

基于多元线性回归的食用油掺假检测模式识别算法设计与优化

丁宇航

浙江理工大学信息科学与工程学院, 浙江 杭州

收稿日期: 2024年6月27日; 录用日期: 2024年7月24日; 发布日期: 2024年7月31日

摘要

目前食用油市场发展迅速但掺假情况严重, 既损害了消费者的利益, 同时又干扰了市场秩序。且国内缺乏一套具有自主知识产权的智能便携式食用油掺假检测系统。本文设计了一款基于多元线性回归算法的便携式食用油掺假检测系统, 该系统在多元线性回归算法(MLR)的基础上进行拓展, 对电子鼻采集到的食用油的气味信号进行处理, 实现对食用油的掺假成分的定性分析和定量计算。实验结果表明所设计的基于MLR的算法对食用油的定性分析准确率可达97%, 对食用油掺假定量识别误差率小于4%。

关键词

食用油掺假, 多元线性回归, 电子鼻, 掺假检测

Design and Optimization of Edible Oil Adulteration Detection Mode Recognition Algorithm Based on Multiple Linear Regression

Yuhang Ding

School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Jun. 27th, 2024; accepted: Jul. 24th, 2024; published: Jul. 31st, 2024

Abstract

The edible oil market is currently developing rapidly but is plagued by severe adulteration issues,

文章引用: 丁宇航. 基于多元线性回归的食用油掺假检测模式识别算法设计与优化[J]. 建模与仿真, 2024, 13(4): 4953-4961. DOI: 10.12677/mos.2024.134448

which harm consumer interests and disrupt market order. Moreover, there is a lack of an intelligent portable edible oil adulteration detection system with independent intellectual property rights in China. This paper designs a portable edible oil adulteration detection system based on the multiple linear regression (MLR) algorithm. The system expands on the MLR algorithm to process the odor signals of edible oils collected by an electronic nose, achieving qualitative analysis and quantitative calculation of the adulteration components in edible oils. Experimental results show that the designed system based on the MLR algorithm has an accuracy rate of up to 97% for qualitative analysis of edible oils and a quantitative recognition error rate of less than 4% for edible oil adulteration.

Keywords

Edible Oil Adulteration, Multiple Linear Regression, Electronic Nose, Adulteration Detection

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,中国食用油市场发展迅速,品种繁多,尤其是一些消费者喜爱的高端油料油脂进口量逐年增加,诸如芝麻、亚麻籽和椰子油等[1]。然而,为了降低食用油的生产成本,一些商家掺假食用油,导致食用油的营养价值贬值甚至对人体有害,比如给高价茶籽油添加廉价的大豆油、有害的地沟油等[2]。因此,掺假食用油不仅侵犯了消费者的权益,而且对消费者的健康造成了危害。所以,掺假食用油的检测鉴定在食品安全领域具有重要意义。

目前,食用油的掺假检测鉴别技术已有传统的理化检测方法到高精度的仪器分析技术。主流的是色谱法和光谱法,色谱法通过色谱测量,对食用油的某些化学成分分析,优点是准确率高,但是需要对样本进行提前处理,造成一定的损失;另一种光谱法,则不需要对样本进行处理,直接通过光谱结合化学测量方法生成模型,但模型需要具有统计学意义的样本数据,极大降低了模型对外部的适用性[3]。而随着新型传感技术电子鼻[4]、电子舌的发展,在食品检测方面占据了一席之地,其具有简便、易操作,结果客观性、准确性的优点,在食品生产、食品监管和日常生活中对食品质量安全发挥着重要作用[5]。

尽管国内电子鼻的研究相较于国际发展较晚,但近年来已呈现出蓬勃的发展态势。曹森[6]等以芦笋为研究对象,采用电子鼻结合气相色谱技术对不同采收期芦笋样品的芳香品质进行了分析鉴定。电子鼻数据的主成分分析和线性判别分析结果验证了区分芦笋不同采收期的能力。张欣[7]等以荆州鱼糕为研究对象,采用了逐步多元线性回归分析,建立了关于鱼糕新鲜度的判别与预测模型,能够针对鱼糕存储过程中产生的各种气体,快速无损地判别鱼糕的新鲜度。冯蕾[8]等使用电子鼻测量了黄瓜和樱桃的新鲜度,并开发了一套检测系统。

在国外电子鼻已广泛应用于油脂质量分析。Wei [9]等人(2018年)使用电子鼻传感器测定了牡丹籽油的掺假情况。研究结果表明,牡丹籽油中含有大量的 α -亚麻酸,可用于鉴别牡丹籽油的纯度和掺假情况。Ayari、Mirzaee-Ghaleh、Rabbani 和 Heidarbeigi [10] (2018年)之前的研究揭示了电子鼻检测不同含量的葵花籽油和牛酥油中的掺假情况。结果表明,主成分分析对葵花籽油和牛酥油掺假的识别率分别达到了96%和97%。Kishimoto 和 Kashiwagi [11] (2019年)研究了电子鼻在识别过滤和未过滤橄榄油挥发性变化中的

应用。研究表明, 过滤橄榄油中的主要挥发性化合物, 即醛、酮、酯和醇明显减少。而目前国内缺少针对我国食用油模式识别的电子鼻系统。

本文提出了一种基于多元线性回归算法的食用油检测模式识别算法设计与优化, 该算法搭建在所设计的便携式电子鼻上, 能通过采样装置将不同食用油散发的气体或气味采集进去, 并利用传感器阵列对多种气体或气味敏感将气味信号转为电信号, 使用实时高速处理器采集传感器阵列输出信号, 有效地完成对食用油掺假成分的定性分析和定量检测的目的。

2. 基于电子鼻模式识别的算法设计

2.1. 多元线性回归算法

设随机变量 Y 于一般变量 $X_j (j=1, 2, \dots, k)$ 的线性回归模型为

$$Y = \beta_0 + \sum_{j=1}^k \beta_j X_j + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \quad (1)$$

其中 Y 称为被解释变量; $X_j (j=1, 2, \dots, k)$ 是 k 个可以精确控制的一般变量, 称为解释变量。是去除 k 个自变量对 Y 影响后的随机误差项(残差)。 $\beta_0, \beta_1, \dots, \beta_k$ 是 $k+1$ 个未知参数, β_0 为回归常数, $\beta_1, \beta_2, \dots, \beta_k$ 称为回归系数[12]。

多元总体线性回归方程, 也就是 Y 的期望与 $X_j (j=1, 2, \dots, k)$ 之间的线性方程为

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (2)$$

其中 X_1 是对 Y 的偏回归系数。

同样, 多元线性样本回归方程为

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j X_j + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \quad (3)$$

其中 $\hat{\beta}_j (j=0, 1, 2, \dots, k)$ 为参数估计值, $\hat{Y}$ 则是实际因变量 Y 的样本拟合值或者是样本预测值。

MLR 算法方程中回归系数 $\beta_1, \beta_2, \dots, \beta_k$ 的估计可以使用普通的最小二乘法求得。残差平方和公式为:

$$SSE = \sum (y - \hat{y})^2 = 0 \quad (4)$$

将 SSE 对 $\beta_1, \beta_2, \dots, \beta_k$ 求偏导数, 然后将它赋值为零, 对它进行简化, 就能得出 $k+1$ 个方程式:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum (y - \hat{y}) = 0 \quad (5)$$

$$\frac{\partial SSE}{\partial \beta_i} = -2 \sum (y - \hat{y}) x_i = 0 \quad (6)$$

通过求解该方程组便可分别得到 $\beta_0, \beta_1, \dots, \beta_k$ 的估计值和 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 回归系数的估计值。而通过化简等方法可以将这一方程组用矩阵方式表示出来:

$$XY = X\hat{\beta} \quad (7)$$

所以通过普通的最小二乘法求得回归系数的求解, 可以用矩阵表示为

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (8)$$

$$\text{其中 } X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}。$$

2.2. 拓展的多元线性回归算法

一般多元线性回归模型进行食用油掺假的定量判别对食用油的判断并不准确，因为其自变量和因变量之间的关系是非线性的，本研究提出了另一种数学模型，对自变量进行扩展。因为非线性的数学模型可由三部分组成：一是个自变量单个变化的影响，二是各自变量自乘方次的影响，三是各自变量互相组合的影响。因此可以得知其数学模型为：

$$Y_i | x_{i,1}, x_{i,2}, \dots, x_{i,m} = \beta_0 + \sum_{k=1}^m \sum_{j=1}^m \beta_{i,j-1} x_{i,k}^j + \sum_{r=2}^m \sum_{\{k_1, k_2, \dots, k_r\} \subset N_m} \beta_{k_1, k_2, \dots, k_r}^* x_{i,k_1} x_{i,k_2} \dots x_{i,k_r} \quad (9)$$

其中 m = 自变量个数，也是自变量的最高幂次； N = 实验所做次数， $i = 1, 2, \dots, N$ ； $N_m = \{1, 2, \dots, m\}$ 。

根据(8)的公式将自变量 x 进行变化之后，接下来对于该公式系数的求解，可以参考普通的多元线性回归算法，这里不再重复书写。

因为优化变化后的模型是用来进行定量判断的，所以样本数据自身携带 y 值。

采用如下步骤实现优化算法：

- (1) 采集食用油的 p 维气体特征数据 $X = (x_1, x_2, \dots, x_p)'$ 的 n 个样本 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ 列出矩阵 $X = (x_{ij})_{n \times p}$ 和 $Y = (y_i)_{n \times 1}$ ；
- (2) 对自变量数据进行标准化，即进行无量纲化处理，使得数值均处在同一数量级别上；
- (3) 按照自变量的扩展方式对自变量数据进行相应变化，将变化后的数据重新组合成新的自变量数据矩阵；
- (4) 根据新的自变量数据矩阵计算出回归系数，从而建立回归模型；
- (5) 将测试数据代入样本数据回归模型中，求得预测值 y ，即可得所需结果。

3. 食用油检测的模式识别的实验设计

3.1. 采集食用油气味信号的电子鼻系统

图 1 是所设计的食用油掺假检测电子鼻系统组成框图。

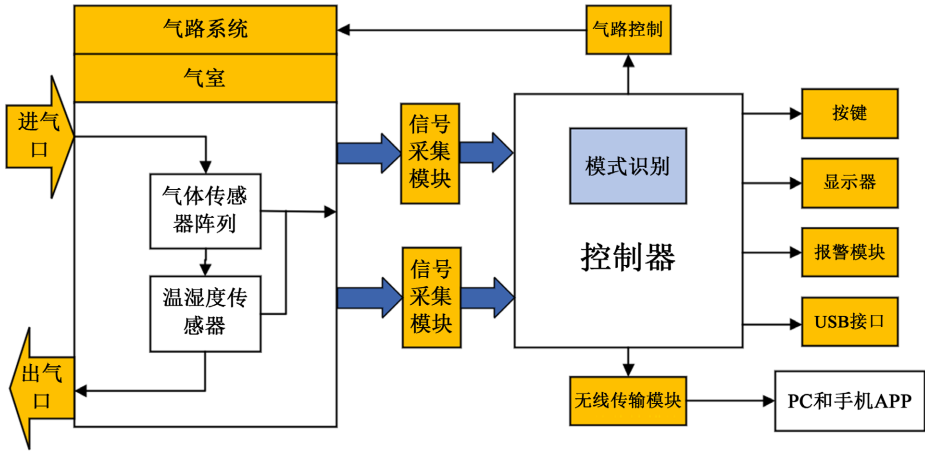


Figure 1. Electronic nose system block diagram
图 1. 电子鼻系统结构框图

所设计的食用油掺假检测专用电子鼻系统由气体采样装置、传感器阵列、温湿控制电路、信号调理电路、控制器、显示模块、报警模块和通讯接口电路组成。系统能通过采样装置将不同食用油散发的气

体或气味采集进去。并利用传感器阵列对多种气体或气味敏感,将气味信号转为电信号,使用实时高速处理器采集传感器阵列输出信号,并经过模式系统能通过采样装置将不同食用油散发的气体或气味采集进去。并利用传感器阵列对多种气体或气味敏感,将气味信号转为电信号,使用实时高速处理器采集传感器阵列输出信号,并进行相应的预处理,能够滤除测量过程中的干扰和噪声,放大传感器微弱信号,同时实现由模拟量到数字量的精确转换,并经过模式识别技术,有效地完成对食用油掺假成分的定性分析和定量检测的目的。控制器采用浮点的 DSP 系统,既要实现电路的控制,还要实现对采集到的食用油气味信号的模式识别处理。系统既能脱机使用,可以通过 USB 接口和 PC 机相连,将数据传给 PC 机进行数据的存储和进一步处理。

3.2. 数据预处理

在进行多元线性回归算法的分析之前,我们都需要将所有的数据进行预处理。在本文中我们采取了肖维涅准则和 Z-score 标准化。

肖维涅准则是一种十分常见的用来剔除可疑数据的方法准则,经过该准则后,在一定程度上可确保数据的稳定性。

首先计算数据($n > 4$)的算数平均值 $\bar{X}$,再计算标准偏差 S ,求得肖维涅系数 $C(n)$,计算 $X_{\min}$ 和 $X_{\max}$,检测所有数据,剔除不在 $(X_{\min}, X_{\max})$ 范围内的坏值。

标准偏差的公式为

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (10)$$

其中 $\bar{X}$ 为所采集数据的 $X_1, X_2, \dots, X_n$ 的均值。

对于肖维涅系数的求解,本文采用了赵九明所提出的一种肖维涅系数的拟合公式,其原理也是根据最小二乘原理而获得的。在给定范围内,拟合度达到了 0.01 [13]。

当 n 大于等于 4 并且同时满足小于等于 70 时,肖维涅系数的拟合公式为

$$C(n) = 0.359 \ln(n - 1.5) + 1.193 \quad (11)$$

当 n 大于 70 并且同时满足小于等于 2000 时,肖维涅系数的拟合公式为

$$C(n) = 0.2688 \ln(n - 20) + 1.63 \quad (12)$$

数值的最大范围为

$$X_{\min} = \bar{X} - (C(n)S) \quad (13)$$

$$X_{\max} = \bar{X} + (C(n)S) \quad (14)$$

具体实现时,在进行食用植物油的算法判别前,将训练集数据分样本进入肖维涅准则程序,以免影响该准则的准确性。在程序内判断读取的数据是否符合肖维涅准则,若超出了该准则所容许的范围则记录该组数据的序列号,而后根据序列号集将对应的训练集数据进行坏值的剔除。

若想将数据进行标准化,其方法有很多,而比较常用的是“Z-score 标准化”、“最小-最大标准化”和“按小数定标标准化”。在本文中所采取的是“Z-score 标准化”。

标准化变换:设有 n 个待测样本单位,每个单位有 p 个指标,则有矩阵 $X = (X_{ij})_{n \times p}$,其中 X_{ij} 表示第 i 单位的第 j 指标值。

从列中的每个值中减去该列的均值,然后将列中的每个值除以该列的标准差。

$$\bar{X}_j = \sum_{i=1}^n \frac{X_i^2}{n} \quad (15)$$

$$S_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (16)$$

$$Z_{ij} = \frac{(X_{ij} - \bar{X}_{ij})}{S_j} \quad (17)$$

通过该 Z-score 标准化处理之后，我们确保了每列数据的平均值等于 0，方差的数值都为 1。我们之所以要进行标准化处理，是为了在测试集样本数据发生变化时，仍然可以得到我们想要的结果，就是在一定程度上可以保证数据的相对稳定性。记标准化之后的数据矩阵为 $Z = (Z_{ij})_{n \times p}$ 。

我们在进行具体实现时，需要将训练集数据经过肖维涅准则之后，再进入标准化模块。在标准化模块中要注意自己想要标准化的是行数据还是列数据。而在进行标准化的过程中，若出现标准差为 0 的情况，则将该标准差值设为 1。

3.3. 多元线性回归算法的实现流程

MLR 算法的实现流程如图 2 所示。

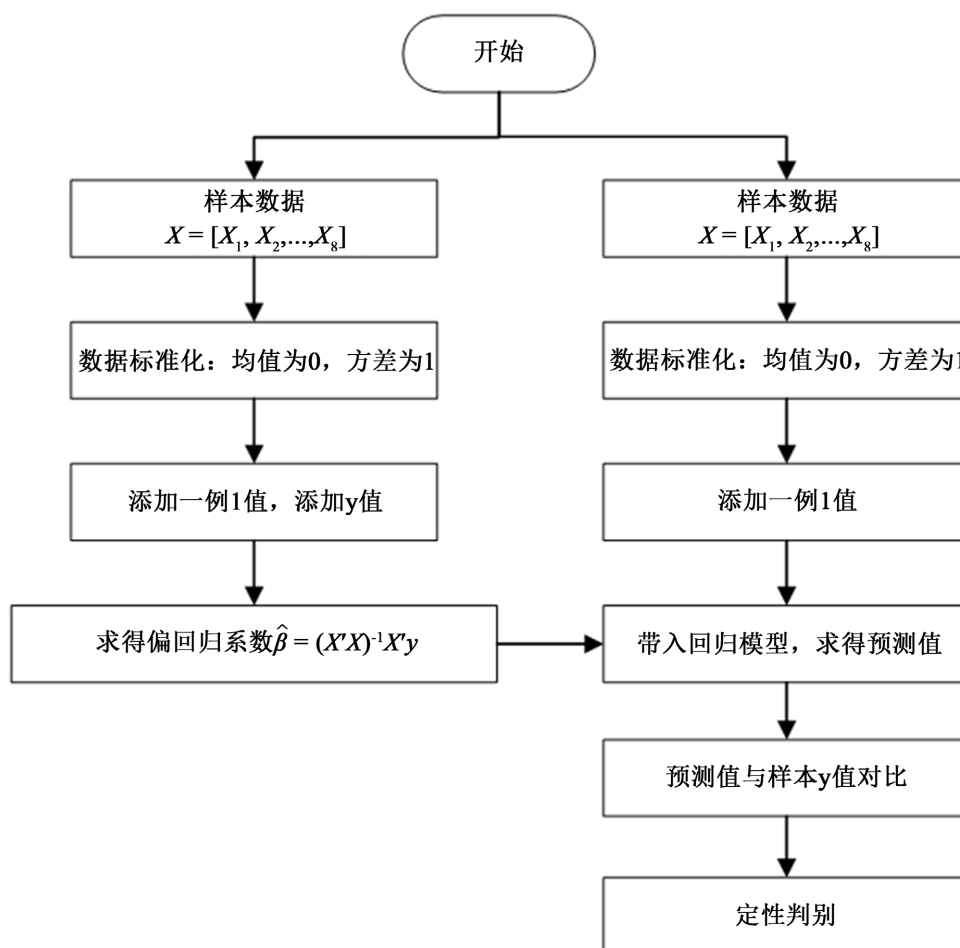


Figure 2. Flowchart of multiple linear regression

图 2. 多元线性回归流程图

根据多元线性回归的偏回归系数的求解方程，我们需要先在训练集和测试集的数据前加一列 1 值，对于多元线性回归中还需要给训练集提供 y 值，以便后期更好地区分测试集。在多元回归分析的核心算法中，所使用的训练集和测试集的程序也是不同的。

训练集：

将经过肖维涅准则和标准化之后的训练集数据按行分别进行多元线性回归分析，得出偏回归系数矩阵，并保留偏回归系数矩阵，从而得到多元线性回归模型。

测试集：

将经过肖维涅准则和标准化之后的测试集数据按行分别代入由训练集数据所得出的每一组偏回归系数中。得到 y 的预测值，计算该训练集的 y 值与测试集所得到的预测值 y 的相对误差值的绝对值，并选取最小误差值，进行保存。

4. 实验结果

4.1. 食用油种类测量结果

对于该种模式识别算法的检验证明，使用食用植物油的训练集和测试集进行。从表 1 中可以得到，通过实际的数据检测之后，本研究所采取的 MLR 算法对食用油类别判别结果的准确率可以达到 97%。

Table 1. Actual detection results of multiple linear regression algorithm

表 1. 多元线性回归算法实际检测结果表

食用植物油	测试个数	错判个数
土茶油	11	0
大豆油	9	1
花生油	12	0
油菜籽油	12	1
玉米油	12	1
橄榄油	9	0
葵花籽油	10	0
芝麻油	12	0
核桃油	10	0
山茶油	12	0

4.2. 食用油掺假定量测试结果

在多元线性回归模型的检测中，本论文采取的训练集数据是芝麻油中混杂了大豆油的混合油数据。图 3 是对下位机传送过来的食用油数据进行的检测结果显示，表 2 是基于拓展多元线性回归算法对芝麻油和大豆油掺杂进行的定量分析测量结果。从图 3 中可以看出所设计的多元线性回归扩展模型算法能对常见的食用油进行定量的判别并显示。

从表 2 可知，采用拓展多元线性回归算法的成分含量识别误差基本在 4%左右。该测试结果表明所设计的多元线性回归扩展模型算法能对常见的食用油进行定量的判别。



Figure 3. Visual interface of the recognition system
图 3. 识别系统的可视化界面

Table 2. Extended multivariate linear regression algorithm actual quantitative detection results
表 2. 拓展多元线性回归算法实际定量检测结果

测试样本	测试结果
测试集 1 芝麻油和大豆油掺杂 20%	芝麻油占 0.83, 大豆油占 0.17
测试集 2 芝麻油和大豆油掺杂 20%	芝麻油占 0.80, 大豆油占 0.21
测试集 3 芝麻油和大豆油掺杂 20%	芝麻油占 0.80, 大豆油占 0.20
测试集 4 芝麻油和大豆油掺杂 10%	芝麻油占 0.95, 大豆油占 0.05

5. 结论

本文针对食用油类别定性判定和掺假定量分析问题,设计了一套基于 MLR 的食用油掺假检测的便携式电子鼻系统,实时采集食用油的气味信号,并利用所设计的多元线性回归算法能对食用油类别进行定性判别,实验表明,准确率达 97%。同时针对一般多元线性回归算法对食用油的定量检测不准确的问题,使用优化后的多元线性回归算法,在大豆与芝麻的混合油的定量测试实验中,测量误差小于 4%,验证了该拓展多元线性回归算法能够准确地对常见的食用油掺假进行定性和定量的判别。

基金项目

国家级大学生创新创业训练计划项目(202310338067)。

参考文献

[1] 王瑞元. 2022 年我国粮油产销和进出口情况[J]. 中国油脂, 2023, 48(6): 1-7.
[2] 孔令琦, 宋佳琪, 陈林林, 等. 食用油掺伪鉴别技术及模型建立的研究进展[J]. 食品安全质量检测学报, 2022, 13(19): 6132-6139.
[3] 张苗. 网络化电子鼻技术研究及实现[D]: [硕士学位论文]. 西安: 西北工业大学, 2007.

-
- [4] Majchrzak, T., Wojnowski, W., Dymerski, T., Gębicki, J. and Namieśnik, J. (2018) Electronic Noses in Classification and Quality Control of Edible Oils: A Review. *Food Chemistry*, **246**, 192-201. <https://doi.org/10.1016/j.foodchem.2017.11.013>
- [5] Lu, L., Hu, Z., Hu, X., Li, D. and Tian, S. (2022) Electronic Tongue and Electronic Nose for Food Quality and Safety. *Food Research International*, **162**, Article 112214. <https://doi.org/10.1016/j.foodres.2022.112214>
- [6] 曹森, 赵成飞, 马风伟, 等. 基于电子鼻和 GC-MS 评价不同采收期天麻的芳香品质[J]. 北方园艺, 2019(19): 87-94.
- [7] 张欣, 陈双宜, 谷惠文, 等. 基于电子鼻技术的荆州鱼糕贮藏过程新鲜度预测[J]. 轻工学报, 2022, 37(3): 17-25.
- [8] 冯蕾. 基于电子鼻及低场核磁共振的黄瓜与樱桃番茄新鲜度智能检测研究[D]: [博士学位论文]. 无锡: 江南大学, 2019.
- [9] Wei, X., Shao, X., Wei, Y., Cheong, L., Pan, L. and Tu, K. (2018) Rapid Detection of Adulterated Peony Seed Oil by Electronic Nose. *Journal of Food Science and Technology*, **55**, 2152-2159. <https://doi.org/10.1007/s13197-018-3132-z>
- [10] Ayari, F., Mirzaee-Ghaleh, E., Rabbani, H. and Heidarbeigi, K. (2018) Detection of the Adulteration in Pure Cow Ghee by Electronic Nose Method (Case Study: Sunflower Oil and Cow Body Fat). *International Journal of Food Properties*, **21**, 1670-1679. <https://doi.org/10.1080/10942912.2018.1505755>
- [11] Kishimoto, N. and Kashiwagi, A. (2019) Evaluation of Filtration on Volatile Compounds in Virgin Olive Oils Using an Electronic Nose. 2019 *IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, Fukuoka, 26-29 May 2019, 1-3. <https://doi.org/10.1109/isoen.2019.8823161>
- [12] Jonsson, A., Winqvist, F., Schnürer, J., *et al.* (1997) Electronic Nose for Microbial Quality Classification of Grains. *International Journal of Food Microbiology*, **35**, 187-193. [https://doi.org/10.1016/S0168-1605\(96\)01218-4](https://doi.org/10.1016/S0168-1605(96)01218-4)
- [13] 赵九明. 肖维涅(Chauvenet)系数的拟合公式[J]. 大学物理, 1987, 1(3): 20.