紫外-可见吸收光谱与增量SVM学习相结合的红葡萄酒快速检测方法

刘军，潘铁军，张正勇

**摘要：**红葡萄酒品质会受到自然环境的影响而发生漂移，因此红葡萄酒在线快速识别中面临高效利用新增样本数据中的新特征来更新识别模型的问题。本文提出了一种将紫外-可见吸收光谱与增量支持向量机学习相结合的识别方法，该方法通过增量学习在每次迭代中将新训练样本包含的新特征引入识别模型，在不显著增加训练时间的前提下，快速更新识别模型。实验结果表明：通过该方法建立的识别模型在训练效率与识别准确率上获得了较好的平衡。

**关键词：**紫外光谱；增量学习；I-SVM；红葡萄酒；在线识别

1. **引言**

随着人们对红葡萄酒消费量的持续增长,大量假冒伪劣红葡萄酒大量涌入市场,为了保护红葡萄酒的品牌效应和经济效益,维护消费者的合法权益,对红葡萄酒进行快速识别至关重要。

紫外-可见吸收光谱法【】通过用紫外分光光度计在一定波长范围内对物质进行紫外扫描所得基于分子内电子跃迁产生的吸收光谱。红葡萄酒中酚酸、氨基酸、单宁等物质含量丰富，葡萄原料、陈酿方式和酿造工艺等因素对红葡萄酒的这些成分及不饱和程度产生影响，造成紫外-可见吸收光谱的差异，从而反映了样品的整体特征。该方法具有灵敏度高、重现性好、效率高、成本低等特点，与模式识别算法结合进行检测已被广泛应用于食品、药品等领域，并且获得了较好的准确率。

传统的模式识别识别方法是在大量的标注样本数据集上训练一个完备的分类器用于识别,这种方法属于离线训练方式。红葡萄酒的品质主要由葡萄原料、陈酿方式和酿造工艺等因素决定，而葡萄原料的品质受产地的气候的影响较大，使得同一品牌不同批次的红葡萄酒的品质存在差异，相应的光谱数据也会发生变化，从而使得离线训练的分类器的分类准确率将随着时间发生波动。解决的办法通常是重新训练分类器，重新训练需要大量的训练样本和训练时间，这种代价是难以承受的。因此，如何使得使用旧标签数据训练的分类器适应新样本的识别是在线识别的一个研究难点。

本文针对红葡萄酒在线识别如何自适应新样本的问题，以不同品种和批次的红红葡萄酒为研究对象，提出了将增量学习与紫外光谱相结合的方法，通过增量学习不断地将新训练样本中包含的新特征引入识别模型，在不显著增加训练时间的前提下，快速更新识别模型。实验结果表明：通过该方法建立的识别模型在训练效率与识别准确率上获得了较好的平衡。

1. **实验材料、设备与方法**
   1. 材料与试剂

9种红葡萄酒样品由国内知名电子商务网站购得，每种红葡萄酒样品分别采购6个批次，共54个红葡萄酒样品。

* 1. 仪器与设备

本实验采用的紫外光谱仪器是xxx公司的T6 新世纪型UV-Vis光谱仪，收集紫外-可见吸收光谱，以纯净水为参比。

* 1. 方法
     1. 样品制备

将红葡萄酒样本在30摄氏度、0.0925MPa真空度条件下旋转蒸发15分钟，，收集馏出液作为待测样。

* + 1. 谱图测试（采集）

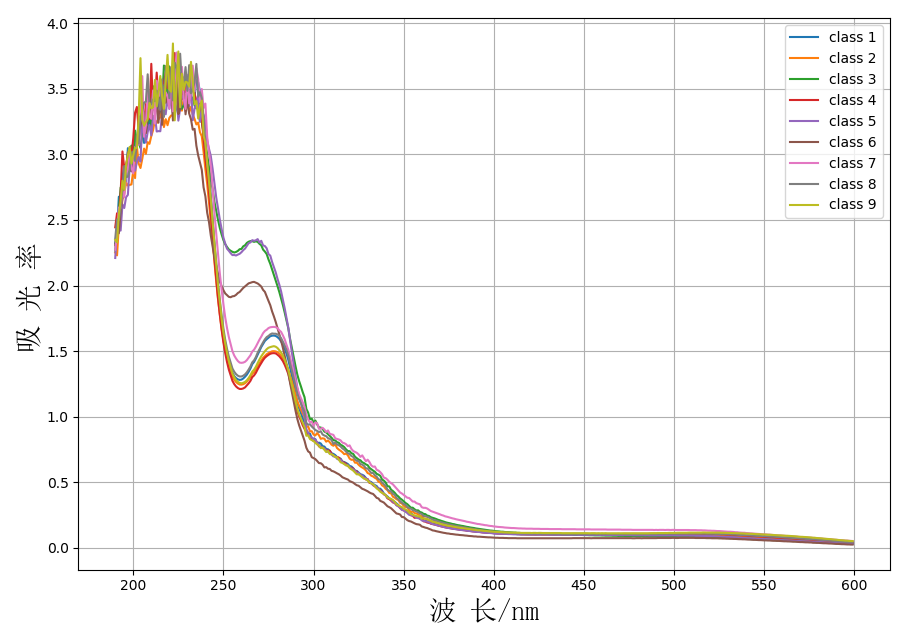


图1 红葡萄酒样品的紫外-可见吸收光谱图

本实验中，对红葡萄酒样本进行扫描，波数扫描范围190~600nm，得到各种红葡萄酒样的紫外-可见吸收光谱图。红葡萄酒样品的扫描时间间隔为0.5秒，积分时间为2秒。由于测量得到的数据中可能包含了一些冗余信息和噪声干扰，影响建模的准确度，需要对获得的光谱波段进行选择。从图1 的原始光谱图中可以看出。红葡萄酒溶液在240~400 nm波段内的吸收光谱数据区分性比较明显，且在230和280 nm附近有一个明显的吸收峰，在260 nm 附近有一个明显的吸收谷。因此，为了更方便更准确地分析红葡萄酒的紫外-可见吸收光谱，本文截取的是波长为190~550 nm 的吸收光谱数据用于建模分析。

由于紫外 - 可见光谱主要反映了红葡萄酒的主要化合物，红葡萄酒的紫外-可见光谱非常相似，难以人工识别。如图1所示。240~550 nm红葡萄酒的光谱类似。在260~275nm附近的UV-Vis吸收可归因于键π轨道与π键轨道之间的pi-pi\*电子跃迁，例如不饱和烃和芳族烃。在传统的光谱分析中，通常使用比尔-朗伯定律，即最大吸收峰值和特定物质之间的关系进行分析，光谱信息的利用率相对较低。事实上，红葡萄酒的紫外 - 可见光谱信息非常丰富，其峰形，面积，宽度等与样品质量密切相关。但是，如果仅仅依靠对光谱的观察，就不宜对红葡萄酒的质量做出直观的判断，因为大多数时候不同的分子对相似的峰值有贡献。为了克服视觉分析的这一限制，统计方法主要用于进一步分析UV-Vis数据。通过统计方法，人们可以通过高亮度的相似性和差异从数据集中提取有用的信息。在这项研究中，我们使用I-SVM进行红葡萄酒的分类，以便高效地处理大量的样本数

* 1. 数据预处理

基于Python语言的Scikit-Learn工具包进行数据处理。为了便于对比增量SVM算法与传统模式识别算法的效率和准确性，使用蒙特卡洛方法将54个原始样本数据扩充为5400个模拟样本数据，用于对比分类算法的效率和准确率。

使用StandardScaler类对样本数据集进行归一化处理，然后将预处理后的红葡萄酒谱图数据随机分为100份（每份包含54个样本数据），使用train\_test\_split函数对每一部分按照6：3的比例分为训练数据集和测试数据，用作初试化分类器和增量学习分类器的训练与测试。采用主成分分析法（Principal Components Analysis，P C A ）对数据进行降维。PCA是常用的高维数据降维方法，通过取累计贡献率大于98%的主成分作为增量SVM分类器的输入变量，降低冗余信息干扰、提高算法运行效率。

* 1. 增量SVM学习模型

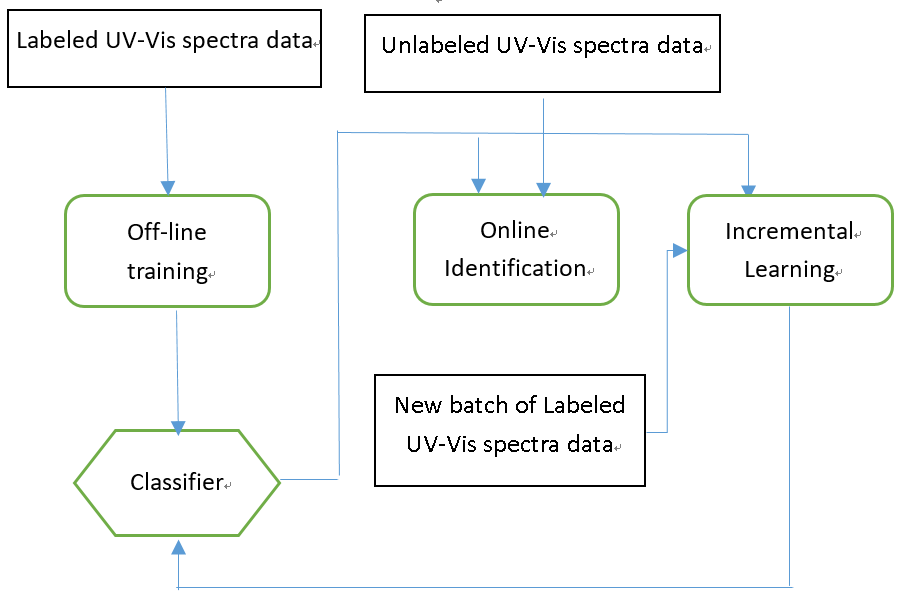


图2 增量SVM算法框架

传统支持向量机是采用固定不变的训练样本形成的固定参数的分类模型来对未来的样本信息进行识别，当新测试样本包含新的特征时，其识别的准确性将会下降。使用，（其中为旧训练样本集合，为新训练样本集合），对识别模型Clf进行训练，又由于传统SVM算法的时间复杂度是（n为样本数量），大量的训练样本使得其训练效率低下。

本文利用增量SVM算法，使用支持向量（Support Vector，SV）代替，与一起对分类模型进行训练，算法框架如图2所示。

具体算法如下：

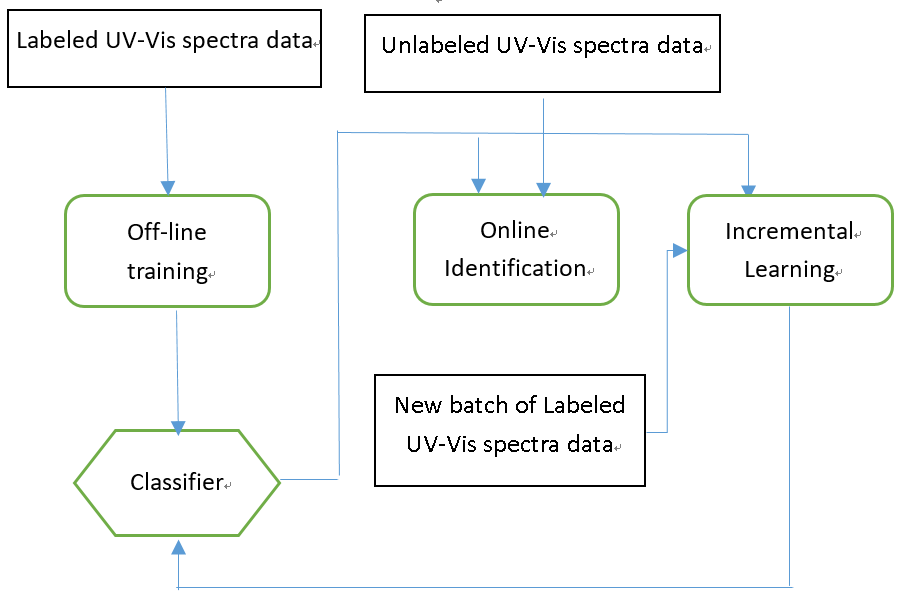
Step 1 使用有标记样本数据集Lini训练得到初始化分类器Clf和支持向量集合SV；

Step 2 获得增量样本数据集,其中L为有标记样本数据集，U为无标记的待分类样本数据集；

Step 3 将有标记样本数据集L与支持向量集合SV合并为训练集，训练分类器Clf，并获得支持向量集合SV；

Step4 使用Clf对待分类样本数据集U进行分类识别；

Step5 转回Step2



1. 结果与分析
   1. 紫外光谱数据分析

结果发现，同红葡萄酒的荧光强度、荧光峰的位置和数量均不同，光光谱数据结合PCA，可以很好地对昌黎产区不同年份、不同酒厂和不同品种的红葡萄酒进行识别。

**分类性能对比**

以ＴＰＲ为纵轴，ＦＰＲ为横轴，绘制标准支持向量机与集成支持向量机的红葡萄酒品质分类器的ＲＯＣ曲线，分别绘制在如图４～５中，曲线下面积记为ＡＵＣ，ＡＵＣ越大则模型分类效果越好．当且仅当ＡＵＣ＞０．５ａｎｄ　ＡＵＣ＜＝１，分类器才是有价值的．ＲＯＣ分析结果显示，标准支持向量机ＡＵＣ约为０．８５±０．０４５，表现为８２．８０％的ＴＰＲ 和１００％的ＦＰＲ，集成支持向量机ＡＵＣ约为０．８９±０．０４５，表现为８６．７２％的ＴＰＲ和１００％的ＦＰＲ，说明基于ＡｄａＢｏｏｓｔ－ＳＶＭ 的红葡萄酒品质分类器有良好的分类性能．

**分类时间对比**

1. 结束语

本实验将紫外光谱与增量学习SVM相结合应用于红葡萄酒的在线识别。红葡萄酒经过xxx处理后，在经紫外光谱仪分析，建立了9种红葡萄酒的紫外光谱指纹图谱库，通过PCA降维后，建立增量SVM模型对红葡萄酒进行识别，红葡萄酒的识别率达到95.31%。同时，通过蒙特卡洛方法模拟100批次样本数据共10000个样本数据进行识别，红葡萄酒的识别率达到93.78%，各批次的识别模型训练的平均时间为xx秒，方差为xx。该方法为在线红葡萄酒的识别提供了一种可靠、稳定、快速、全新的方法，可为红葡萄酒品质评价和质量控制提供方法依据。

参考文献

1. 刘玉平,孙宝国. 我国食用香料香精的基本状况与发展趋势[J]. 食品科学,2004,(10):373-375. [2017-09-28].
2. 孙宝国,田红玉,刘玉平,谢建春,郑福平. 食品香料香精对食品安全的影响[J]. 现代科学仪器,2006,(01):49-51. [2017-09-28].
3. 孔梦红,吴杜轩,陈相柏. 拉曼光谱定性和定量检测青蒿素研究[J]. 光谱学与光谱分析,2017,37(03):778-782. [2017-09-28].
4. 朱颖洁,郭磊,刘易,龚莹,邱泽武,吴剑峰,谢剑炜. 基于壳层隔绝纳米粒子和在线裂解-吹扫捕集的血液氰化物表面增强拉曼光谱快速检测方法[J]. 分析化学,2017,45(05):627-632. [2017-09-28].
5. 赵迎,李明,肖兹兰,任立志,王雷. 基于拉曼光谱快速鉴别新陈大米的方法研究[J]. 光谱学与光谱分析,2016,36(S1):303-304. [2017-09-28].
6. 周秀军. 基于拉曼光谱的食用植物油定性鉴别与定量分析[D].浙江大学,2013.
7. 周秀军,戴连奎,李晟. 基于拉曼光谱的食用植物油快速鉴别[J]. 光谱学与光谱分析,2012,32(07):1829-1833. [2017-10-06].
8. 翟晨,彭彦昆,李永玉,DHAKAL Sagar,徐田锋,郭浪花. 基于拉曼光谱的苹果中农药残留种类识别及浓度预测的研究[J]. 光谱学与光谱分析,2015,35(08):2180-2185. [2017-09-28].
9. 李晶,徐济仓,李雪梅,周建光,朱岩,缪明明. 超高效液相色谱法同时测定香精香料中14种禁限用物质[J]. 色谱,2012,30(08):816-821. [2017-10-02].
10. 邓其馨,黄朝章,张建平,蔡国华,吴清辉,黄华发,许寒春,刘泽春,谢卫. 液相色谱串联质谱法测定烟用香精香料中的亚硝胺[J]. 现代食品科技,2014,30(01):195-199. [2017-10-02]. DOI：10.13982/j.mfst.1673-9078.2014.01.016
11. 李长于,李祖光,周示玉,叶丹凤,刘文涵. 气相色谱-串联质谱法测定香精香料中的香豆素和黄樟素[J]. 质谱学报,2011,32(05):265-270. [2017-10-02].
12. 孟冬玲,刘畅,李小兰. 离子液体双水相萃取-高效液相色谱法测定香精香料中的抗氧化剂[J]. 分析科学学报,2013,29(04):547-550. [2017-10-02].
13. 吴利敏. 近红外光谱法快速检测某些中药及中成药品质的应用研究[D].西南大学,2013.
14. 陈小康,孙素琴,李隆弟. 中药注射剂荧光光谱法的快速鉴别和热稳定性研究[J]. 分析化学,2002,(10):1168-1173. [2017-10-06].
15. 张慧敏,马书荣,王娜,张衍亮. 拉曼光谱法快速检测化妆品[J]. 分析仪器,2016,(01):33-37. [2017-10-06].