

聚类算法结合 LIF 技术用于葡萄酒鉴定的研究

周孟然, 王骋*, 胡锋, 来文豪, 卞凯

安徽理工大学电气与信息工程学院, 安徽 淮南 232001

摘要 针对现有葡萄酒检测技术无法快速、高效鉴别葡萄酒品质的问题, 基于无需建立标签、调参简单的聚类算法, 提出了一种利用激光诱导荧光技术获取葡萄酒光谱信息的方法, 实现了酒样鉴别。选取三个品牌、两个年份的四个葡萄酒样本, 在同一酒精度的前提下, 与水进行 1:10 体积配比后, 对每个酒样采集 100 个光谱信息。利用 K -均值、自组织竞争网络和自组织特征映射神经网络三个聚类算法进行酒样的鉴别, 实验结果表明: 在光谱信息分析中, 三个聚类算法均表现出较优越的性能, 识别准确率均达到 99% 以上, 自组织特征映射神经网络的分类准确率更是达到了 100%, 平均用时 5.875 s, 具有较高的鲁棒性和泛化能力。研究结论证明聚类算法对葡萄酒品质的检测是切实可行的。

关键词 光谱学; 激光诱导荧光; 聚类算法; 荧光光谱识别; 葡萄酒; 自组织特征映射神经网络

中图分类号 O657.3

文献标志码 A

doi: 10.3788/LOP202158.1630001

Clustering Algorithm Combined with LIF Technology for Wine Identification

Zhou Mengran, Wang Cheng*, Hu Feng, Lai Wenhao, Bian Kai

College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China

Abstract Aiming at the problem that the existing wine detection technology cannot quickly and efficiently identify the quality of wine, this paper proposes a method to capture the spectral information of wine samples by laser-induced fluorescence technology, which is based on the clustering algorithm without label build and complex tuning. Four wine samples from three brands and two vintages were selected. After being mixed with water at 1:10, under the premise of the same alcohol content, 100 spectral informations were collected for each wine sample. We used K -means, self-organizing competition network and self-organizing feature mapping neural network (SO-FMNN) to identify wine samples. The experimental results show that the three clustering algorithms have superior performances in spectral information analysis, and the recognition accuracy rate can reach more than 99%. The classification accuracy rate of SO-FMNN is even 100%, the average time is 5.875 s, and it has high robustness and generalization ability. It is verified that the clustering algorithm for wine quality detection is feasible.

Key words spectroscopy; laser-induced fluorescence; clustering algorithm; fluorescence spectrum recognition; wine; self-organizing feature mapping neural network

OCIS codes 300.2530; 300.6170; 100.5010

收稿日期: 2020-10-19; 修回日期: 2020-11-15; 录用日期: 2020-12-17

基金项目: 国家重点研发计划(2018YFC0604503)、国家“十二五”科技支撑计划重点项目(2013BAK06B01)、安徽省科技重大专项(201903a07020013)、安徽省自然科学基金青年项目(1808085QE157)

通信作者: *Engr_wangcheng@163.com

1 引言

随着社会的发展,葡萄酒逐渐活跃在人们的生活中,也越来越受到人们的喜爱^[1]。葡萄酒市场的火爆不可避免地带来假酒,假酒不仅让市场经济受损,更对消费者的健康造成恶劣的影响。由于葡萄酒的生产过程及成分较复杂且易受环境因素的影响,如何快速准确地鉴定葡萄酒品质是亟需解决的问题。

目前,葡萄酒的鉴别主要为人工识别法,该方法的准确率与识别速度不高。其他方法如电子鼻分析检测法^[2-3]、气相色谱法^[4-5]以及质谱法^[6-7]因灵敏度不高、检测不全面而难以得到推广。激光诱导荧光(LIF)检测方法具有快速、准确、灵敏度高等特点,目前已用于气体浓度检测、液体水源检测、医学疾病诊断等。文献[8-11]介绍了激光诱导荧光在各种油类识别领域中的应用,LIF技术在定量分析、分类识别中表现出优越的性能。文献[12]利用聚类算法对蜂蜜样本进行了快速识别。文献[13]利用深度学习结合LIF技术对白酒酒样进行了识别,但该方法对硬件要求较高且调参较复杂、耗时较多。聚类算法属于无监督学习算法,无需建立标签,省去了特征预处理,相比有监督学习算法,具有参数调节简单、用时相对较少及对硬件设备要求较低等优点。文献[14-17]利用各类算法并结合LIF技术对矿井突水水源进行了检测,对不同水源种类的鉴别准确率均能达到100%。本文将聚类算法与LIF技术结合,实现了葡萄酒样的快速准确分类,为葡萄酒样的鉴定提供了一种可行的方法。

2 材料与方 法

2.1 激光诱导荧光技术

葡萄酒样的激光诱导荧光机理为分子诱导荧光机理,相比原子诱导荧光机理更加复杂。假设分子被激发后,由振动-转动能级 (v'_k, J'_k) 跃迁到达低能级 (v'_j, J'_j) 并发射荧光 I_{kj} ,其中 k, j 为能级, v'_k, v'_j 为振动量子数, J'_k, J'_j 为转动量子数。 I_{kj} 可表示为

$$I_{kj} \propto N_k A_{kj} h \nu_{kj}, \quad (1)$$

式中: N_k 为布居数密度; A_{kj} 为跃迁几率; ν_{kj} 为频率。跃迁几率 A_{kj} 可表示为

$$A_{kj} \propto |\mathbf{R}_e|^2 |\mathbf{R}_{\text{vib}}|^2 |\mathbf{R}_{\text{rot}}|^2, \quad (2)$$

式中: \mathbf{R}_e 为电子矩阵; \mathbf{R}_{vib} 为振动矩阵; \mathbf{R}_{rot} 为转动矩阵。因此,三个矩阵的矩阵元不同时为0是荧光发生的必要条件。

针对分子荧光光谱的特性,荧光光谱的采集步骤为:固定发射激光的强度,不断调节发射激光的波长,并采集某一波长位置处的荧光强度,即记录下不同波长位置处葡萄酒分子的荧光强度。不同品牌以及年份的葡萄酒所含的化学成分及含量各不相同,这为分子荧光光谱的采集提供了可能。

2.2 仪器与设备

本文选用发射波长为405 nm的蓝紫光半导体激光器及美国Ocean optics公司的USB2000+光谱仪采集葡萄酒样数据,光谱仪的像素为2048 pixel,接收波段为400~790 nm,荧光光谱检测的全波段范围为340.72~1021.086 nm。实验步骤如图1所示,激光器发射的激光经由荧光光纤、石英光纤和浸入式微型荧光探头射入酒样中,探头接收回传的激光并将其送至光谱仪,上位机采集汇总数据,从而得到光谱数据。

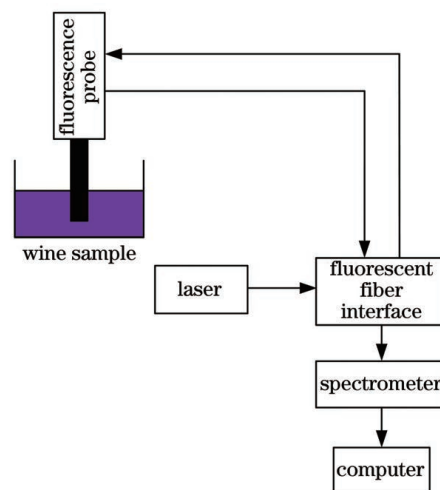


图1 LIF系统示意图

Fig. 1 Schematic of LIF system

2.3 材料和酒样

本文选取了四种酒精度(体积分数,全文同)均为12%的赤霞珠干红葡萄酒进行研究,分别为生厂商未知的葡萄酒(记为酒样A)、张裕葡萄酒一年(记为酒样B)、长城葡萄酒一年(记为酒样C)和长城葡萄酒五年(记为酒样D)。每种酒样各采集100组样本,共计400组。

由于葡萄酒色泽浓郁,若用原酒样进行激光诱导荧光实验会出现激光探头接收不到反射荧光的情况,因此本文对葡萄酒进行稀释操作。在四种葡萄酒样中随机抽取一个酒样进行配比实验,其中酒样C与纯净水在不同体积配比(V_r)情况下的荧光光谱如图2所示。由图2(a)可以明显看出,随着纯净水占比的逐渐增大,400~700 nm波段内的波峰强度

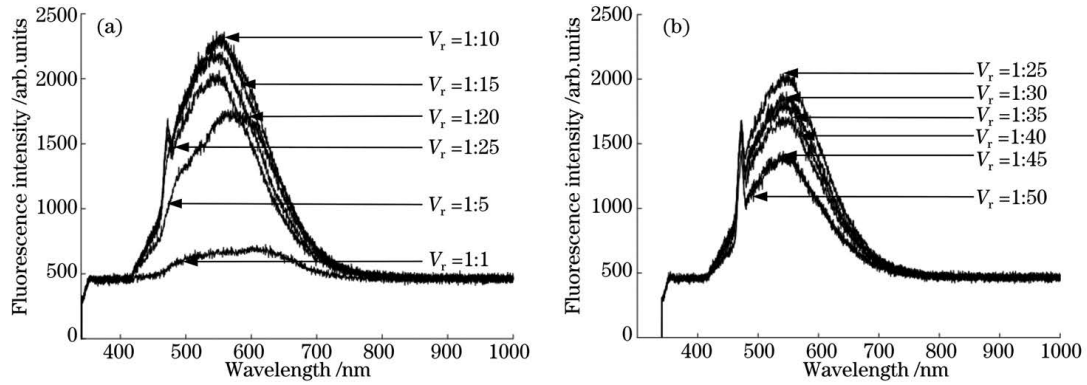


图 2 不同 V_r 下酒样 C 的光谱图。(a) 第一阶段 V_r ; (b) 第二阶段 V_r

Fig. 2 Spectra of wine sample C under different V_r . (a) V_r in first stage; (b) V_r in second stage

逐渐增大,并在葡萄酒与纯净水体积配比 V_r 为 1:10 时达到最大值。由图 2(b) 容易看出,当 V_r 超过 1:10 之后,400~700 nm 波段内的波峰强度逐渐减小。因此,结合图 2(a)、(b),本文所有的葡萄酒样与纯净水的体积配比均为 1:10。

本文在环境不变的情况下对每个葡萄酒样采集了 100 组样本,共计 400 组。图 3 为四种葡萄酒样各 100 个样本的荧光光谱图。

由图 3 可以看出,不同酒样荧光强度的差异主要集中在 450~650 nm 波段,其中酒样 A 与其他酒样的差异肉眼可辨,其他三种酒样的光谱大致相似,酒样 C 与酒样 D 的光谱不易区分。

为了更清晰地观察不同品牌、年份的葡萄酒样的荧光光谱,截取了酒样 B、C、D 在 450~650 nm 波段的局部光谱图,如图 4 所示。图 4(a) 是酒样 B 与酒样 C 的光谱对比,两个酒样的年份相同但品牌不同,荧光强度略有差异。图 4(b) 是酒样 C 与酒样 D

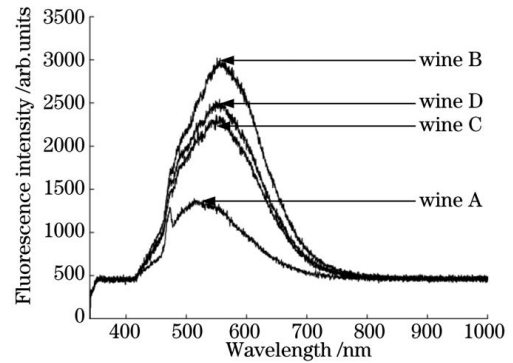


图 3 不同葡萄酒样的光谱

Fig. 3 Spectra of different wine samples

的光谱对比,两个酒样的品牌相同但年份不同,荧光强度与波峰位置非常近似。

为了证明 LIF 技术与聚类算法的结合在酒样鉴定方面的可行性,本文不对输入数据进行预处理。为了尽量贴合实际应用场景,输入的 400 组酒样样本数据均随机分布。

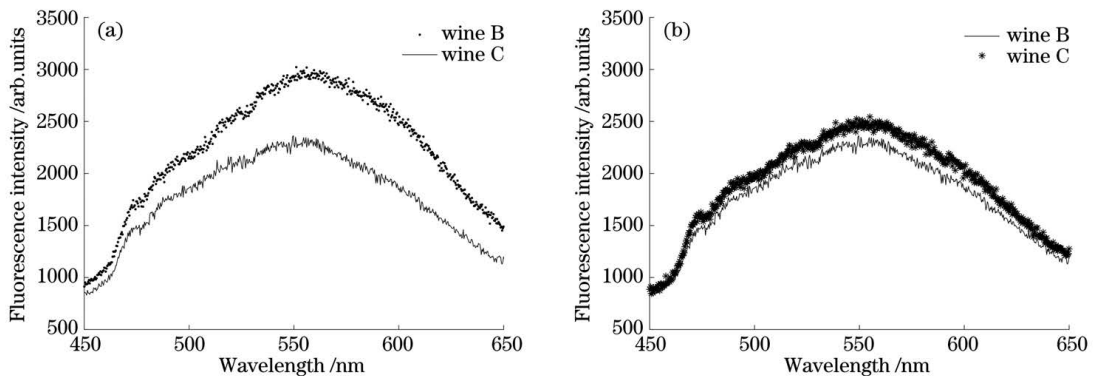


图 4 不同酒样在 450~650 nm 波段的光谱对比。(a) 酒样 B 与酒样 C; (b) 酒样 C 与酒样 D

Fig. 4 Spectral comparison for different wine samples at 450-650nm. (a) Samples B and C; (b) samples C and D

3 聚类算法

聚类算法是一种无标签无监督的算法,目标是对感知样本间的相似度进行类别归纳。本文采取聚

类算法中常用的 K -均值(K -Means)、自组织竞争网络(Self-Organizing Competition Network)以及自组织特征映射(Self-Organizing Feature Mapping, SOM)神经网络来探索 LIF 技术在葡萄酒样鉴别方

面的可行性。

在进行算法介绍之前,定义输入数据集为 $X = \{x_i | x_i \in \mathbb{R}^p, i = 1, 2, \dots, n\}$, 其中 x_i 为输入样本, \mathbb{R}^p 为实数域, p 为维度, i 为样本标号, n 为样本总数。

3.1 K-均值

K-均值算法作为最简单高效的聚类算法,核心思想是:指定 K 个初始质心作为聚类的类别,重复迭代直至算法收敛。

算法流程如下。

1) 选择 K 个类别 (C_1, C_2, \dots, C_K);

2) 从数据集中随机选择 K 个聚类中心 ($\mu_1, \mu_2, \dots, \mu_K$);

3) 计算数据集中的每个点向量 x_i 到聚类中心向量 μ_q 的欧氏距离 $d_{iq} = \sqrt{(x_i - \mu_q)^T (x_i - \mu_q)}$, 根据 d_{iq} 判别该点的类别属性;

4) 重复迭代聚类中心直至满足收敛要求。

K-均值算法容易实现,调参简单,对于簇之间区分明显的数据集识别效果很好。但它需要提前确认类别个数,对于未知样本不适用,且初始质点选取具有随机性,不同的选取方法对结果的影响不同。

3.2 自组织竞争网络

自组织竞争网络分为输入层和竞争层,竞争层中的神经元互相竞争,唯一获胜的神经元代表当前输入样本的分类属性。

算法流程如下。

1) 参数(网络权值 ω 、最大训练次数 T 、学习率 η 等)初始化;

2) 对输入样本向量 \mathbf{Y} 和内星权向量 \mathbf{W}_m 进行归一化,其中 m 为内星权向量的个数;

3) 确定获胜神经元 j^* , 获胜神经元的输出为 1, 其余输出为 0;

4) 利用获胜神经元调整 \mathbf{W}_m , 即

$$\mathbf{W}_m = \begin{cases} \mathbf{W}_m + \Delta \mathbf{W}_m, & m = j^* \\ \mathbf{W}_m, & m \neq j^* \end{cases};$$

5) 判断终止条件,达到最大训练次数后训练自行终止,否则返回步骤 2)。

自组织竞争神经网络可以在很少的训练次数下达到较好的效果,对数据进行分类时首先要设定输入向量的类别总数,再确定类别个数。但每次激发的神经元不一样,相似类别激发的神经元邻近,差别很大的类别激发的神经元距离也比较远。

3.3 自组织特征映射神经网络

SOM 神经网络由输入层和竞争层组成,与自组织竞争神经网络相比,SOM 神经网络的优势是可以学习输入向量的拓扑结构。

算法流程如下。

1) 随机设定输入层与竞争层连接权值的初始值并进行初始化处理,建立初始优胜邻域 N_{j^*} (0) 和学习率 η 的初值;

2) 计算竞争层的权值向量与输入向量之间的欧氏距离;

3) 修正输出神经元 j^* 与邻接神经元的权值 ω_{ij} ;

4) 判定结束条件,若学习率低于设定的最低学习率,则自行终止,否则返回步骤 2)。

SOM 神经网络虽然能够进行有效分类,但学习速度和最终权值的选择要根据不同的需求进行折中处理,因此需要根据不同的需求选择合适的参数。

4 模型建立与结果分析

4.1 聚类模型的建立

为了对比不同聚类算法的识别效果,本文选取以下三种聚类算法:K-均值、自组织竞争网络和自组织特征映射神经网络。所有算法都运行 10 次并取平均值以尽可能消除随机性对结果的影响,为了清晰表述分析类别与实际类别的差异,效果图中的“o”是输入样本的实际类别,“*”为模型预测的结果,若“*”与“o”重合,则表示样本被正确识别。

4.1.1 K-均值算法

将 400 个样本随机分配输入后,进行 K-均值聚类,由于已知四种酒样,K-均值算法初始设置类别数量为 4,重复次数为 10。识别效果如图 5 所示,可以看出,所有样本均被正确识别,聚类准确率为 100%,平均运行时间为 0.135 s。

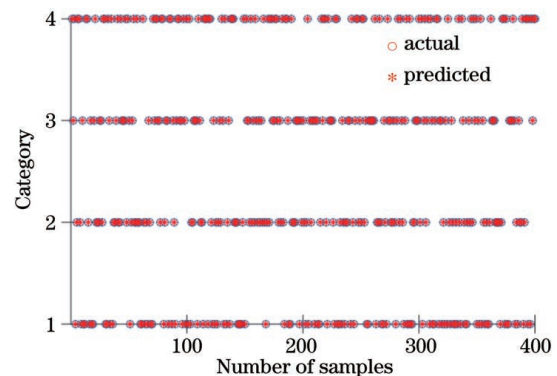


图 5 K-均值算法的识别效果

Fig. 5 Recognition effect of K-means algorithm

4.1.2 自组织竞争网络

自组织竞争网络在分析过程中将酒样数据随机划分训练集和测试集。其中每 100 组样本,80 组为训练集,20 组为测试集,竞争网络初始参数采用默认数值,神经元参数为 4,迭代次数设置为 200。同样取 10 次运行结果的平均值,所得训练集的准确率为 99.34%,测试集的准确率为 99.63%,平均运行时间为 33 s。分类准确率如表 2 所示,测试集效果

表 2 训练集和测试集的分类准确率

Table 2 Classification accuracy rates of training and test sets

Set	Accuracy rate / %
Training	99.34
Test	99.63

图如图 6 所示。如图 6(a)所示,样本被 100% 准确识别;如图 6(b)所示,1 个样本被错误识别。

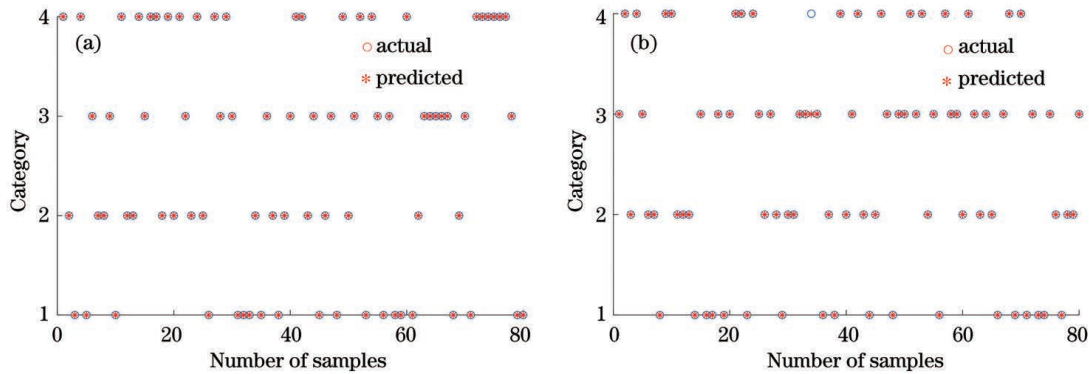


图 6 自组织竞争网络测试集的分类效果。(a)第三次识别效果;(b)第五次识别效果

Fig. 6 Recognition effects of self-organizing competition network testing set. (a) The third recognition effect; (b) the fifth recognition effect

4.1.3 自组织特征映射神经网络

自组织特征映射神经网络在分析过程中将网络初始参数设置为默认数值,将网络层维数设定为 4×1 ,迭代次数设置为 200。运行 10 次后的效果图如图 7 所示。对于四种酒样,每种 100 个样本,均已被分类至四个不同输出神经元,识别准确率为 100%,平均运行时间为 5.875 s。

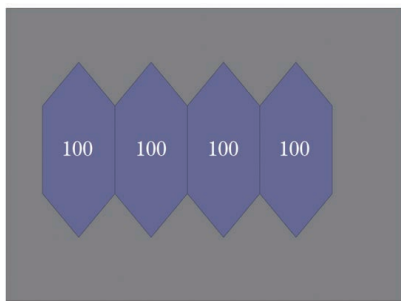


图 7 SOM 神经网络的识别效果

Fig. 7 Recognition effect of SOM neural network

为了贴近实际应用场景,探索在葡萄酒样未知的情况下 SOM 神经网络的有效性,将网络层的维数设定为 5×1 ,其他参数不变。效果图如图 8 所示,可以看出,对于四种酒样,每种 100 个样本,均已被分类至五个输出神经元中的四个,另一个输出神经元未得到样本,酒样被 SOM 算法识别为四类,这表明 SOM 神经网络在样本种类与网络层维数不一

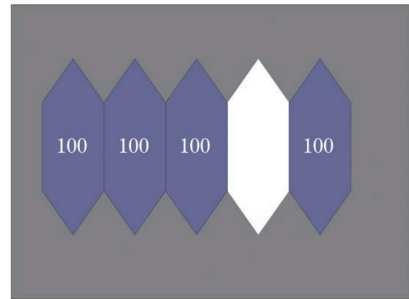


图 8 5×1 SOM 神经网络的效果

Fig. 8 Recognition effect of 5×1 SOM neural network

致时,同样表现出优越的性能。

为了更直观地展示本文所选取的葡萄酒样在聚类过程中的距离差异,作出四种葡萄酒样的聚类谱系图,如图 9 所示。

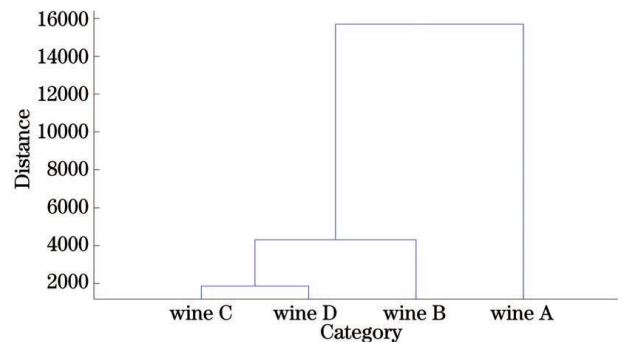


图 9 聚类谱系图

Fig. 9 Cluster pedigree

4.2 结果分析

综上所述,聚类模型在酒样光谱信息处理方面表现良好,本文未对数据进行任何预处理,识别工序简单。如表 3 所示, K -均值算法与自组织特征映射神经网络在识别准确率上表现更好,但自组织特征映射神经网络的鲁棒性更优越。

表 3 聚类算法的性能比较

Table 3 Performance comparison among clustering algorithms

Algorithm	Average recognition rate /%
K -means algorithm	100
Self-organizing competitive network	99.34
SOM neural network	100

5 结 论

首先利用 LIF 技术获取不同品牌、年份葡萄酒的光谱信息,然后将光谱信息输入到聚类算法中进行分析识别,最后通过对比 K -均值算法、自组织竞争网络和自组织特征映射神经网络三种算法的性能,得出如下结论。

1) 以酒精度为 12% 的赤霞珠干红葡萄酒为统一标准,按照不同年份、品牌挑选四种酒样,将透光性弱的原酒与水进行配比,对得到的光谱进行无监督分类学习。结果表明,对于不同品牌、年份、酒精度的酒样,LIF 光谱存在差异。

2) 利用无监督学习的三种算法即 K -均值算法、自组织竞争网络和自组织特征映射神经网络识别葡萄酒样,验证算法对酒样识别的有效性,通过对比发现, K -均值算法在事先已知葡萄酒样数量时表现最好,自组织竞争网络的识别效果稍差,自组织特征映射神经网络在酒样识别(即使葡萄酒样数量未知)中表现出较理想的分类效果。

3) 无监督学习在分析未处理数据时表现良好,相比深度学习,具有调参简单、对硬件要求不高且耗时较少的优点。

葡萄酒样的快速准确识别对葡萄酒市场具有重要意义,激光诱导荧光技术结合无监督学习的方法对其他液体水源的检测也具有一定的参考价值。

参 考 文 献

- [1] OIV. State of the vitiviniculture world market [EB/OL]. (2019-04-11) [2020-10-15]. <http://www.oiv.int/js/lib/pdfjs/web/viewer.html?file=/public/medias/6679/en-oiv-state-of-the-vitiviniculture-world-market-2019.pdf>.
- [2] Zhang S M, Yang Y, Ni Y Y. Combination of near infrared spectroscopy and electronic nose for alcohol quantification during the red wine fermentation [J]. *Spectroscopy and Spectral Analysis*, 2012, 32(11): 2997-3001.
张树明, 杨阳, 倪元颖. 近红外光谱和电子鼻技术用于葡萄酒发酵过程中酒精度的定量分析 [J]. *光谱学与光谱分析*, 2012, 32(11): 2997-3001.
- [3] Banerjee R, Tudu B P, Bandyopadhyay R, et al. Application of electronic nose and tongue for beverage quality evaluation [M] // Grumezescu A M, Holban A M. *Engineering tools in the beverage industry. Volume 3: the science of beverages*. Amsterdam: Elsevier, 2019: 229-254.
- [4] Vyviurska O, Špánik I. Assessment of Tokaj varietal wines with comprehensive two-dimensional gas chromatography coupled to high resolution mass spectrometry [J]. *Microchemical Journal*, 2020, 152: 104385.
- [5] Zhou J G, Yang X S, Zhu X, et al. Influence of different impregnation process of cabernet sauvignon dry red wine aroma [J]. *Journal of Food Science and Biotechnology*, 2019, 38(9): 50-59.
周继巨, 杨学山, 祝霞, 等. 不同浸渍工艺对赤霞珠干红葡萄酒香气的影响 [J]. *食品与生物技术学报*, 2019, 38(9): 50-59.
- [6] Ji C, Feng F, Chen Z X, et al. Determination of five synthetic sweeteners in wines using high performance liquid chromatography-tandem mass spectrometry [J]. *Chinese Journal of Chromatography*, 2010, 28(8): 749-753.
嵇超, 冯峰, 陈正行, 等. 高效液相色谱-串联质谱法测定葡萄酒中的 5 种人工合成甜味剂 [J]. *色谱*, 2010, 28(8): 749-753.
- [7] Šuklje K, Carlin S, Antalick G, et al. Regional discrimination of Australian shiraz wine volatome by two-dimensional gas chromatography coupled to time-of-flight mass spectrometry [J]. *Journal of Agricultural and Food Chemistry*, 2019, 67(36): 10273-10284.
- [8] Zhou M R, Wang J G, Song H P, et al. Application of kernel extreme learning machine and laser induction fluorescence technique in edible oil identification [J]. *Laser & Optoelectronics Progress*, 2020, 57(20): 173002.
周孟然, 王锦国, 宋红萍, 等. 核极限学习机和激光诱导荧光技术在食用油识别中的应用 [J]. *激光与光电子学进展*, 2020, 57(20): 173002.
- [9] Zhu Q S, Hao S G, Luo N N, et al. Detection and

- quantification of vegetable oil adulteration based on laser-induced fluorescence spectroscopy[J]. Chinese Journal of Lasers, 2019, 46(12): 1211002.
- 朱泉水, 郝仕国, 罗宁宁, 等. 基于激光诱导荧光的植物油掺假检测与量化分析[J]. 中国激光, 2019, 46(12): 1211002.
- [10] Li T, Chen S Y, Zhang Y C, et al. Quantification of adulterated extra virgin olive oil using laser induced fluorescence[J]. Spectroscopy and Spectral Analysis, 2018, 38(S1): 285-286.
- 李婷, 陈思颖, 张寅超, 等. 激光诱导荧光初榨橄榄油掺杂定量分析[J]. 光谱学与光谱分析, 2018, 38(S1): 285-286.
- [11] Chen Z K, Guo R, Cheng P F. Application of LIF technology-based spectral feature extraction in oil detection [J]. Laser & Optoelectronics Progress, 2020, 57(13): 133002.
- 陈至坤, 郭蕊, 程朋飞. 基于 LIF 技术的光谱特征提取在油类检测中的应用[J]. 激光与光电子学进展, 2020, 57(13): 133002.
- [12] Zhang X H, Zhang S Y, Qing X D, et al. A new strategy for rapid classification of honeys by simple cluster analysis method based on combination of various physicochemical parameters [J]. Chemical Research in Chinese Universities, 2019, 35(3): 390-394.
- [13] Lai W H, Zhou M R, Wang Y, et al. Application of counterfeit liquor recognition based on deep learning and laser induced fluorescence[J]. Laser & Optoelectronics Progress, 2018, 55(4): 043001.
- 来文豪, 周孟然, 王亚, 等. 深度学习与激光诱导荧光在假酒识别中的应用[J]. 激光与光电子学进展, 2018, 55(4): 043001.
- [14] Lai W H, Zhou M R, Li D T, et al. Application of unsupervised learning AE and MVO-DBSCAN combined with LIF in mine water inrush recognition [J]. Spectroscopy and Spectral Analysis, 2019, 39(8): 2437-2442.
- 来文豪, 周孟然, 李大同, 等. 无监督学习 AE 和 MVO-DBSCAN 结合 LIF 在煤矿突水识别中的应用[J]. 光谱学与光谱分析, 2019, 39(8): 2437-2442.
- [15] Zhou M R, Hu F, Yan P C, et al. Laser induced fluorescence spectrum analysis of water inrush in coal mine based on FCM[J]. Spectroscopy and Spectral Analysis, 2018, 38(5): 1572-1576.
- 周孟然, 胡锋, 闫鹏程, 等. 基于 FCM 的煤矿突水激光诱导荧光光谱分析[J]. 光谱学与光谱分析, 2018, 38(5): 1572-1576.
- [16] Zhou M R, Li D T, Hu F, et al. Research of the AdaBoost arithmetic in recognition and classifying of mine water inrush sources fluorescence spectrum[J]. Spectroscopy and Spectral Analysis, 2019, 39(2): 485-490.
- 周孟然, 李大同, 胡锋, 等. AdaBoost 算法在矿井突水水源的荧光光谱识别中的研究[J]. 光谱学与光谱分析, 2019, 39(2): 485-490.
- [17] Zhou M R, Lai W H, Wang Y, et al. Application of CNN in LIF fluorescence spectrum image recognition of mine water inrush[J]. Spectroscopy and Spectral Analysis, 2018, 38(7): 2262-2266.
- 周孟然, 来文豪, 王亚, 等. CNN 在煤矿突水水源 LIF 光谱图像识别的应用[J]. 光谱学与光谱分析, 2018, 38(7): 2262-2266.