XGBoost 在气体红外光谱识别中的应用

陶孟琪[1,2]，刘家祥[1]，吴越[1,2]，宁志强[1,2]，方勇华[1,2*]

[1]中国科学院安徽光学精密机械研究所环境光学与技术重点实验室，安徽 合肥 230031；

[2]中国科学技术大学，安徽 合肥 230026

摘要 为解决气体红外光谱识别问题，引入提升算法中较新的研究成果——极端梯度提升(XGBoost)算法。选用实测的三氯甲烷、对二甲苯、四氯乙烯气体的红外光谱数据进行实验。首先在对原始数据进行预处理后，通过特征工程提取光谱特征，生成特征向量；然后建立 XGBoost 模型，并对模型参数进行调优；最后基于分类准确率指标，将所提模型与随机森林(RF)、支持向量机(SVM)、前馈神经网络(FNN)、卷积神经网络(CNN)模型进行对比。实验结果表明，XGBoost 在气体红外光谱识别领域有着广阔的应用前景。

关键词 光谱学；模式识别；红外光谱；提升算法；特征工程

中图分类号 O433 **文献标志码** A **doi**：10.3788/AOS202040.0730002

Application of XGBoost in Gas Infrared Spectral Recognition

Tao Mengqi[1,2], Liu Jiaxiang[1], Wu Yue[1,2], Ning Zhiqiang[1,2], Fang Yonghua[1,2*]

[1]*Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, Anhui* 230031, *China*;

[2]*University of Science and Technology of China, Hefei, Anhui* 230026, *China*

Abstract To address the problem of gas infrared spectral identification, a new lifting algorithm named eXtreme gradient boosting (XGBoost) is introduced. Infrared spectral data of chloroform, p-xylene, and tetrachloroethylene are selected for experiments. After these original data are preprocessed, the spectral features are first extracted by feature engineering to generate feature vectors. Then, the XGBoost model is established and its parameters are optimized. Finally, based on a classification accuracy index, the XGBoost model is compared with random forest (RF), support vector machine (SVM), feedforward neural network (FNN), and convolutional neural network (CNN). The experimental results show that XGBoost has a broad application prospect in the field of gas infrared spectral identification.

Key words spectroscopy; pattern recognition; infrared spectroscopy; lifting algorithm; feature engineering

OCIS codes 300.6340; 070.5010; 150.1135

1 引 言

气体种类识别在科学研究、生产及环境检测等领域有着广泛的应用。通过傅里叶变换红外(FTIR)光谱仪采集气体的红外光谱数据，再对这些光谱数据进行特征提取并分析，从而识别出气体种类[1-3]。这种方式具有检测重复性好、分析速度快、不需要消耗样本、可在线实时监测等特点，在气体识别领域有着广泛的应用。

当前，随着计算机科学技术的快速发展，人工智能算法在各种领域中得到了广泛的应用，尤其在分类识别领域取得了巨大的研究成果，因此红外光谱识别与人工智能算法相结合的方式在气体红外光谱识别领域逐渐成为了研究热点。白鹏等[4]提出了基于支持向量机(SVM)二值分类识别的逐一气体种类识别法，该方法在对天然气气体种类识别实验中取得了良好的效果。刘美娟等[5]将小波多尺度分析与反向传播神经网络算法结合，实现了对甲基膦酸二甲酯(DMMP)光谱的快速识别。余段辉[6]采用径向基神经网络算法，对 NH_3、HCL、CO、NO、CH_4 这 5 种气体光谱进行了有效识别。2016 年，Chen 等[7]提出了极端梯度提升(XGBoost)算法，该算法由于具有快速、高效、准确等特点，在医疗、生物、环境检测、商业生产、军事等领域有着广泛的应用，并

收稿日期：2019-12-03；**修回日期**：2019-12-19；**录用日期**：2019-12-30

* **E-mail**：yhfang@aiofm.ac.cn

取得了优异的效果。因此，将 XGBoost 算法应用于红外光谱识别领域有着重要的研究意义。本文将 XGBoost 算法应用于气体红外光谱种类识别中，从气体红外光谱数据预处理、特征提取、建模分析这三个方面开展工作，并挑选了实测的三氯甲烷、对二甲苯、四氯乙烯气体的红外光谱数据对 XGBoost 算法进行验证。实验结果表明，XGBoost 算法具有良好的工作特性和模式识别能力，可完成气体种类识别任务，具有重要的工程应用价值。

2　基本原理

2.1　XGBoost 算法

XGBoost 算法是提升算法的一种，通过残差拟合生成多个弱学习器，最后将生成的弱分类器累加起来得到一个强学习器[8-12]。XGBoost 在优化过程中将损失函数二阶泰勒展开，引入二阶导数信息，使得模型在训练过程中的收敛速度更快。此外，XGBoost 还在损失函数中添加正则化项来抑制模型复杂度，防止过拟合现象的发生。XGBoost 算法的具体推导过程如下。令 $D=\{(x_i,y_i)\}$ 是一个拥有 n 个样本、每一个样本拥有 d 个特征的数据集，其中 x_i 表示第 i 个样本数据，y_i 表示第 i 个样本的标签。基模型选用分类与回归树（CART）。XGBoost 的集成模型利用由 K（树的数目）个基模型组成的一个加法运算式来预测最终结果：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i),\tag{1}$$

式中：k 为树的编号；$f_k(\cdot)$ 为第 k 棵树的表达式。

模型的预测精度是由模型的偏差和方差共同决定的。损失函数可以反映模型的偏差。为了控制模型的方差，使模型更简单，添加抑制模型复杂度的正则化项。正则化项与模型的损失函数构成 XGBoost 算法的目标函数：

$$\begin{cases}\mathrm{Obj}^{(k)}=\sum_{i=1}^{n}l\left[y_i,\hat{y}_i^{(k-1)}+f_t(x_i)\right]+\sum_{k}\Omega(f_k)\\ \Omega(f)=\gamma T+\frac{1}{2}\lambda\left\|\boldsymbol{w}\right\|^2\end{cases},\tag{2}$$

式中：$\hat{y}_i^{(k-1)}$ 为前 $(k-1)$ 棵树的输出值之和；$f_k(x_i)$ 为第 k 棵树的输出结果；l 为衡量预测值 $\hat{y}_i$ 和真实值 y_i 之间差异的可微凸损失函数；$\Omega(\cdot)$ 为模型复杂度的惩罚项；γ 为叶子数目的正则化参数；λ 为叶子权重的正则化参数；$\boldsymbol{w}$ 为叶子节点的取值；T 为叶子节点的个数。定义 I_j 为第 j 个叶子节点上的样本集合，将损失函数在 $\hat{y}_i^{(k-1)}$ 处利用泰勒公式展开。定义 g_i 和 h_i 为泰勒展开式的一阶导数和二阶导数，去掉常数项，则泰勒展开后的目标函数变为

$$\mathrm{Obj}^{(k)}=\sum_{j=1}^{T}\left[\left(\sum\nolimits_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum\nolimits_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T,\tag{3}$$

式中：w_j 为叶子节点 j 的权重。定义 $G_i=\sum_{i\in I_j}g_i$，$H_i=\sum_{i\in I_j}h_i$，代入(3)式，目标函数化简为

$$\mathrm{Obj}^{(k)}=\sum_{j=1}^{T}\left[G_iw_j+\frac{1}{2}(H_i+\lambda)w_j^2\right]+\gamma T。\tag{4}$$

在(4)式中，叶子节点 w_j 是一个不确定值。故将目标函数 $\mathrm{Obj}^{(k)}$ 对 w_j 求一阶导数，解出叶子节点 j 的最优值 w_j^*：

$$w_j^*=-\frac{G_i}{H_i+\lambda}。\tag{5}$$

将 w_j^* 代回目标函数，$\mathrm{Obj}^{(k)}$ 取得最小值：

$$\mathrm{Obj}^{(k)}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T。\tag{6}$$

2.2　XGBoost 模型构建流程

XGBoost 在构建 CART 时，使用贪心算法进行特征分裂，设树的最大深度为 m，算法流程如图 1 所示。

Algorithm: exact greedy algorithm for split finding
Input: I, instance set of current node
Input: s, feature dimension
gain $\leftarrow 0$
$G\leftarrow\sum_{i\in I}g_i, H\leftarrow\sum_{i\in I}h_i$
for $k=1$ to m **do**
  $G_{\mathrm{L}}\leftarrow 0, H_{\mathrm{L}}\leftarrow 0$
  for j in sorted (I, by x_{jk}) **do**
    $G_{\mathrm{L}}\leftarrow G_{\mathrm{L}}+g_j, H_{\mathrm{L}}\leftarrow H_{\mathrm{L}}+h_j$
    $G_{\mathrm{R}}\leftarrow G-G_{\mathrm{L}}, H_{\mathrm{R}}\leftarrow H-H_{\mathrm{L}}$
    $\mathrm{score}\leftarrow\max\left(\mathrm{score},\frac{G_{\mathrm{L}}^2}{H_{\mathrm{L}}+\lambda}+\frac{G_{\mathrm{R}}^2}{H_{\mathrm{R}}+\lambda}-\frac{G^2}{H+\lambda}\right)$
  end
end
Output: split with max score

图 1　XGBoost 算法流程

Fig. 1　Flow chart of XGBoost algorithm

贪心算法每次从根节点开始对每一节点遍历所有特征，选出得分最高的点作为分裂节点，分裂到树的最大深度后停止分裂，并开始构建下一棵树的残差。最后将所有生成的树进行集合，得到 XGBoost 模型。XGBoost 算法构建示意图如图 2 所示[13]。

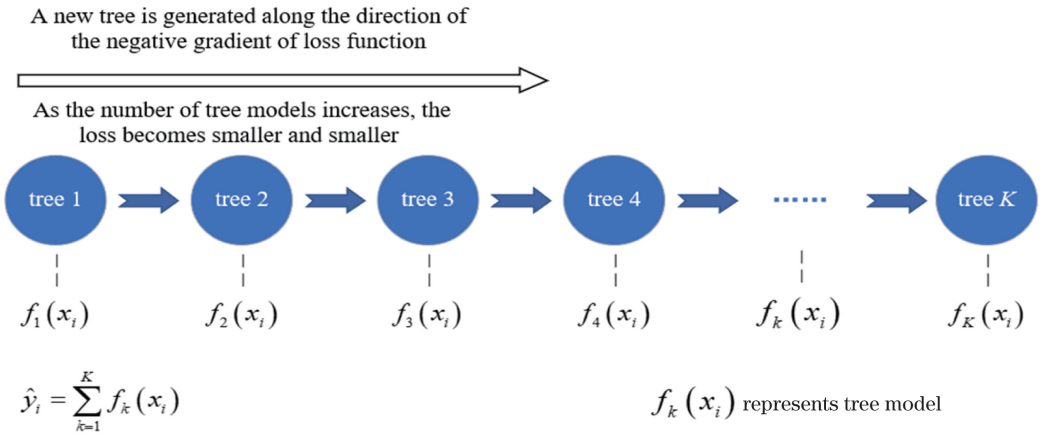

图 2　XGBoost 算法示意图[13]

Fig. 2　Schematic of XGBoost algorithm[13]

3　实验测试与分析

3.1　数据处理

本实验数据集选取实际测量的三氯甲烷、对二甲苯、四氯乙烯这三种气体的红外光谱数据。光谱采集的波数范围是 650～1350 cm^{-1}，光谱分辨率为 2.35 cm^{-1}。共收集了 8616 条数据，其中三氯甲烷 3025 条、对二甲苯 2855 条、四氯乙烯 2736 条，将数据按照 7∶3的比例随机划分训练集和测试集。

由于数据在采集过程中不可避免地受到仪器响应、背景变换、噪声等干扰，FTIR 光谱仪采集到的数据含有基线和高频噪声成分，这些成分会在不同程度上影响模型的性能。为了消除这些影响，需要对采集的气体红外光谱数据进行预处理。第一步，选用迭代多项式拟合法，消除原始数据中的基线漂移现象[14]；第二步，将去除基线的数据通过 Savitzky-Golay 滤波器，消除光谱数据中的高频噪声，使光谱曲线更加平滑[15]；第三步，为使所有的气体光谱数据处于同一量级上，并便于模型分析，对每一条光谱数据进行归一化处理。三种气体红外光谱数据预处理前后对比结果如图 3 所示。

在对原始数据进行去基线、滤波、归一化处理后，需要提取特征向量来训练 XGBoost 模型。通过查阅相关文献[16-18]并结合三种气体红外光谱数据，选取了几种能够很好地反映不同种气体吸收峰信息的特征。选取的特征和意义如表 1 所示。

3.2　XGBoost 模型训练

在对原始光谱数据进行处理得到特征向量后，开始对 XGBoost 模型进行训练。模型训练步骤如下。

表 1　用于气体光谱分类的特征

Table 1　Features for gas spectral data classification

Feature	Meaning
Width	Full width at half maximum of characteristic peak
Kurtosis	Sharpness of characteristic peak
Skewness	Symmetry of characteristic peak
Correlation	Correlation coefficient with standard spectrum on NIST
SNR	Signal to noise ratio of characteristic peak

1）将得到的特征向量和标签输入到模型中；

2）根据当前已学习到的基学习器的预测值之和与样本真实值得到残差，初始预测值为 0；

3）初始化待分割的特征列表，对于潜在的分割点，得出分割前和分割后的目标函数变化情况；

4）判断当前基学习器的深度是否达到最大分裂深度，若未达到最大深度，则寻找最优分割点，并基于最优分割点分配样本到左右叶子节点，然后回到步骤 3），若达到最大分裂深度，则停止分裂，计算每个叶子节点权重，完成当前基学习器的建立；

5）判断当前模型训练是否达到终止条件（基学习器数量到达设定的最大值等），若未达到，则回到步骤 2），若达到，则将所有训练得到的基学习器集合起来，得到 XGBoost 模型，模型训练结束。

XGBoost 模型训练流程如图 4 所示。

在模型的训练过程中，为了确定模型参数，利用网格搜索结合十折交叉验证法对模型的超参数进行了调优，以确保得到最优的模型架构。

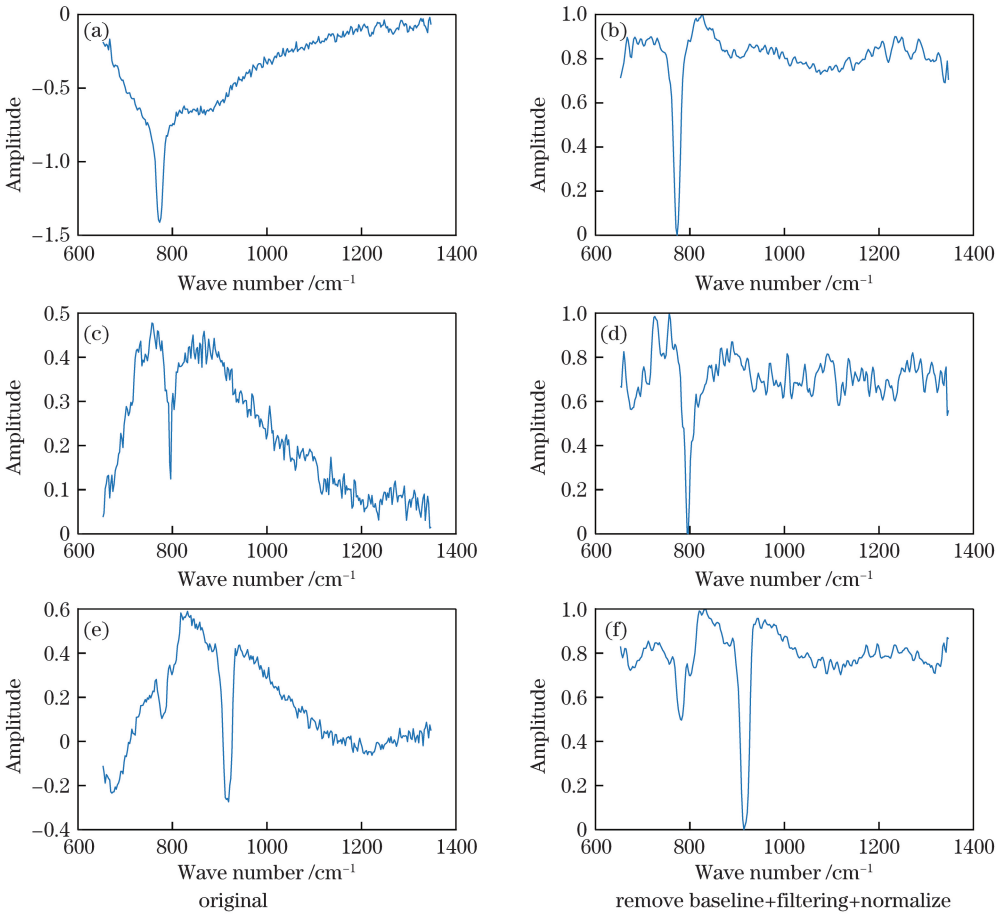

图 3　光谱预处理前后对比。(a)(b)三氯甲烷;(c)(d)对二甲苯;(e)(f)四氯乙烯

Fig. 3　Comparison before and after spectral pretreatment. (a)(b) Trichloromethane; (c)(d) paraxylene; (e)(f) tetrachloroethylene

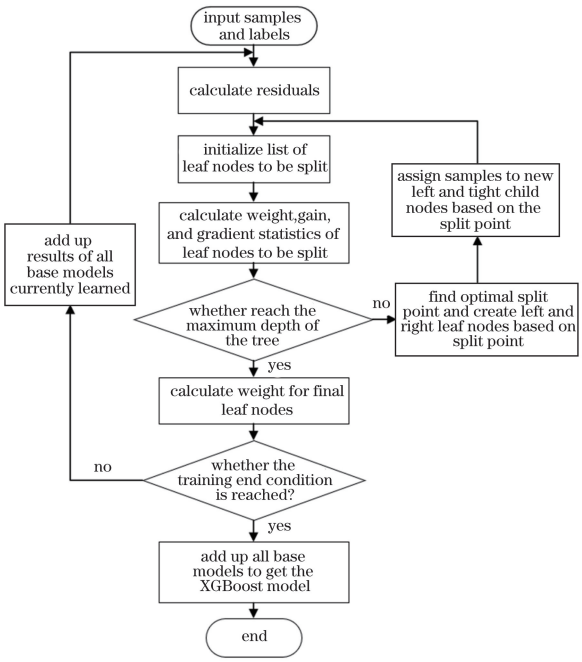

图 4　XGBoost 模型训练流程图

Fig. 4　Flow chart of XGBoost model training

3.3　模型性能分析

在完成 XGBoost 模型训练后，得到三种气体光谱分类误差矩阵，如表 2 所示。误差矩阵的每一列代表数据的预测类别，每一行代表数据的真实类别。从表 2 可以看出，模型对三氯甲烷、对二甲苯及四氯乙烯的分类准确率分别为 97.79%、95.09% 及 97.32%。三种气体的综合分类准确率为 96.75%。

表 2　三种气体分类误差矩阵

Table 2　Classification error matrix for three kinds of gases

Gas name	Trichloromethane	Paraxylene	Tetrachloroethylene
Trichloromethane	887	15	5
Paraxylene	31	814	11
Tetrachloroethylene	9	13	799

为了更好地评估 XGBoost 模型在气体红外光谱分类上的性能，引入随机森林(RF)、SVM、卷积神经网络(CNN)、前馈神经网络(FNN)这 4 种常用的分类算法作为对比。使用上述训练集和测试集对这几种模型进行训练和测试，几种模型的性能如

表 3 所示。

表 3　5 种模型的分类准确率

Table 3　Classification accuracy for five models

Model	Accuracy /%
RF	96.35
SVM	79.48
CNN	80.37
FNN	95.61
XGBoost	96.75

从表 3 可以看出：SVM 模型和 CNN 模型的分类准确率在 80% 左右，表现较差；RF、FNN、XGBoost 这 3 种模型的分类准确率表现较好，均达到了 95%以上；其中 XGBoost 模型的分类准确率最高，说明 XGBoost 模型在气体光谱识别上有着良好的性能表现。

4　结　论

通过实际测得的三种气体红外光谱数据，研究了基于 XGBoost 算法的气体红外光谱数据分类问题。首先对原始光谱数据进行预处理，得到了用于模型训练的特征向量；然后通过网格搜索和交叉验证等方法对模型进行优化；最后基于分类准确率指标，将 XGBoost 与随机森林、支持向量机、卷积神经网络、前馈神经网络这 4 种常用的分类模型进行对比。实验结果表明，XGBoost 算法的分类正确率达到了 96.75%，与其他模型性能相比有不同程度的提升。因此，XGBoost 模型在红外光谱识别上性能表现优越，具有广阔的应用前景。

参　考　文　献

[1] Sheehe S L, Jackson S I. Identification of species from visible and near-infrared spectral emission of a nitromethane-air diffusion flame[J]. Journal of Molecular Spectroscopy, 2019, 364: 111185.

[2] Han Y Z, Zhang Y X, Chang S J, et al. Recognition for the nonlinear fluorescence spectra based on optimal wavelet transform and artificial neural network[J]. Journal of Optoelectronics · Laser, 2005, 16(6): 718-721.
韩应哲, 张延炘, 常胜江, 等. 基于最佳小波变换和神经网络的气体非线性荧光光谱的识别[J]. 光电子·激光, 2005, 16(6): 718-721.

[3] Bai P, Xie W J, Liu J H. Method of infrared spectrum analysis of hydrocarbon mixed gas based on multilevel and SVM-subset[J]. Spectroscopy and Spectral Analysis, 2008, 28(2): 299-302.
白鹏, 谢文俊, 刘君华. 层次式 SVM 子集含烃类混合气体光谱分析方法[J]. 光谱学与光谱分析, 2008, 28(2): 299-302.

[4] Bai P, Wang J H, Wang H K, et al. A method of mixed gas component infrared spectrum recognition based on SVM regression model[J]. Acta Photonica Sinica, 2008, 37(4): 754-757.
白鹏, 王建华, 王宏柯, 等. 基于 SVM 回归模型的混合气体组分种类光谱识别方法[J]. 光子学报, 2008, 37(4): 754-757.

[5] Liu M J, Feng W W, Shi F R, et al. Fast algorithm for feature extraction and identification of infrared spectra of polluted gases[J]. Spectroscopy and Spectral Analysis, 2006, 26(10): 1854-1857.
刘美娟, 冯巍巍, 史丰荣, 等. 污染气体红外光谱特征的快速提取与识别[J]. 光谱学与光谱分析, 2006, 26(10): 1854-1857.

[6] Yu D H. Research on gas recognition and concentration detection algorithm based on infrared spectrum[D]. Chengdu: University of Electronic Science and Technology of China, 2018: 13-58.
余段辉. 基于红外光谱的气体识别与浓度检测算法研究[D]. 成都: 电子科技大学, 2018: 13-58.

[7] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, California. New York: ACM, 2016: 785-794.

[8] Zopluoglu C. Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost)[J]. Educational and Psychological Measurement, 2019, 79(5): 931-961.

[9] Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine learning: XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017, 4: 159-169.

[10] Li D Z, Wang C, Li Y Y. Evaluation of fan blade icing based on XGBoost algorithm[J]. Electric Power Science and Engineering, 2019, 35(9): 43-48.
李大中, 王超, 李颖宇. 基于 XGBoost 算法的风机叶片结冰状态评测[J]. 电力科学与工程, 2019, 35(9): 43-48.

[11] Zhang X, Luo A. XGBOOST based stellar spectral classification and quantized feature[J]. Spectroscopy and Spectral Analysis, 2019, 39(10): 3292-3296.
张枭, 罗阿理. 基于 XGBOOST 的恒星光谱分类特征数值化[J]. 光谱学与光谱分析, 2019, 39(10): 3292-3296.

[12] Zhang W W, Liu D, Jia X Y. Three classified coupon prediction based on XGBoost algorithm[J]. Journal of

Nanjing University of Aeronautics & Astronautics, 2019, 51(5): 643-651.
张微薇, 刘盾, 贾修一. 基于XGBoost的三分类优惠券预测方法[J]. 南京航空航天大学学报, 2019, 51(5): 643-651.

[13] Mo H, Sun H J, Liu J J, et al. Developing window behavior models for residential buildings using XGBoost algorithm[J]. Energy and Buildings, 2019, 205: 109564.

[14] Wang X, Lu S L, Li Y, et al. Automatic baseline correction of gas spectra based on baseline drift model [J]. Spectroscopy and Spectral Analysis, 2018, 38(12): 300-305.
王昕, 吕世龙, 李岩, 等. 基于基线漂移模型的气体光谱自动基线校正[J]. 光谱学与光谱分析, 2018, 38(12): 300-305.

[14] Wang X, Lü S L, Li Y, et al. Automatic baseline correction of gas spectra based on baseline drift model [J]. Spectroscopy and Spectral Analysis, 2018, 38(12): 3946-3951.
王昕, 吕世龙, 李岩, 等. 基于基线漂移模型的气体光谱自动基线校正[J]. 光谱学与光谱分析, 2018, 38(12): 3946-3951.

[15] Liu J, Koenig J L. A new baseline correction algorithm using objective criteria[J]. Applied Spectroscopy, 1987, 41(3): 447-449.

[16] Zhao Y S, Xue X M, Song X J, et al. Comparison and analysis of FT-IR spectra for six broad-leaved wood species[J]. Journal of Forestry Engineering, 2019, 33(5): 40-45.
赵阅书, 薛晓明, 宋小娇, 等. 6种阔叶树材红外光谱特征的比较[J]. 林业工程学报, 2019, 33(5): 40-45.

[17] Yang S Q, Yan L J, Liu N, et al. Asphalt index based on characteristic spectral analysis of infrared spectrum[J]. Journal of Jiangsu University(Natural Science Edition), 2019, 40(2): 244-248.
杨三强, 颜立景, 刘娜, 等. 基于红外光谱图特征峰分析的沥青指标[J]. 江苏大学学报(自然科学版), 2019, 40(2): 244-248.

[18] Zhuang L, Song X J, Xu Y H. Study on the infrared spectral characteristic of tetracentron sinense wood [J]. Hubei Agricultural Sciences, 2017, 56(7): 1334-1339, 1344.
庄琳, 宋小娇, 徐燕红. 水青树木材红外光谱特征研究[J]. 湖北农业科学, 2017, 56(7): 1334-1339, 1344.