DOI:10.19651/j.cnki.emt.2208766

基于随机森林的气味感知分类研究*

蒋丹凤[1] 温腾腾[1] 吴黎明[1,2] 王 立[1]
(1.广东工业大学机电工程学院 广州 510006;2.佛山沧科智能科技有限公司 佛山 528228)

摘 要: 机器嗅觉是一种基于传感器阵列与计算机算法模拟生物嗅觉的新兴仿生技术,气味物质气味表征是机器嗅觉值得研究的领域,目前嗅觉感知处于初级研究阶段,气味的通用分类理论基础还不成熟。本文从物质气味电子信息角度出发,利用采集样本中相对均衡香型数据,通过机器学习算法及参数调整、网格搜索等模型优化手段,提出基于电子鼻数据的物质气味分类模型,建立物质气味电子鼻信息与感知联系,实验结果表明,基于随机森林的气味分类在各评价指标上表现突出,平均准确率达到 93.6%,随机森林模型相比其他机器学习算法表现优异。

关键词: 气味分类;机器嗅觉;电子鼻;随机森林
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 520.2050

Research on classification of odor perception based on random forest

Jiang Danfeng[1] Wen Tengteng[1] Wu Liming[1,2] Wang Li[1]
(1. School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China;
2. Foshan Cangke Intelligent Technology Co., Ltd., Foshan 528228, China)

Abstract: Machine olfaction is an emerging bionic technology based on sensor arrays and computer algorithms to simulate biological olfaction. The characterization of odor substances is a field worthy of research in machine olfaction. At present, olfactory perception is in the preliminary research stage, and the general classification theory of odor is not yet mature. In this paper, starting from the electronic information of material odor, aiming at the relatively balanced fragrance data in the collected data, using machine learning algorithms and parameter adjustments, grid search and other model optimization methods, the material odor classification model based on electric nose data is proposed, and the connection between the information and perception of the material odor electronic nose is established. The experimental results show that the random forest model performs better than other machine learning algorithms in each evaluation index, and the average accuracy of odor classification based on random forest reaches 93.6%.

Keywords: smell classification; machine smell; electronic nose; random forest

0 引 言

气味的存在影响着人们的生活,比如情绪、记忆甚至行为[1-2]。在所有感官中,嗅觉是最难以理解与描述的,人们对气味的认知主要借助言语描述词,对于气味的感知存在极大的主观意志,受生活经验、语言[3]、敏感度、感知疲劳、适应能力等因素影响,个体之间对于嗅觉的判断亦存在着认知差异。气味特性的测量,如质量、强度、相似性等,还未形成通用统一的定义标准。若利用机器解决气体或气味的检测问题可以避免人工感知的主观因素,降低人工感知气味的危险性,同时基于机器嗅觉的气体及气味的检测识别随着传感器与人工智能技术的发展日益强大。

机器嗅觉目前主要集中在具体物质分类识别方面的研究,在食品、医疗、化工、环境以及机器人等领域都有一系列较为成熟的研究发展。如通过采集电子鼻数据利用采用局部切空间排列与线性判别分析方法识别柑橘品种[4],利用电子鼻和气相色谱质谱联用仪研究烘烤水平对大豆分为特征的影响[5],探讨电子鼻在呼吸医学中的应用[6],利用电子鼻监测家禽养殖场对不同恶臭进行分类和实时监测[7],设计机器人主动嗅觉平台实现气味的快速探测和定位[8]。由于气味认知的复杂性,通过电子或化学信息数据预测气味是一项富有挑战性的任务。

近几年,一些研究者开始研究化学结构与气味的关系,采用机器学习模型从气相色谱质谱仪获得的气味分子参数预测气味信息[9],结合分子描述符信息和化学结构图像预

收稿日期:2022-01-05
* 基金项目:广东省科技计划项目(2019B101001017)、佛山广工大研究院创新创业人才团队计划项目(20191108)资助

测化学物质的气味[10]，通过图神经网络捕获结构和气味之间的潜在关系[11]。除了化学结构信息，气味的预测还可以根据电子鼻的传感器数据开展研究，通过电子鼻进行气味愉悦度的研究，区分愉悦程度相关性超 90%，在绝对令人愉快和绝对不愉快的气味分类准确率达到 99%[12]。电子鼻数据的应用更多在于具有挥发性气体的物质分类[13-15]，然而气味的种类千变万化，仅仅对特定气味分类的手段不能达到感知的目的。因此，本文提出采用香型标签来描述气味，为气味贴上一个或多个标签，实现机器嗅觉感知。机器嗅觉的气味感知预测的研究对于气味检测与评价具有重要意义，能够减少时间和人力成本，同时提升嗅觉感知的客观性，将逐渐成为机器嗅觉领域研究的热点，对于人工智能和机器嗅觉的未来发展具有研究意义和应用价值。

1　基于随机森林算法的气味预测方法及原理

1.1　建立气味数据集

根据采集的香料电子鼻数据及林翔云标注的气味类别建立一个气味数据集，原始样本数据结构为 $\{\boldsymbol{X},y\}$，其中 $\boldsymbol{X}$ 为 120×10 的样本数据矩阵，$\boldsymbol{X}=\{x_1,x_2,\cdots,x_{10}\}$，$x_i$ 是第 i 个传感器的响应值序列，$y=\{y_1,y_2,\cdots,y_{32}\}$ 是该样本的气味标签，实际气味标签的预测是一个多标签问题。提取各传感器后 40 s 的均值即稳态值作为传感器特征，为第 i 个传感器的稳态值，建立以稳态值为特征的气味多标签数据集 $\{\overline{\boldsymbol{X}},y\}=\{\overline{x_1},\overline{x_2},\cdots,x_{10},y\}$，即最终输入模型的样本数据结构为 10 个传感器响应平均稳态值矩阵和数据标签。其中：

$$\overline{x_i}=\frac{\sum_{t=81}^{120}x_{it}}{40} \tag{1}$$

1.2　构建随机森林预测模型

随机森林(random forest)是一种以决策树为基分类器的集成学习算法，可以集成多种决策规则对数据集进行学习，尤其在多维数据应用广泛并有良好表现[16-20]。随机森林是由很多决策树分类模型组成的组合分类模型，每个决策树分类模型都具有投票权从而获得最优的分类结果，以提高模型的准确性。

算法的基础实现流程如下：1)通过装袋重采样方法从原始数据集中有放回地抽取 n 个样本，作为训练数据集；2)选择特征对抽取到的 N 个样本集建立决策树；3)设样本集 X 包含 M 个特征属性，从 M 个属性中随机选取 m 个属性作为子集($m<M$)，根据最小 Ginni 系数原则从这个子集中选取最优属性作为分裂变量；4)重复以上步骤，建立多个决策树组成随机森林，分别利用 n 个决策树进行分类预测，遵循投票机制输出气味预测结果。随机森林在训练阶段，使用自助法重采样技术从输入训练数据集中采集多个不同的子训练数据集来依次训练多个不同决策树，将这些决策树拟合到数据集的各个子样本上。将测试样本数据输入随机森林，并取每个决策树预测结果的平均值作为最终的预测结果，以提高预测的准确性。

1.3　模型性能提升

为模型选择合适的参数可以提升模型性能，本文随机森林模型通过交叉验证和网格搜索改善模型效果，最终选择最小的平均均方根误差(RMSE)和最高准确率所对应的最佳优化参数，提高模型的稳定性及泛化性。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{2}$$

其中，y_i 为样本真实值，为 $\hat{y}_i$ 测试样本的预测值。

网格搜索实际上是一种列举搜索的参数优化方法。首先设置空间各参数，通过遍历所有网格参数组合节点确定模型的均方根误差或准确率进行模型优化。只有遍历网格平面的所有节点得到使准确率最高或均方根误差最小的参数节点，也就是模型的最优参数组合。

本文主要对以下关键参数进行搜索：决策树的数量 n_estimators、每棵树的最大特征变量 max_features、树的最大深度 max_depth 和最小子节点样本数 min_samples_leaf。最终选择表现最好的超参数来训练预测模型，同时在测试集上对模型的有效性进行评价，网格搜索参数设置范围如表 1 所示。

表 1　网格搜索参数设置表

超参数	搜索范围
max_depth	range(5,20)
n_estimators	range(10,200,10)
min_samples_leaf	range(1,10)
max_features	range(1,10,2)

2　实验与结果分析

2.1　数据来源

实验数据采集仪器为德国 AIRSENSE 公司生产的 PEN3 电子鼻，实物如图 1 所示，是一种通用的气体采样仪，主要由传感器阵列、气路流量控制系统和信号处理系统组成。基于金属氧化物气体传感器阵列对气体进行感知、分析判断的检测系统，能够得到快速、客观、可靠、稳定的测量结果。PEN3 电子鼻内置 10 个不同的金属氧化物传感器，检测类型及敏感气体信息如表 2 所示。

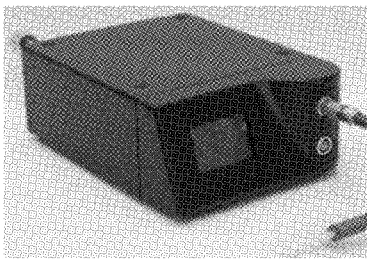

图 1　PEN3 电子鼻实物图

表 2　PEN3 电子鼻传感器信息

序号	传感器名称	敏感气体
1	W1C	芳香成分
2	W5S	灵敏度大，对氮氧化合物很灵敏
3	W3C	氨水，对芳香成分灵敏
4	W6S	主要对氢气有选择性
5	W5C	烷烃，芳香成分
6	W1S	对甲烷灵敏
7	W1W	对硫化物灵敏
8	W2S	对乙醇灵敏
9	W2W	芳香成分，对有机硫化物灵敏
10	W3S	对烷烃灵敏

本文以电子鼻采集数据为基础展开嗅觉感知方面研究，实验数据采集使用德国 PEN3 电子鼻设备，其内置 10 个具有交义灵敏性的传感器。实验采集过程如图 2 所示，连接电子鼻设备各进气、洗气接口及显示设备，使用移液器取液滴入 150 ml 烧杯，用保鲜膜密封静置 20 min，使其气味分子得到充分挥发，尽量将环境温湿度保持在数据收集过程的稳态一致性。测量开始前设置电子鼻测量参数，采用顶空采样法，对每个样本进行 1 次/s 的气味信息采集，采集时间为 120 s，完成一次数据采集过程，故一次样本采集获得数据格式为：120×10 阵列数据。每个样本取样量分别为 5、10、20 μl，电子鼻测量流速设置为 50、100、200 ml/min，由于传感器响应与取样量、流速的非线性特性，丰富取样与测试过程使得样品测量数据具有多样性，拓展模型的适应能力。在 WinMuste 软件中进行参数设置，具体参数如表 3 所示。

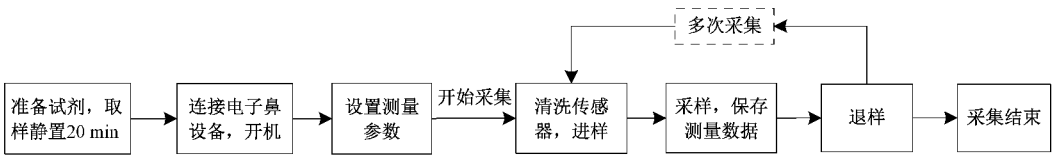

图 2　数据采集实验流程图

表 3　实验参数设置

参数名称	数值
静置时间/min	20
采样时间/s	120
采样间隔/s	1
洗气时间/s	120
样本质量/μl	5、10、20
进气流量/(ml/min)	50、100、200

每组样品测量次数为 10 次，故每种样本全部采集结束后均可获得 90 个 nos 文件。本文实验对象为甲酸甲酯、丁酸乙酯、苯甲醛、异丙醇等 34 种香料，获得 34×90＝3 060 个数据样本文件。经过数据预处理进行数据可视化展示，如图 3 所示为一次丁酸甲酯采样数据的可视化显示。

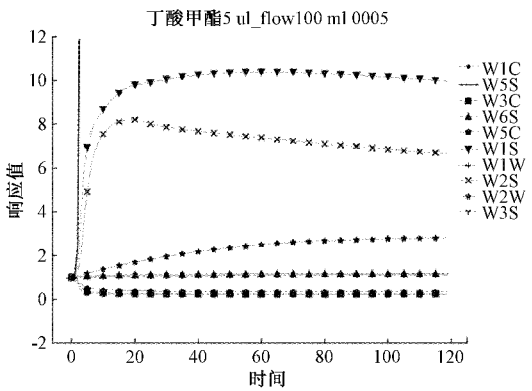

图 3　香料数据可视化展示

数据标签以林翔云定义的自然界气味关系中的 32 种香型为原始气味标签，是在研究香精香料的配制与应用后，从电子信息的角度建立气味分布体系。

2.2　数据标签分布

气味的描述与分类模式复杂多样，可从心理学、神经科学、植物学等进行分类，仍难以全面描述自然界的所有气味。林翔云参考了多种香气分类体系和理论，结合几十年调香评香经验以及综合调香师经验，经过多次调整修改形成自然界气味关系图[21]，可对上千种香料赋予气味描述，是一种较为通用的气味描述体系。本文即采用自然界关系图以气味 ABC 方法定义的自然界 32 种气味作为气味标签，所谓气味 ABC 方法表征的是物质气味的混合特性，例如，乙酸甲酯具有水果味、香橼味、坚果味和醛香味。

本文采集了 34 种香料化合物数据，拥有 21 种气味标签，构成 3 060 个 csv 文件样本数据集。由于香料数据选择的局限性和气味多标签分布不均衡的特性，对数据标签分布进行统计，图 4 给出 34 种香料气味标签的分布情况，本文对分布频率较高的前 8 种气味标签（水果、醛香、乳酪、青气、芳烃、油脂、坚果、玫瑰）进行实验研究，建立电子鼻数据与气味标签的关系，同时保证数据的相对均衡性。

2.3　实验结果及对比

实验基于 Python 语言使用 scikit-learn 库中 Ensemble 框架建立随机森林模型，由 sklear. model_selection 导入 GridSearchCV 模块执行网格搜索对模型进行优化，另外与其他机器学习模型进行对比，实验结果采用准确率、精确

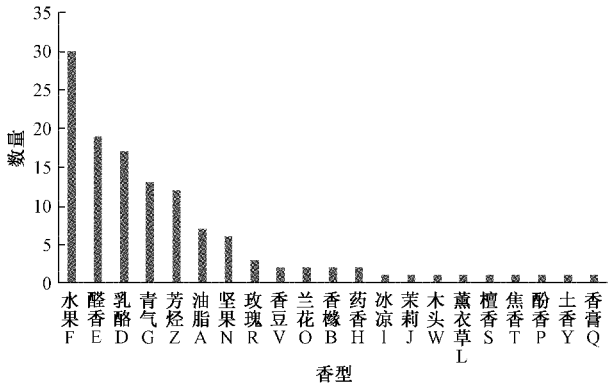

图 4　采集数据标签分布情况

度、召回率、F1 作为评价指标，最终由各气味评价指标的平均值判定各模型效果。

由问题转换策略将原始数据多标签问题转换为多个二分类任务，构建随机森林物质气味预测模型，通过网格搜索和交叉验证手段调整模型进行性能优化，与其他常用的机器学习方法 KNN、DT、GBDT 进行比较研究，进一步验证随机森林算法预测模型的有效性。本文采用常用的准确率(acc)、精确度(p)、召回率(r)、F1 作为评价指标，由表 4 可知，与其他机器学习算法模型对比随机森林模型表现最为优异，对于所研究的 8 种气味标签的平均准确率达到 93.6%，在准确率、精确度等评价指标上均有较为突出的表现。

表 4　预测实验结果对比

odor	KNN				GBDT				RF			
labels	acc	p	r	F1	acc	p	r	F1	acc	p	r	F1
乳酪 D	0.75	0.71	0.77	0.74	0.90	0.88	0.9	0.89	0.93	0.93	0.92	0.93
青气 G	0.78	0.68	0.68	0.68	0.84	0.76	0.78	0.77	0.83	0.76	0.74	0.75
油脂 A	0.86	0.71	0.62	0.66	0.86	0.71	0.62	0.66	0.97	0.96	0.94	0.95
醛香 E	0.72	0.7	0.82	0.75	0.90	0.88	0.94	0.91	0.91	0.90	0.92	0.91
芳烃 Z	0.81	0.75	0.68	0.71	0.92	0.94	0.83	0.88	0.92	0.93	0.82	0.87
水果 F	0.98	0.96	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.994	0.995	0.994
坚果 N	0.92	0.87	0.67	0.76	0.94	0.93	0.76	0.837	0.95	0.94	0.8	0.866
玫瑰 R	0.99	0.89	0.93	0.91	0.99	0.97	0.91	0.94	0.99	0.937	0.978	0.957
平均值	0.851	0.783	0.77	0.774	0.916	0.883	0.841	0.859	**0.936**	**0.918**	**0.889**	**0.903**

3　结　　论

近年气味的受关注程度越来越高，不管是特定物质气味、特殊场景气味还是空气质量等一系列问题中都突出其重要地位。现有的气味检测多集中在特定的气味预测，尚未形成通用的气味表征系统。本文将电子鼻数据与专家提供的香料气味描述建立气味预测模型，为气味的通用表征提供一种思路，但理论和应用上还存在一些不足之处，后续能够在以下方面深入研究：

1)由于受到实验条件限制，数据样本量不足导致气味标签不均衡，本文仅对 8 种气味进行研究，后续还可以进行数据扩充以实现多样气味标签的预测；

2)随机森林虽在高维非线性数据上有优异的表现，但网格搜索方法耗时较长，对气味检测效率有一定影响；

3)目前机器嗅觉的研究更多在于对具有挥发性物质的分类，气味表征研究还处于初级发展阶段，在探索理论研究的基础上还应该立足于实际应用中，进一步丰富气味检测的应用场景。

参考文献

[1]　王娟，沈树华，张积家. 大学生的气味词分类——基于语义相似性和知觉相似性的探讨[J]. 心理学报，2011，43(10)：1124-1137.

[2]　聂春艳，宋晓兵，孟佳佳. 环境气味对消费者产品评价和购买意向的影响研究[J]. 管理科学，2016，29(5)：93-105.

[3]　SOROKOWSKA A，ALBRECHT E，HUMMEL T. Reading first or smelling first? Effects of presentation order on odor identification[J]. Attention Perception & Psychophysics，2015，77(3)：1-6.

[4]　彭珂，骆德汉，夏必亮. 基于机器嗅觉的柑橘品种无损检测与识别[J]. 江西农业大学学报，2017，39(5)：1017-1024.

[5]　CAI J S，ZHU Y Y，MA R H，et al. Effects of roasting level on physicochemical，sensory，and volatile profiles of soybeans using electronic nose and HS-SPME-GC-MS[J]. Food Chemistry，2021，DOI：10.1016/j.foodchem.2020.127880.

[6]　GASPARRI R，SEDDA G，SPAGGIARI L. The electronic nose's emerging role in respiratory medicine[J]. Sensors，2018，18(9)：3029，DOI：10.3390/s18093029.

[7]　AUNSA-ARD W，POBKRUT T，Kerdcharoen T，et al. Electronic nose for monitoring of livestock farm odors (poultry farms)[C]. 2021 13th International

Conference on Knowledge and Smart Technology (KST), 2021:176-180.

[8] 康张琦. 三维机器人主动嗅觉仿真平台设计[D]. 天津:天津大学,2018.

[9] LIANG S, LIU C, TOMIURA Y, et al. Machine-learning-based olfactometer: prediction of odor perception from physicochemical features of odorant molecules[J]. Analytical Chemistry, 2017, 89(22): 11999-12005.

[10] SHARMA A, KUMAR R, RANJTA S, et al. SMILES to Smell: Decoding the structure-odor relationship of chemical compounds using the deep neural network approach[J]. Journal of Chemical Information and Modeling, 2021, 61(2):676-688.

[11] SANCHEZ-LENGELING B, WEI J N, LEE B K, et al. Machine learning for scent: Learning generalizable perceptual representations of small molecules[J]. ArXiv Preprint, 2019, DOI:10.48550/arXiv.1910.10685.

[12] 吴丹莉. 基于电子鼻的气味愉悦度评估模型研究[D]. 广州:广东工业大学,2020.

[13] 李超, 周博. 自制电子鼻检测霉变大米[J]. 食品工业科技, 2021, 42(12):218-224.

[14] 岳盈肖, 闫子茹, 赵江丽, 等. 利用电子鼻解析采后深州蜜桃品质变化[J/OL]. 保鲜与加工, 2021, 21(8):101-108.

[15] 李志远, 舒涵, 靳梦亚, 等. 基于电子鼻和人工神经网络的沉香与沉香曲鉴别[J]. 中国现代中药, 2021, 23(2):286-289.

[16] 吕红燕, 冯倩. 随机森林算法研究综述[J]. 河北省科学院学报, 2019, 36(3):37-41.

[17] 杨傲雷, 刘佳奇, 徐昱琳, 等. 融合随机森林模型的单目视觉人体空间定位方法[J]. 仪器仪表学报, 2020, 41(11):207-215.

[18] 王毅, 陈进, 李松浓, 等. 基于时频域分析和随机森林的故障电弧检测[J]. 电子测量与仪器学报, 2021, 35(5):62-68.

[19] 尤志军, 俞秋峰, 汪晓晖, 等. 基于随机森林法的精神分裂症患者病情复发的预测[J]. 国际精神病学杂志, 2021, 48(4):631-636.

[20] 徐肖伟, 李鹤健, 于虹, 等. 基于随机森林的变压器油中溶解气体浓度预测[J]. 电子测量技术, 2020, 43(3):66-70.

[21] 林翔云, 王丽萍, 林君如, 等. 自然界气味关系图[J]. 香料香精化妆品, 2015(1):66-73.

作者简介

蒋丹凤,硕士研究生,主要研究方向为机器嗅觉、深度学习。

E-mail:2111901127@mail2.gdut.edu.cn

温腾腾,博士后,主要研究方向为机器嗅觉、智能感知。

E-mail:wentt@gdut.edu.cn

吴黎明(通信作者),教授,硕士生导师,主要研究方向为仪器科学与技术、智能测控技术。

E-mail:jkyjs@gdut.edu.cn