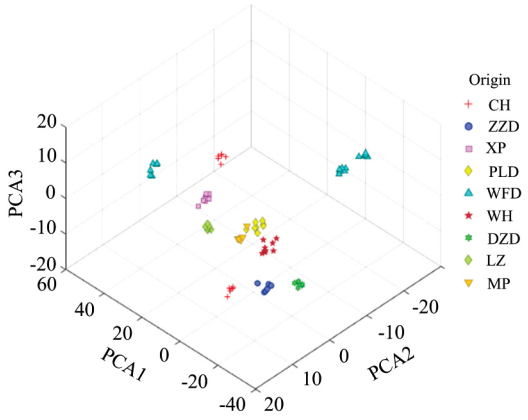

图 5 氨基酸碳稳定同位素数据主成分分析结果

Fig. 5 PCA results of amino acids carbon stable isotope data

2.3 机器学习方法分类结果

利用 Accord.NET 与 Math.NET 框架下的机器学习程序集设计并优化 24 个不同方法的分类器。运用 GPSO 与交叉验证方法，在参数区间内随机设置每个模型参数的初始值，进行种群规模为 100，进化代数为 100，自我学习因子 c_1 为 1.496 18，社会学习因子 c_2 为 1.496 18，权重 w 为 0.752 9的模型参数优化，得到性能最优的单体分类模型。

相对于传统的粒子群优化算法，通过引入遗传算法中的交叉变异算子，在每次遗传进化中以粒子不同 K 值各 100 次交叉验证的平均准确率为适应度，前一半粒子直接进行下一代演化，后一半粒子与前一半粒子进行交叉遗传。这样不断有新的粒子进入到种群中，提高了种群的多样性与全局寻优能力，在保证收敛速度的同时，也防止了模型陷入局部最优解的问题。

图 6 为 24 个单体分类模型的优化结果，每个矩形的上边界为最优项的精度，下边界为第 100 项的精度，矩形中的红线为前 100 项的平均值。最佳的前 9 个模型均使用氨基酸碳稳定同位素数据，体现了 CSIA 更加优秀的产地认证特性；最佳方法为高斯径向基为核的支持向量机与 K 邻近算法，两者的前 100 项精度都达到了 100%。图 7 为氨基酸碳稳定同位素模型的优化过程，证明了 GPSO 结合交叉验证能够快速高效地提高模型性能。

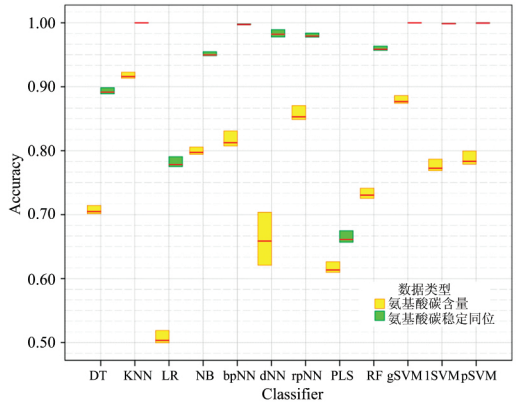

图 6 单体分类模型优化结果

Fig. 6 Monomer classification model optimization results

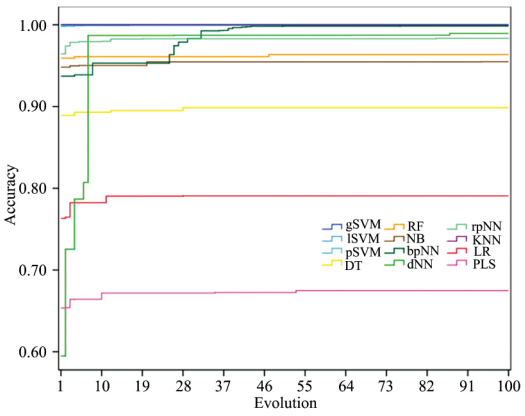

图 7 氨基酸碳稳定同位素模型优化过程

Fig. 7 Optimization process of amino acids carbon stable isotope

2.4 认证模型集成结果

以训练好的 24 个单体模型的最优项精度为权重，选择

出 100 个用于集成的单体分类器，再从对应模型的前 100 项参数中随机选取出每个分类器的参数，最后利用不同的集成学习规则进行 100 个好而不同分类器的集成。选取了 5 种不同的集成规则进行认证，规则的具体描述如表 3 所示。

表 3　集成规则的描述

Table 3　The descriptions of ensemble rules

规则	描述
多数投票	每一类结果为所有分类器预测结果的计数
求和	每一类结果为所有分类器预测概率之和
求积	每一类结果为所有分类器预测概率的乘积
最大	每一类结果为所有分类器预测概率的最大值
最小	每一类结果为所有分类器预测概率的最小值

表 4 为不同集成规则认证模型进行不同 K 值 100 次交叉验证的结果，多数投票规则的认证准确率明显优于其他规则。因此，选取多数投票规则构建产地认证模型，平均准确率为 99.67%，形成了融合多源数据处理方法认证仿刺参产地信息的完整体系。

表 4　不同集成规则交叉验证结果

Table 4　Cross validation results for different ensemble rules

主成分	$K=3$ /%	$K=7$ /%	$K=11$ /%	$K=26$ /%	$K=78$ /%	AVG /%
多数投票	98.577	99.773	100.000	100.000	100.000	99.670
求和	98.288	99.318	99.571	99.833	100.000	99.402
求积	98.346	99.227	99.571	99.833	100.000	99.396
最大	83.865	86.727	88.071	89.833	91.500	87.999
最小	80.038	86.000	85.143	87.500	83.500	84.436

2.5　产地信息认证系统

为了防治食品欺诈事件，主管部门采取了加装防伪标识的手段，但不法商家伪造标识以次充好，更为严重的是部分从业者将其他产地的仿刺参运输到地理标志产地，养殖几天后佩戴上合法标识进行销售。通过构建产地信息认证系统，改变只能预防无法治理的局面，为行业监管与消费者维权提供可靠技术支撑。用户按照指南从终端提交仿刺参样品的氨基酸数据，后台进行分析运算得出认证结果，最后生成检测报告返回给前台，在线生成的检验报告如图 8 所示。

仿刺参产地信息认证系统检验报告

Inspection Report on Apostichopus Japonicus Origin Information Authentication System

委托单位 (Client)	大连海事大学		
地址 (ADD)	大连市凌水路 1 号		
邮箱 (Email)	1113261355@qq.com	电话 (Tel)	18840866641
水产品名称 (Name of Aquatic Products)	海参	加工方式 (Processing Methods)	未加工
样本数量 (Number of Samples)	1	检验方式 (Inspection Methods)	集成认证法
提交日期 (Submission Date)	2019 年 7 月 11 日 9 时 12 分		
数据类型 (Data Types)	氨基酸含量 + 氨基酸碳稳定同位素组成		
检验结论 (Conclusion)			
该海参样本的产地为：獐子岛			

图 8　仿刺参产地信息认证系统检验报告

Fig 8　Inspection report on apostichopus japonicus origin information authentication system

3　结　论

通过氨基酸数据对仿刺参营养富集的详尽刻画，采用主成分分析法降低数据维度，聚集产地认证特性，选取 8 个家族的 12 个分类方法，共建立出 24 个单体分类模型。运用基于遗传交叉因子改进的粒子群优化算法，结合交叉验证与 MCMC 采样，得到性能最佳的单体分类器，最后利用集成学习汇聚单体分类器优势，构建了平均准确率为 99.67%的仿刺参产地信息认证模型。

结果表明，基于氨基酸的多源融合认证方法，能够挖掘出数据背后的价值，保证产地认证准确率的同时，有效提升模型的稳定性与泛化能力。借助互联网技术构建了产地信息认证系统，有效防治了仿刺参产地欺诈事件的发生，促进了整个行业的平稳健康发展。

References

[1] Ma D Y, Yang H S, Sun L N, et al. Acta Oceanologica Sinica, 2014, 33(8): 55.

[2] Zhao Y C, Xue C H, Zhang T T, et al. Journal of Agricultural and Food Chemistry, 2018, 66(28): 7222.

[3] Moyer D C, DeVries J W, Spink J. Food Control, 2017, 71: 358.

[4] WU Peng, LI Ying, LIU Yu, et al(吴　鹏，李　颖，刘　瑀，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(5): 1604.

[5] Van Ruth S M, Huisman W, Luning P A. Trendsin Food Science & Technology, 2017, 67(6): 70.

[6] Cantalapiedra-Hijar G, Ortigues-Marty I, Schiphorst A M, et al. Journal of Agricultural and Food Chemistry, 2016, 64(20): 4058.

[7] Zhao X D, Liu Y, Li Y, et al. Food Control, 2018, 91: 128.

[8] Boschker H T S, Kromkamp J C, Middelburg J J. Limnologyand Oceanography, 2005, 50(1): 70.

[9] Paolini M, Ziller L, Laursen K H, et al. Journal of Agricultural and Food Chemistry, 2015, 63(25): 5841.

[10] Persson J, Nasholm T. Physiologia Plantarum, 2001, 113(3): 352.

[11] Metges C C, Petzke K J, Hennig U. Journal of Mass Spectrometry, 1996, 31(4): 367.

[12] Fernandez-Delgado M, Cernadas E, Barro S, et al. Journal of Machine Learning Research, 2014, 15: 3133.

[13] Wolpert D H. Neural Computation, 1996, 8(7): 1341.

[14] Ghamisi P, Benediktsson J A. IEEE Geoscience and Remote Sensing Letters, 2015, 12(2): 309.
[15] Sun Z B, Song Q B, Zhu X Y, et al. Pattern Recognition, 2015, 48(5): 1623.

Study on the Origin Information Authentication Method of Apostichopus Japonicus Based on Amino Acids

WU Peng[1,2], LI Ying[1,2*], LIU Yu[2,3], CHEN Chen[1,2], RAN Ming-qu[1,2], LI Ya-fang[1,2], ZHAO Xin-da[3]
1. Navigation College, Dalian Maritime University, Dalian 116026, China
2. Environmental Information Institute, Dalian Maritime University, Dalian 116026, China
3. College of Environmental Science and Engineering, Dalian Maritime University, Dalian 116026, China

Abstract The apostichopus japonicus is rich in a variety of active substances, has extremely high medicinal value and economicvalue, and it is an indispensable aquaculture resource for the fishery Industry. There are significant differences in the geographical environment and trophic structure of different producing areas, consequently, the growth cycle and culturing cost of the apostichopus japonicus vary greatly. When consumers buy apostichopus japonicus, they will use the origin information as the primary factor of choice, because the origin of the apostichopus japonicus directly reflects the nutritional value of the food. The price gap between apostichopus japonicus from different producing areas is wide. In the face of the temptation of interest, it is difficult to prevent the occurrence of origin fraud incidents completely. Therefore, a method of apostichopus japonicus origin information authentication with high accuracy, good stability and excellent generalization ability is studied, which effectively protects the vital interests of brand origin practitioners and consumers. Amino acids are the main substances in the nutrient enrichment of apostichopus japonicus. The amino acid characteristics can be used to analyze the composition of primary producers, and as an effective tool for origin information authentication of apostichopus japonicus. Gas Chromatography-Mass Spectrometry (GC-MS) technology produces unique chemical fingerprints for identification of origin information. The 156 samples of the apostichopus japonicus from 9 producing areas were subjected to acid hydrolysis, derivatization and esterification, and finally, the amino acids content and amino acids carbon stable isotope data were determined by GC-MS. Perform a Tukey's test with a 95% confidence level, and the box-plot were used to check the data distribution, and screen 13 amino acids content and 10 amino acids carbon stable isotope data. Principal component analysis can reduce the data dimension, valuable mine information, aggregate the origin information identification characteristics, and improve the calculation speed and authentication accuracy of the model at the same time. Through cross-validation, the first five principal components were selected as the input of amino acids content and amino acids carbon stable isotope model, and the accumulative contribution rates were 98.727% and 95.982%, respectively. In order to fully exploit the value hidden behind the amino acids data, this paper selected 12 machine learning methods from 8 families, built a total of 24 monomer classifiers, and found the optimal authentication method according to the characteristics of the data itself. The particle swarm optimization algorithm based on genetic crossover factor improvement was used to optimize the model parameters, and the best performance monomer classifier was obtained. The results show that the carbon of the amino acid stable isotope data has better origin authentication characteristics. The support vector machine (Gaussianradial basis as the kernel function) and the k-nearest neighbor algorithms are the best two classification methods. Finally, leverage ensemble learning to bring together the advantages of monomer classifiers, a method for origin information authentication of apostichopus japonicus with fusioning multi-source data processing methods is constructed. The average accuracy of the model is 99.67%. An origin information authentication system for the apostichopus japonicus is established, which provides a simple and feasible mean for the supervision of the competent authorities and consumer rights protection. The occurrence of the apostichopus japonicus origin fraud incidents is effectively prevented and controlled, and the stable and healthy development of the aquaculture industry is ensured.

Keywords Apostichopus japonicus; Origin authentication; Amino acids; Principal component analysis; Machine learning; Ensemble learning

* Corresponding author

(Received Jul. 25, 2019; accepted Dec. 2, 2019)