紫外光谱与增量学习SVM结合的在线葡萄酒快速检测方法

基于在线增量学习SVM的葡萄酒增快速识别方法研究

**摘要：**针对葡萄酒在线快速识别中,如何在不增加模型训练时间的前提下,利用新增样本数据中的新特征来更新识别模型的问题,提出了一种紫外光谱与增量学习支持向量机相结合的识别方法。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；CCH-SVM；葡萄酒；在线识别

1. 引言

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

紫外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。紫外光谱曲线的峰形、峰高、峰面积的差异刻画了样本所含物质的成分及成分的不饱和程度的差异，反映了样品整体特征。紫外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，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。葡萄酒的品质主要由葡萄原料和酿造工艺决定，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葡萄酒的口感存在差异，相应的光谱数据也会发生变化，在这种情形下离线训练的分类器的分类准确率将大大降低。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。

基于结构风险最小化原则的支持向量机(Support Vector Machine，SVM),由于具有很强的学习能力和较好的泛化性能,能够较好地解决小样本、高维数、非线性、局部极小等问题,可以有效地进行分类。增量训练由SV样本和新样本组成,再训练只需要进行一次即可完成,所有的非SV样本点都被抛弃。使用SVM增量学习算法,与标准SVM方法相比,可以在保证分类性能的前提下,大大节约计算时间,加快训练速度。

传统的支持向量机方法存在两方面的问题：

1）分类模型无法反映新增数据的特征。

对于新增已标记样本数据{xi,yi}（xi为葡萄酒紫外谱图的特征值，yi为对应的标签），如果该样本数据满足KKT条件，则其包含的特征信息在原分类器中已被反映。当新增已标记样本不满足KKT条件时，有可能使得原样本数据集中的非支持向量转化为支持向量，在这种情况下，原有的支持向量集已经不能刻画最优分类面，必须进行更新以反映最新的特征。

2）更新分类模型的效率低下。

根据SVM方法中对偶最优化问题的求解过程与结果进行分析可知，求解该二次优化问题的时间复杂度为Q3，时间会随着数据的规模增加而成3次指数增长。大量的训练样本数据，也会占用大量的内存，导致分类模型更新的效率低下。

本文针对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种葡萄酒样品由国内知名电子商务网站购得，每种葡萄酒样品分别采购6个批次，共54个葡萄酒样品。

* 1. 仪器与设备

本实验采用的光谱仪型号是Ocean Optics公司的USB2000 + 型UV 光谱仪，它的光谱波段扫描为190~550 nm。光源为氘灯，采样波长间隔约0.34 nm，因此在190~550 nm 波段间总共被分成了2048个波长点。本实验采用的是石英比色皿，其光程长度为1 cm。为减小单次随机误差的影响，每个样本共扫描50次，从而得到每个样本的50组光谱数据，最终谱图数据由其平均而得。

* 1. 方法
     1. 样品制备

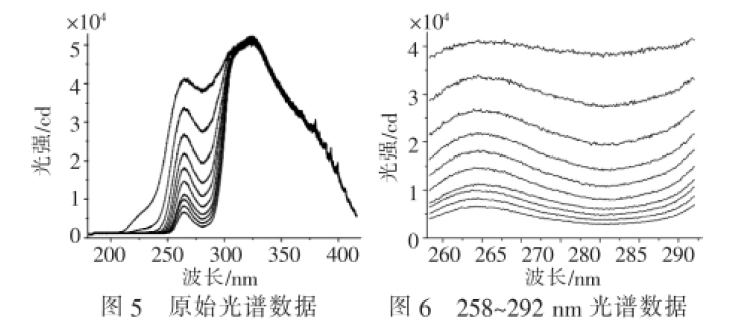
将葡萄酒样本在30摄氏度、0.0925MPa真空度条件下旋转蒸发15分钟，，收集馏出液作为待测样。

* + 1. 谱图测试（采集）

对馏出液进行全波扫描，波数扫描范围190~550nm，得到各种葡萄酒样的紫外-可见吸收光谱图。所有样品的扫描时间间隔为0.5秒，积分时间为2秒。图x为葡萄酒样品的近红外透射光谱，190～550 nm 范围内有较丰富的吸收峰，其中xxx～xxx nm主要是葡萄酒各组分中 Ｃ—Ｈ 键的合频吸收峰，xxx～xxx nm是 Ｃ—Ｈ键的二倍频吸收峰。

本实验中54个样品所获得的光谱波段为190~550 nm。由于测量得到的光谱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。原始光谱数据和258~292 nm 波段的光谱数据如图5、6 所示。从图5 的原始光谱图中可以看出，葡萄酒溶液在xxx~xxx nm波段内的吸收光谱数据区分性比较明显，且在xxx nm附近有一个明显的吸收峰，在282 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葡萄酒的紫外吸收光谱，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。

通过随机森林算法，使用信息熵确定对于分类最为有效的波长范围为：xxx~yyynm。



在实验中，选取10、40、70、100 mg /L 四种TOC 标准溶液的紫外吸收光谱作为训练样本来构建浓度反演模型，将20、30、50、60、80、90 mg /L 六种浓度的紫外光谱数据进行测试，其中每种浓度均有100 组数据。目前较多文献中使用的是254nm 单波长点光谱数据及单波段光谱数据两种方式进行分析建模。为了对比两种方法的有效性，本文对单波长点和单波段光谱数据分别采用普通最小二乘法( ordinary least square，OLS) 、BP 神经网络( back propagation neural network，BP) 、支持向量回归( support vector regression， SVＲ) 方法进行对比，其实验结果如表1、2 所示。表2 中加粗字体表示同一测试浓度条件下的相对误差最大项。在此基础上，本文又对单波段光谱数据采用不同的特征提取方法进行分析对比。特征提取方法主要包括主成分分析法( principal component analysis，PCA) 、深度信念网络、随机子空间+ 深度信念网络等。

* 1. 数据处理

基于Python语言的sklearn包进行数据处理。为了模拟大量数据情形下的增量学习，首先通过添加白噪音的方法，将54个原始光谱样本数据扩充为5400个光谱样本数据，作为样本数据集。使用标准SVM算法对其进行分类测试，验证扩充后的样本的可分性。

使用xx模块中xxx函数对样本数据集进行归一化处理，然后使用model\_select模块中的xx函数将预处理后的葡萄酒谱图数据随机分为10部分（每部分包含540个样本数据），对每一部分按照6：4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。

* 1. 增量SVM学习模型

在大数据在线学习的情境下，减少训练数据集样本的数量是提高

支持向量机方法是建立在统计学习理论的基础上,强调结构风险最小化。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。

本文选择的增量SVM算法，当模型精度

采用把原有的支持向量集和新增样本作为新的训练集的方法,具体增量学习算法如下：

Step 1 使用初始被标记样本数据集L0训练得到初始分类器clf0。

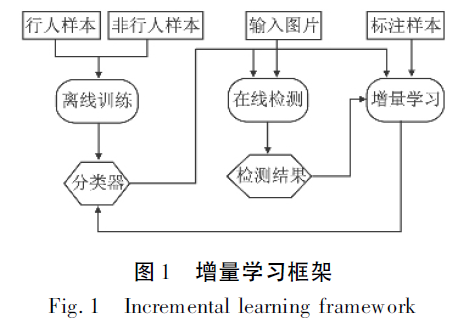
Step 2

1. 结果与分析
   1. 紫外光谱数据分析

葡萄酒中酚酸、氨基酸、单宁等物质含量丰富，些物质与葡萄酒的品种、产地、酿造工艺等有关，都能产生荧光现象，因此可以通过分析葡萄酒的荧光光谱，到葡萄酒产地溯源的目的。

结果发现，同葡萄酒的荧光强度、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相结合应用于葡萄酒的在线识别。葡萄酒经过xxx处理后，在经紫外光谱仪分析，建立了9种葡萄酒的紫外光谱指纹图谱库，通过PCA降维后，建立增量SVM模型对葡萄酒进行识别，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，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。该方法为在线葡萄酒的识别提供了一种可靠、稳定、快速、全新的方法，可为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].