紫外-可见吸收光谱与CCH-SVM学习相结合的

红葡萄酒快速检测方法

刘军

（南京财经大学管理科学与工程学院，江苏，南京 210023）

**摘要：**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，在每次增量学习的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-可见吸收光谱；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可见光谱仪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。紫外-可见吸收光谱法具有灵敏度高、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，具有广阔的应用前景。

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干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**

本实验采用的光谱信号采集设备是北京普析通用仪器有限公司生产的T6 新世纪型紫外可见光谱仪，以去离子水为参比。

* 1. **方法**
     1. **样品制备**
     2. 取红酒样品100 μL，置于3 mL去离子水中，混合均匀，即得到待测样品。而后利用紫外光谱仪进行测量，得到相应标号的红酒样品的紫外可见光谱数据。**谱图测试（采集）**

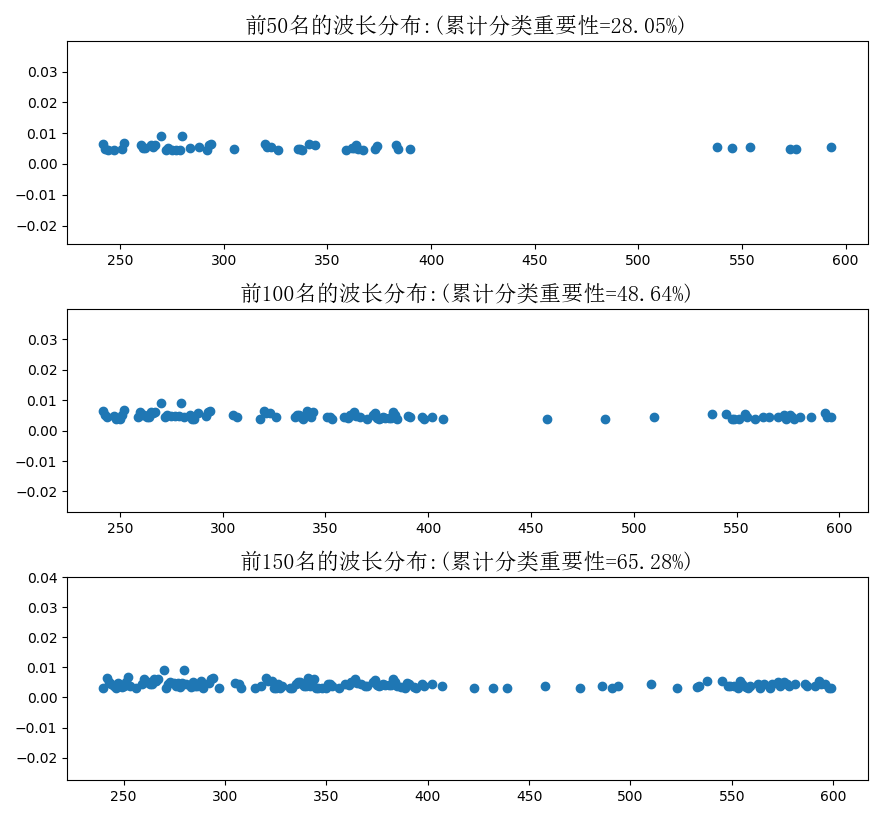


图1 通过信息熵分析波长对分类的贡献度

本实验中，对红葡萄酒样本进行扫描，波数扫描范围190~600 nm，得到红葡萄酒样的紫外-可见吸收光谱图，样品的扫描时间间隔为1 nm。为了更为客观的分析波长对分类的贡献度，我们还通过计算各个波长包含的信息熵对其在分类中的贡献度进行分析。由图1可以看出，对于分类贡献较高的前50个波长主要分布在190~400 nm以及530~600 nm之间，而贡献度在51~150名之间的波长主要分布在400~530 nm之间。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190~600 nm 的吸收光谱数据用于建模分析。

* 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，PCA ）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**

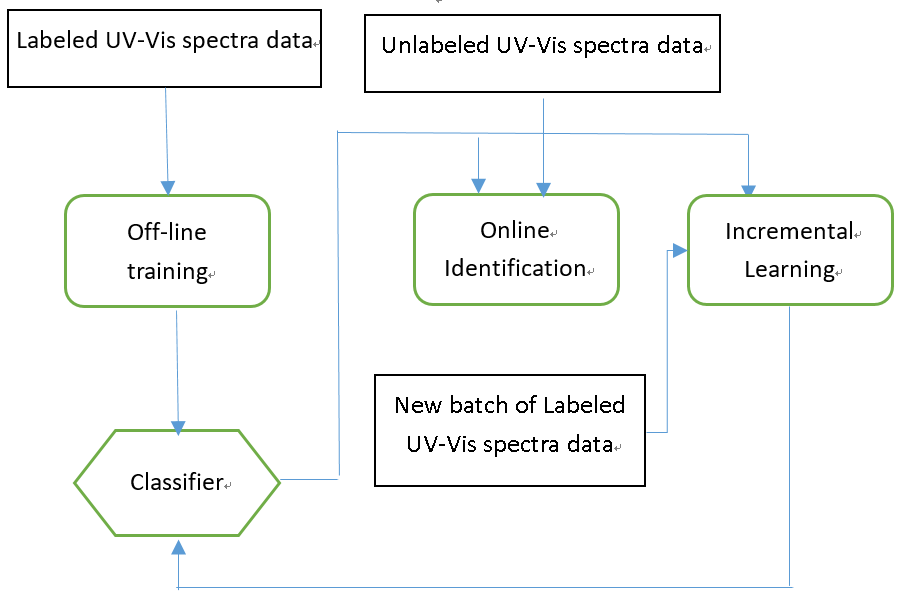


图2 增量SVM算法框架

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CCH-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 使用有标记的初始化样本数据集训练得到初始化分类器Clf和支持向量集合 ；

Step 2 获得增量样本数据集,其中为新获取的标记样本数据集，为新获取的无标记待分类样本数据集；

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

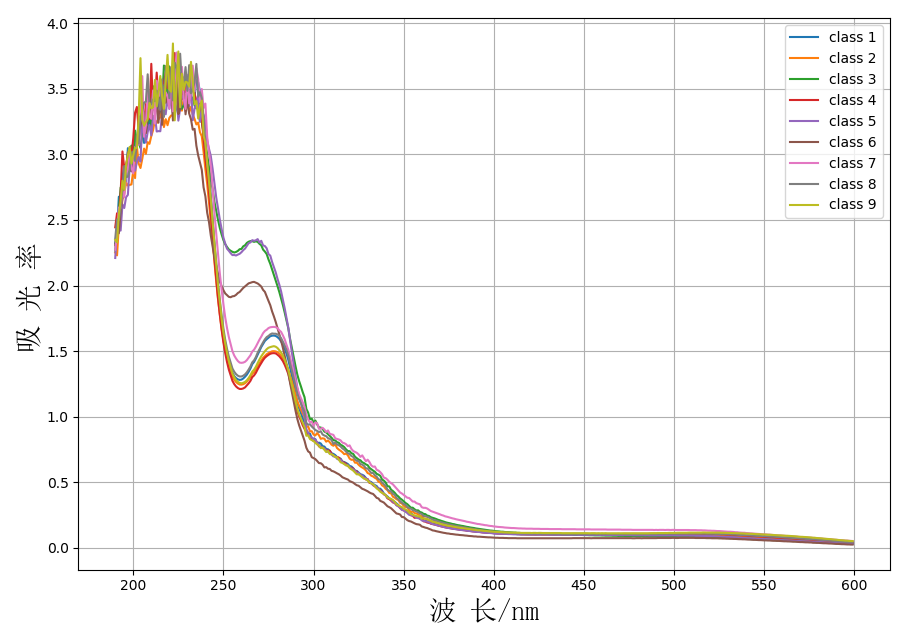


图2 红葡萄酒样品的紫外-可见吸收光谱图

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2 的样品光谱图中可以看出，class 3与class 5的谱图曲线较为相似，class 1、2、4、7、8和9的谱图较为相似。样品谱图由于不饱和双键等的存在，产生的pi-pi\*电子跃迁，在260~275nm附近产生明显的吸收峰，在260 nm附近有一个明显的吸收谷，光谱数据区分性比较明显。

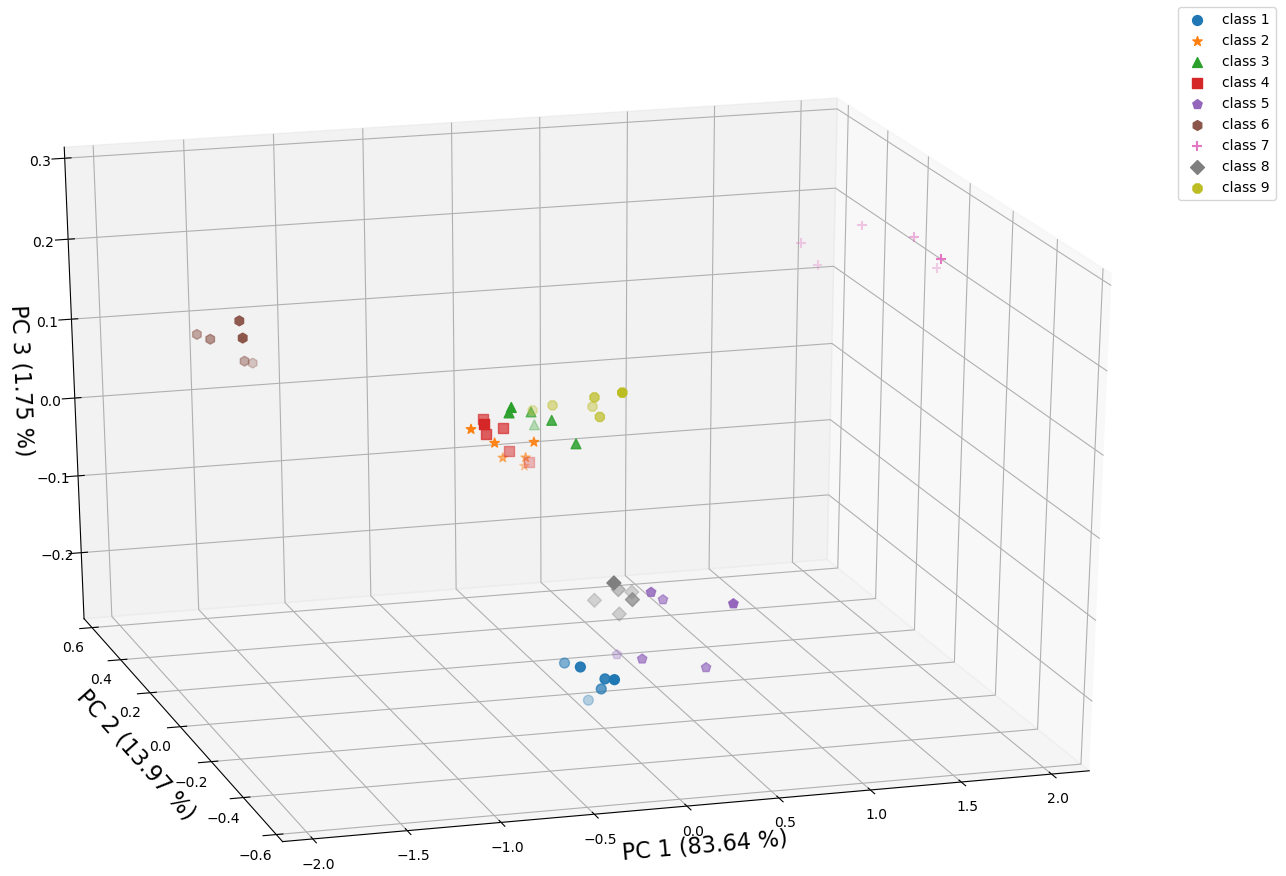
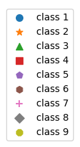


图3 9种红葡萄酒紫外-可视光谱PCA分析图

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，结合模式识别的算法可以从数据中最大限度的提取信息，样品集进行分类。为降低谱图数据的维度，选取190~600 nm波段内的紫外-可见吸收光谱数据为样本，进行PCA降维。由图3所示，9种红葡萄酒中，class2、3、4和9相互重叠聚在一起， class1、5和8相互重叠聚在一起，class 6单独聚为一类。

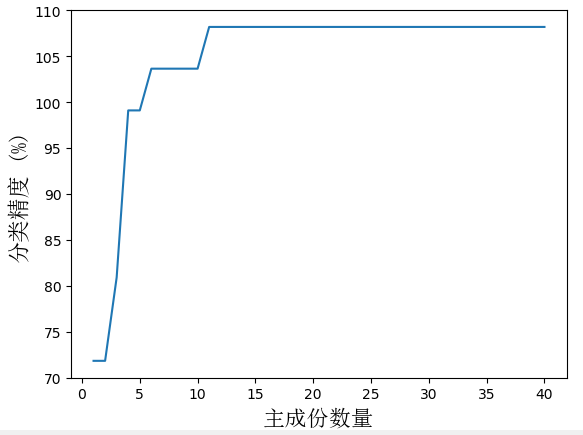


图4 分类精度与主成份数量关系图

由PCA生成的特征空间的维度取决于最终分类的准确率，根据图3，我们使用15个主成份构成新的特征向量，作为I-SVM分类器的输入。

按照图2所示的I-SVM算法流程进行增量学习。分类器的初始化阶段：将初始化训练用的紫外-可见吸收光谱数据进行归一化处理，然后进行PCA降维，降维后的数据维度为15。然后对SVM分类器进行训练，并获得分类器和支持向量集合。增量学习阶段：合并新获取的标记样本数据集与支持向量集合SV，作为训练数据对分类器进行训练。经100次随机实验，每次均进行100次增量学习，增量学习的效果如图5所示，识别率方面：传统SVM分类器的平均识别率为97.27%，增量学习的平均识别率为93.78%；训练时间方面：传统SVM算法评价训练耗时为1.12秒，方差为0.76，增量SVM算法评价训练耗时为0.12秒，方差为0.03。增量SVM学习算法在识别效率方面略低于传统SVM算法，训练耗时为传统SVM算法的13.4%。实验结果表明：增量SVM算法在保证识别率的前提下，显著提高了识别模型的训练效率。

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94.9%。同时，通过蒙特卡洛方法模拟100批次样本数据共5400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0.12秒，方差为0.03。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].
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].