



(12)发明专利申请

(10)申请公布号 CN 109358022 A

(43)申请公布日 2019.02.19

(21)申请号 201811399669.7

(22)申请日 2018.11.22

(71)申请人 陕西中烟工业有限责任公司

地址 710065 陕西省西安市高新区沣惠南路38号

(72)发明人 何媛 李小斌 彭军仓 黄扬明
张凤侠 张萌萌 康世平 孙赵麟
何育萍 师金鸽 王颖 闵顺耕

(74)专利代理机构 西安通大专利代理有限责任公司 61200

代理人 徐文权

(51)Int.Cl.

G01N 21/359(2014.01)

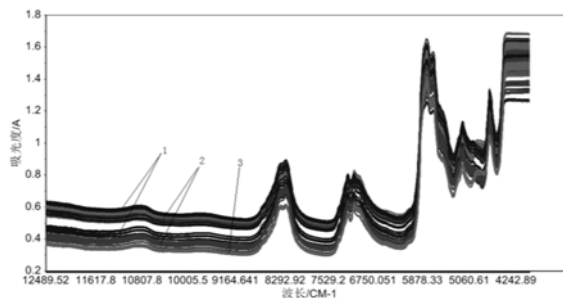
权利要求书2页 说明书7页 附图1页

(54)发明名称

一种快速判别烟用爆珠类型的方法

(57)摘要

本发明公开了一种快速判别烟用爆珠类型的方法,通过对不同类型烟用爆珠进行原始光谱采集,建立不同类型烟用爆珠的标准样本,对采集的不同类型烟用爆珠原始光谱进行预处理,以达到消除各种噪声和干扰,降低样品表面不均匀等因素影响,然后对预处理后的不同类型烟用爆珠光谱进行建立分类模型,利用分类模型对待检测烟用爆珠进行识别,采用近红外光谱技术不仅适用于产品组分的品质控制和定量分析,还能够利用光谱所反映的组成和结构信息进行定性判别,能够快速、准确的对烟用爆珠类型进行判别,样品无需复杂前处理、分析速度快、无损、操作简单方便、模型预测准确率高,对烟用爆珠类别快速分析进行初步探索,并可进一步为烟用爆珠质量分析提供技术支撑。



1. 一种快速判别烟用爆珠类型的方法,其特征在于,包括以下步骤:

步骤1)、对不同类型烟用爆珠进行原始光谱采集;

步骤2)、对采集的不同类型烟用爆珠原始光谱进行预处理;

步骤3)、对预处理后的不同类型烟用爆珠光谱进行建立分类模型,利用分类模型对待检测烟用爆珠进行识别。

2. 根据权利要求1所述的一种快速判别烟用爆珠类型的方法,其特征在于,在室温下对不同类型烟用爆珠样品进行近红外光谱采集,谱区采集范围:12000 cm^{-1} ~4000 cm^{-1} ,光谱分辨率:2~32 cm^{-1} ,扫描次数8~128次,分不同的天对样品进行光谱数据采集。

3. 根据权利要求1所述的一种快速判别烟用爆珠类型的方法,其特征在于,采用矢量归一化方法,对未知光谱 x 进行预处理,其矢量归一化算法如下:

$$X_{\text{normalized}} = \frac{X - \bar{X}}{\sqrt{\sum_{k=1}^m X_k^2}}$$

其中, $\bar{X} = \frac{\sum_{k=1}^m X_k}{m}$, $k=1,2,\dots,m$, m 为波长点数。

4. 根据权利要求1所述的一种快速判别烟用爆珠类型的方法,其特征在于,采用光谱均值中心化变换进行预处理,将样品光谱减去校正集的平均光谱,经过变换的校正集光谱阵的列平均值为零;

首先计算校正集样品的平均光谱 \bar{x} :

$$\bar{x}_k = \frac{\sum_{i=1}^n x_{i,k}}{n}$$

式中, n 为校正集样品数, $k=1,2,\dots,m$, m 为波长点数。

5. 根据权利要求1所述的一种快速判别烟用爆珠类型的方法,对光谱数据进行标准正态变量变换:

标准正态变量变换法的光谱按下式计算:

$$x_{SNV} = \frac{x - \bar{x}}{\sqrt{\frac{\sum_{i=1}^m (x_k - \bar{x})^2}{m-1}}}$$

其中, $\bar{x} = \frac{\sum_{k=1}^m x_k}{m}$, $k=1,2,\dots,m$, m 为波长点数。

6. 根据权利要求1所述的一种快速判别烟用爆珠类型的方法,其特征在于,采用PLS-DA有监督模式识别算法或者SIMCA有监督模式识别方法以及线性判别分析,对三类烟用爆珠建立分类模型。

7. 根据权利要求6所述的一种快速判别烟用爆珠类型的方法,其特征在于,PLS-DA方法:分解光谱数据矩阵与类别数据矩阵,并使得二者最大程度的相关,采用Kennard-Stone法选择代表性样本,基于变量之间的欧氏距离计算得到的代表性样本集,选择最具差异性的样本用于模型的建立。

8. 根据权利要求6所述的一种快速判别烟用爆珠类型的方法,其特征在于,SIMCA方法:对训练集中每一类样本的光谱矩阵分别进行主成分分析,建立每一类的主成分分析数学模

型,然后在此基础上对未知样本进行分类。

9. 根据权利要求6所述的一种快速判别烟用爆珠类型的方法,其特征在于,线性判别分析:分别计算类内和类间的协方差矩阵,计算未知样品 x_{un} 与每类均值的距离平方,然后将其判别到距离最小的类别中即可。

一种快速判别烟用爆珠类型的方法

技术领域

[0001] 本发明属于烟用材料分析技术领域,涉及一种快速判别烟用爆珠类型的方法。

背景技术

[0002] 烟用爆珠在卷烟滤棒中的添加,能够减少外界环境对吸味的影响和造成的香气损失,实现卷烟吸食过程中人为可控的特色香味释放,对于改善卷烟口感,弥补和增强卷烟香气,突出卷烟风格具有重要作用。因此,烟用爆珠的开发和研究对于卷烟企业的发展具有重要意义。目前,对于烟用爆珠的研究主要是新型烟用爆珠内含物的开发与制备、烟用爆珠对主流烟气有害成分释放量和滤嘴截留的影响等。而对于烟用爆珠的评价仅限于物理指标如爆珠直径、胶皮厚度、破碎压力及外观评价。由于卷烟生产自动化程度高,快速通过内含物评价烟用爆珠质量的研究却鲜见报道,而且目前没有针对烟用爆珠的评价有好的效果的方法或方式。因此,建立快速评价烟用爆珠方法具有十分重要的意义。

发明内容

[0003] 本发明的目的在于提供一种快速判别烟用爆珠类型的方法,以克服现有技术的不足。

[0004] 为达到上述目的,本发明采用如下技术方案:

[0005] 一种快速判别烟用爆珠类型的方法,包括以下步骤:

[0006] 步骤1)、对不同类型烟用爆珠进行原始光谱采集;

[0007] 步骤2)、对采集的不同类型烟用爆珠原始光谱进行预处理;

[0008] 步骤3)、对预处理后的不同类型烟用爆珠光谱建立分类模型,利用分类模型对待检测烟用爆珠进行识别。

[0009] 首先,在室温下对不同类型烟用爆珠样品进行近红外光谱采集,谱区采集范围: $12000\text{cm}^{-1}\sim 4000\text{cm}^{-1}$,光谱分辨率: $2\sim 32\text{cm}^{-1}$,扫描次数 $8\sim 128$ 次,分不同的天对样品进行光谱数据采集。

[0010] 1) 采用矢量归一化方法,对未知光谱x进行预处理,其矢量归一化算法如下:

$$[0011] \quad X_{\text{normalized}} = \frac{X - \bar{X}}{\sqrt{\sum_{k=1}^m X_k^2}}$$

[0012] 其中, $\bar{X} = \frac{\sum_{k=1}^m X_k}{m}$, $k=1,2,\dots,m$,m为波长点数。

[0013] 2) 采用光谱均值中心化变换进行预处理,将样品光谱减去校正集的平均光谱,经过变换的校正集光谱阵的列平均值为零;

[0014] 计算校正集样品的平均光谱 \bar{x} :

$$[0015] \quad \bar{x}_k = \frac{\sum_{i=1}^n x_{i,k}}{n}$$

[0016] 式中,n为校正集样品数, $k=1,2,\dots,m$,m为波长点数。

[0017] 3) 采用标准正态变量变换法对原始光谱数据进行预处理:

[0018] 标准正态变量变换法的光谱按下式计算:

$$[0019] \quad x_{SNV} = \frac{x - \bar{x}}{\sqrt{\frac{\sum_{i=1}^m (x_k - \bar{x})^2}{m-1}}}$$

[0020] 其中, $\bar{x} = \frac{\sum_{k=1}^m x_k}{m}$, $k=1,2,\dots,m$, m 为波长点数。

[0021] 进一步的,采用PLS-DA有监督模式识别算法或者SIMCA有监督模式识别方法或者线性判别分析,对三类烟用爆珠建立分类模型。

[0022] PLS-DA方法:分解光谱数据矩阵与类别数据矩阵,并使得二者最大程度的相关,采用Kennard-Stone法选择代表性样本,基于变量之间的欧氏距离计算得到的代表性样本集,选择最具差异性的样本用于模型的建立。

[0023] SIMCA方法:对训练集中每一类样本的光谱矩阵分别进行主成分分析,建立每一类的主成分分析数学模型,然后在此基础上对未知样本进行分类。

[0024] 线性判别分析:分别计算类内和类间的协方差矩阵,计算未知样品 x_{un} 与每类均值的距离平方,然后将其判别到距离最小的类别当中即可。

[0025] 与现有技术相比,本发明具有以下有益的技术效果:

[0026] 本发明一种快速判别烟用爆珠类型的方法,通过对不同类型烟用爆珠进行原始光谱采集,建立不同类型烟用爆珠的标准样本,对采集的不同类型烟用爆珠原始光谱进行预处理,以达到消除各种噪声和干扰,降低样品表面不均匀等因素影响,然后对预处理后的不同类型烟用爆珠光谱进行建立分类模型,利用分类模型对待检测烟用爆珠进行识别,能够快速、准确的对烟用爆珠类型进行判别,样品无需复杂前处理、分析速度快、无损、操作简单方便、模型预测准确率高,对烟用爆珠类别快速分析进行初步探索,并可进一步为烟用爆珠质量分析提供技术支撑。

[0027] 进一步的,采用近红外光谱技术不仅适用于产品组分的品质控制和定量分析,还能够利用光谱所反映的组成和结构信息进行定性判别,同时,近红外光谱穿透能力强,在1cm-3cm之间,因此,可以获得爆珠内含物质的光谱信息,简单快捷。

附图说明

[0028] 图1为实施例烟用爆珠未预处理近红外光谱数据图。

[0029] 图2为实施例烟用爆珠预处理Savitzky-Golay一阶导的近红外光谱数据图。

具体实施方式

[0030] 下面结合附图对本发明做进一步详细描述:

[0031] 一种快速判别烟用爆珠类型的方法,包括以下步骤:

[0032] 步骤1)、原始光谱采集

[0033] 在室温下对样品进行近红外光谱采集,谱区采集范围:12000 cm^{-1} ~4000 cm^{-1} ,光谱分辨率:16 cm^{-1} ,每个样品一个点扫描,扫描次数:64次,通过样品杯旋转,一个样品扫描五个点,测试一个样品(不包括重复)为1min,不同类型烟用爆珠放置仅一层;在此上方添加金属

盖用于光的反射,分不同的天对样品进行光谱数据采集。

[0034] 步骤2)、预处理方法

[0035] 近红外光谱信号会受样品的状态、光的散射、杂散光及仪器响应等的干扰,导致近红外光谱的基线漂移,影响重复性,运用合理的光谱预处理方法可提取近红外光谱的特征信息,消除各种噪声和干扰,降低样品表面不均匀等因素影响,因此,对原始光谱进行预处理是非常必要的,具体采用以下预处理方法:

[0036] a. 归一化法:

[0037] 采用矢量归一化方法,对未知光谱 $x(1 \times m)$,其矢量归一化算法如下:

$$[0038] \quad X_{\text{normalized}} = \frac{X - \bar{X}}{\sqrt{\sum_{k=1}^m X_k^2}}$$

[0039] 其中, $\bar{X} = \frac{\sum_{k=1}^m X_k}{m}$, $k=1,2,\dots,m$, m 为波长点数,采用矢量归一化算法校正由微小光程差异引起的光谱变化。

[0040] b. 均值中心化法:

[0041] 采用光谱均值中心化变换(mean centering)将样品光谱减去校正集的平均光谱,经过变换的校正集光谱阵的列平均值为零;

[0042] 首先计算校正集样品的平均光谱 \bar{x} :

$$[0043] \quad \bar{x}_k = \frac{\sum_{i=1}^n x_{i,k}}{n}$$

[0044] 式中, n 为校正集样品数, $k=1,2,\dots,m$, m 为波长点数。

[0045] 对未知光谱 $x(1 \times m)$: $x_{\text{centered}} = x - \bar{x}$

[0046] c. 标准化法:

[0047] 标准化(auto-scaling)又称均值方差化,光谱标准化变换是将均值中心化处理后的光谱再除以校正集光谱阵的标准偏差光谱。

[0048] 首先计算校正集样品的平均光谱 \bar{x} ,然后计算校正集样品的标准偏差光谱 s :

$$[0049] \quad s_k = \sqrt{\frac{\sum_{i=1}^n (x_{i,k} - \bar{x}_k)^2}{n-1}}$$

[0050] 式中, n 为校正集样品数, $k=1,2,\dots,m$, m 为波长点数。

[0051] 对未知光谱 $x(1 \times m)$ 首先进行均值中心化,然后再除以标准偏差光谱 s ,就得到了标准化处理后的光谱:

$$[0052] \quad x_{\text{autoscaled}} = \frac{x - \bar{x}}{s}$$

[0053] 经过标准化处理后的光谱,其列均值为零,方差为1,在对低浓度成分建立模型时,该方法特别适用。

[0054] d. 多元散射校正(MSC)法:

[0055] 多元散射校正(MSC)先分离光谱中的物理光散射信息和化学光吸收信息,然后消

除不同光谱之间的物理散射信息差别,从而使所有样品中的光谱校正信息在同一水平线上。

[0056] MSC算法主要是消除颗粒分布不均及颗粒大小产生的散射影响,在固体漫反射和浆状物透射光谱中应用较广泛。

[0057] e. 标准正态变量变换 (SNV) 法:

[0058] 对需SNV变换的光谱按下式计算:

$$[0059] \quad x_{SNV} = \frac{x - \bar{x}}{\sqrt{\frac{\sum_{k=1}^m (x_k - \bar{x})^2}{m-1}}}$$

[0060] 其中, $\bar{x} = \frac{\sum_{k=1}^m x_k}{m}$, $k=1,2,\dots,m$, m 为波长点数。

[0061] 步骤3)、分类方法

[0062] 采用PLS-DA有监督模式识别算法或者SIMCA有监督模式识别方法或者线性判别分析法,对三类烟用爆珠建立分类模型:

[0063] PLS-DA方法:PLS-DA是建立在偏最小二乘法(PLS)基础上的一种有监督模式识别方法,该算法的基本思路为分解光谱数据矩阵与类别数据矩阵,并使得二者最大程度的相关;代表性样本选择算法:优秀的模型往往需要具备代表性的样本参与建模;本发明采用Kennard-Stone(K-S)法选择代表性样本,该算法基于变量之间的欧氏距离计算得到的代表性样本集,选择最具差异性的样本用于模型的建立。

[0064] SIMCA方法:SIMCA方法又称相似分析法,它是建立在主成分分析(PCA)基础上的一种有监督模式识别方法,该算法的基本思路是对训练集中每一类样本的光谱矩阵分别进行主成分分析,建立每一类的主成分分析数学模型,然后在此基础上对未知样本进行分类。

[0065] 线性判别分析:线性判别分析(LDA)也属于有监督的模式识别方法,分别计算类内和类间的协方差矩阵,对于未知样品 x_{un} ,只需计算它与每类均值的距离平方,然后将其判别到距离最小的那类即可,在本发明中,采用的是马氏距离。

[0066] 具体针对现有产品做以下实验:

[0067] 采用傅立叶变换近红外光谱仪(Bruker,型号:MPA),取蜜甜型、清甜型和薄荷型3类烟用爆珠样品不同批次共27个批次样品,采用The Unscrambler 10.3软件进行预处理和建模,对清甜型、蜜甜型和薄荷型三类烟用爆珠的不同批次27个样品的408条光谱(清甜型6批次90条,蜜甜型14批次211条,薄荷型7批次107条)进行分析;

[0068] 原始光谱数据和进行预处理后的数据,采用隔4选1的方法,选择出每类中20%的样品光谱(清甜型18条,蜜甜型42条,薄荷型21条)用于外部预测,剩余的80%的样品光谱(清甜型72条,蜜甜型169条,薄荷型85条)作为训练集建立模型。

[0069] 原始光谱数据可进行波长范围选择及预处理。图1为所有样品的光谱叠加部分波数区间表现出平头峰以及光谱区间较大,通常波数在 10000cm^{-1} 以上基团吸收很弱,故可剔除平头峰与主要噪声引起的光谱区间。通过SIMCA算法选择建模样本数并使用外部验证集进行模型验证,采用准确度(accuracy, %),误判率(error, %)以及未归类(not-assigned, %)三个指标进行评价。准确度越高,误判率与未归类越小,说明模型效果越好。

[0070] SIMCA分类模型的建立

[0071] 采用SIMCA方法(5个主成分,90%置信度),分别对清甜型、蜜甜型和薄荷型三类烟用爆珠进行建模。

[0072] 表1.原始数据及各个预处理方法后的光谱数据SIMCA建模分类效果表

[0073]

SMICA 模型	预处理		无预处理		归一化			均值中心化		
	总样本数	识别数量	识别率	误判率	识别数量	识别率	误判率	识别数量	识别率	误判率
清甜模型	92	47	51.09%	16.30%	29	31.52%	0	47	51.09%	16.30%
蜜甜模型	212	43	20.28%	0	97	45.75%	0.47%	43	20.28%	0
薄荷模型	112	107	95.54%	0	103	91.96%	0	107	95.54%	0
SMICA 模型	预处理		标准化		MSC			SNV		
	总样本数	识别数量	识别率	误判率	识别数量	识别率	误判率	识别数量	识别率	误判率
清甜模型	92	47	51.09%	16.30%	92	100%	0	92	100%	0
蜜甜模型	212	43	20.28%	0	194	91.51%	0	194	91.51%	0
薄荷模型	112	107	95.54%	0	107	95.54%	0	107	95.54%	0

[0074] 表1为原始数据和经五种预处理方法后的光谱数据的SIMCA建模分类效果,得到如下结论:无处理的建模效果较不好,清甜型、蜜甜型识别率较低且有误判的情况;归一化、均值中心化、标准化处理后,建模效果未得到明显改善,说明这三种预处理方法并不适用于本模型;MSC、SNV的建模分类效果显著提高,识别率均超过90%,三种类型爆珠可以成功被分开,SIMCA所建立起的模型不存在误判现象。

[0075] 2.1.2 SIMCA分类模型的验证

[0076] 用预测集对建立的SIMCA模型进行预测。

[0077] 表2.原始数据及各个预处理方法后的光谱数据SIMCA建模分类预测表

[0078]

SMICA 模型	预处理		无预处理		归一化			均值中心化		
	总样本数	识别数量	识别率	误判率	识别数量	识别率	误判率	识别数量	识别率	误判率
清甜模型	19	10	52.63%	15.79%	5	26.32%	0	10	52.63%	15.79%
蜜甜模型	43	8	18.60%	0	19	44.19%	0	8	18.60%	0
薄荷模型	23	22	95.65%	0	21	91.30%	0	22	95.65%	0
SMICA 模型	预处理		标准化		MSC			SNV		
	总样本数	识别数量	识别率	误判率	识别数量	识别率	误判率	识别数量	识别率	误判率
清甜模型	19	10	52.63%	15.79%	19	100%	0	19	100%	0
蜜甜模型	43	8	18.60%	0	39	90.70%	0	39	90.70%	0
薄荷模型	23	22	95.65%	0	22	95.65%	0	22	95.65%	0

[0079] 表2为预测集对SIMCA模型的预测结果,从表中可以看到,MSC、SNV方法效果很好,对清甜型、蜜甜型的识别率均在90%以上,结果表明所建的SIMCA模型具有较高的预测准确度和预测稳定性。

[0080] 图1,图2分别为烟用爆珠未经处理和预处理S-G的近红外光谱数据图。从图中可以明显看出,S-G预处理可以达到明显的消除基线漂移的作用。

[0081] 2.2线性判别分析模型

[0082] 2.2.1 LDA模型的建立

[0083] 采用LDA方法(PCA与马氏距离,5个主成分),对不同类型的烟用爆珠进行评价。

[0084] 表3.原始数据及各个预处理方法后的光谱数据LDA建模表

[0085] SIMCA 模型	准确度预处理					
	无预处理	归一化	均值中心化	标准化	MSC	SNV
清甜模型	100%	100%	100%	100%	100%	100%
蜜甜模型	100%	100%	100%	100%	99.41%	99.41%
薄荷模型	100%	100%	100%	100%	100%	98.88%

[0086] 表3为原始数据及各个预处理方法后的光谱数据的LDA建模,从表中可以看到,五种预处理方法和未处理的光谱数据的LDA模型分类效果均很好,都基本接近100%,三种类型的各个批次均能被很好的分开。我们需用外部检验的方法,寻求最优的方法。

[0087] 2.2.2 LDA模型的验证

[0088] 用预测集对建立的各个LDA模型进行外部预测。

[0089] 表4. 外部预测正确率

[0090] SIMCA 模型	正确率 预处理					
	无预处理	归一化	均值中心化	标准化	MSC	SNV
清甜模型	94.74%	94.74%	94.74%	94.74%	100%	100%
蜜甜模型	100%	97.67%	100%	100%	97.67%	97.67%
薄荷模型	100%	100%	100%	100%	100%	98.88%

[0091] 表4为外部预测正确率结果表,我们可以看到,所有的LDA模型的预测正确率均很好,其中MSC和SNV较其他结果稍好,对三种类型不同批次的烟用爆珠均能很好的分开。

[0092] 采用近红外光谱分析结合SIMCA模式识别技术对3种烟用爆珠的不同批次共27个样品建立了分类模型,可将3种烟用爆珠成功分开,同时进行了预测,蜜甜型、清甜型和薄荷型烟用爆珠不同批次样品的准确度分别为97.67%、100%和100%。该方法操作简单、快速及无损,对烟用爆珠类别快速分析进行初步探索,并可进一步为烟用爆珠质量分析提供技术支持。

[0093] 实施案例1

[0094] 采集三类爆珠的近红外光谱图,分不同天测试,一个批次样品重复测试3次。波数设置范围 $10000\text{cm}^{-1}\sim 4520\text{cm}^{-1}$,数据优先不考虑预处理。通过K-S选择不同校正集的样本数所对应的结果如下表5所示。可以发现,通过K-S选择不同的校正集样本,当选择样本数达到60个或者70个,对于验证集样本的预测准确率为100%,说明该模型存在使用价值能力。但是当选择样本数为50个时,准确率下降,说明建模时光谱数据存在干扰因素。

[0095] 表5无预处理的校正集样本个数与准确率的关系

[0096]

	15 个样本	20 个样本	30 个样本	40 个样本	50 个样本	60 个样本	70 个样本
准确度%	95.1	97.55	98.78	100	98.78	100	100
误判率%	4.9	2.45	1.22	0	1.22	0	0
未归类%	0	0	0	0	0	0	0

[0097] 实施案例2

[0098] 采集三类爆珠的近红外光谱图,分不同天测试,一个批次样品重复测试3次。波数设置范围 $10000\text{cm}^{-1}\sim 4520\text{cm}^{-1}$ 。光谱数据由于仪器等原因,数据中普遍存在噪声或者干扰成分的现象,结果如下表6所示,数据可预处理以提升模型的稳健性。本例中选择的预处理方法为Savitzky-Golay一阶导(9点法)。当选择样本数为30时,准确度就达到100%,说明有效地扣除了光谱数据的噪声,提高校正模型的稳健性。

[0099] 表6预处理的校正集样本个数与准确率的关系

[0100]

	15 个样本	20 个样本	30 个样本	40 个样本	50 个样本	60 个样本	70 个样本
准确度%	96.81	99.26	100	100	100	100	100
误判率%	3.19	0.49	0	0	0	0	0
未归类%	0	0.25	0	0	0	0	0

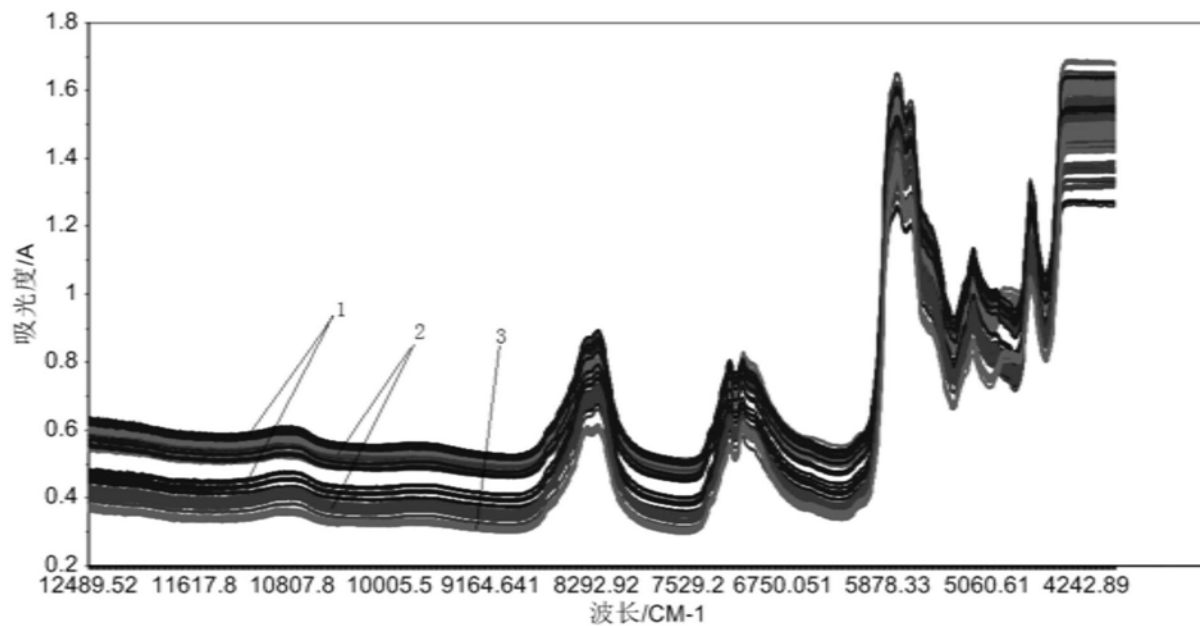


图1

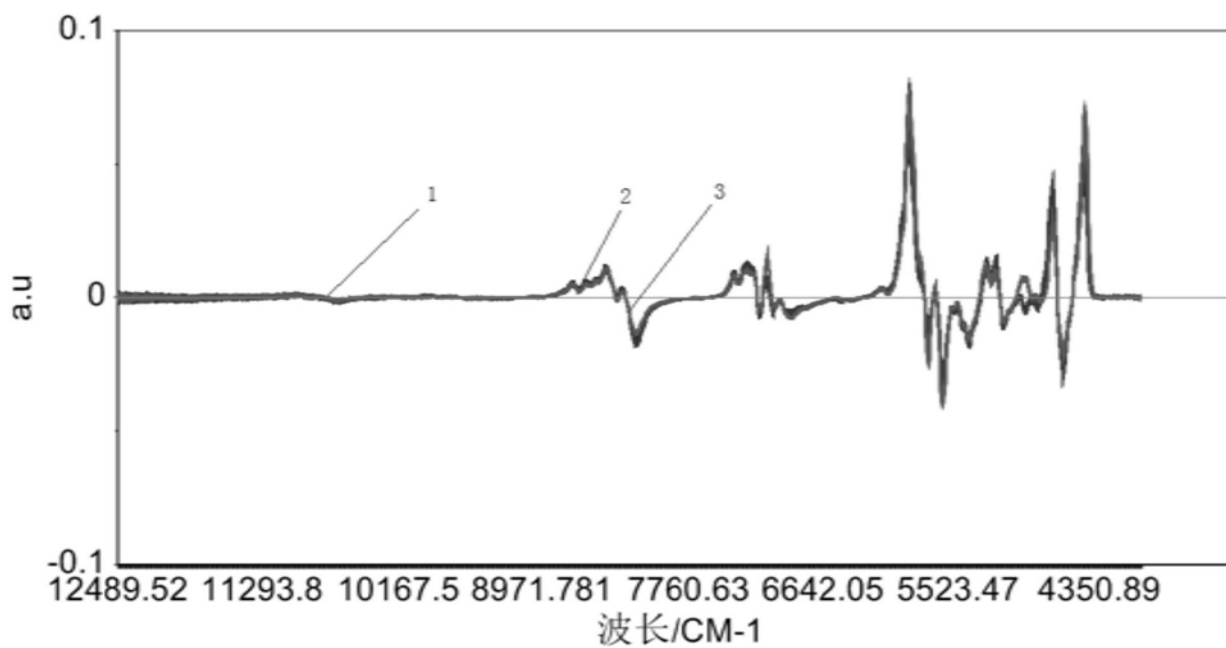


图2