一种结合直接正交信号校正与蒙特卡罗的波长选择方法

谢林江,洪明坚*,余志荣

重庆大学大数据与软件学院,重庆 401331

摘 要 在近红外光谱数据分析中,全光谱数据具有波长点多、冗余量大、共线性关系严重的特点,导致了 部分波长点对建立校正模型没有积极作用,甚至还会降低模型的预测能力。波长选择被证明是有效避免上 述问题的重要方法。针对近红外光谱的特性,提出了一种基于直接正交信号校正(DOSC)与蒙特卡罗方法 (Monte Carlo, MC)结合的波长选择算法。与大多数根据波长的"重要性"进行选择的方法不同, MC-DOSC 依据波长的"不重要"性进行选择。波长"不重要"性通过 DOSC 的权重 w 来度量。首先将 w 归一化作为波长 被滤除的概率,以此建立波长选择的概率模型,并使用蒙特卡罗随机抽样得到 N 个波长子集的集合。在每 一次抽样过程中,用选择的波长点建立 PLS 模型,计算相应的交叉验证均方根误差(RMSECV)。经过 N 次 随机抽样后,以 RMSECV 最小时的 PLS 模型对应的波长子集作为备选子集。将备选子集包含的光谱数据作 为新的光谱阵, 重复上述过程直到 RMSECV 不再下降为止。迭代停止后, 将 RMSECV 最小的备选子集作 为最佳波长子集。采用玉米数据集和汽油数据集对该算法进行测试,同时与蒙特卡罗无信息变量消除法 (MCUVE)、遗传算法(GA)、竞争性自适应权重取样法(CARS)三种算法进行比较。实验结果表明:该算法 能大幅度减少波长点个数,并且相应的 PLS 模型的预测能力也提高了。玉米数据集的实验运行结果,波长 点个数从全光谱的 700 个减少到 15 个,预测集相关系数从 0.828 2 提高到 0.931 4, RMSEP 从 0.109 8 减少 到 0.071 3。汽油数据集的实验运行结果,波长点个数从全光谱的 301 个减少到 31 个,预测集相关系数从 0.987 5 提高到 0.993 9, RMSEP 从 0.255 5 减少到 0.178 8。该算法在 2 个数据集中的表现均优于对比的三 种算法。

关键词 近红外光谱;波长选择;正交信号校正;蒙特卡罗;偏最小二乘

中图分类号: O657.33 文献标识码: A DOI: 10.3964/j. issn. 1000-0593(2022)02-0440-06

引言

近红外光谱分析技术因为其分析速度快、无损、无污染和成本低等优点,在农业、制药、石化、环境和医学等领域有着广泛的应用^[1]。近红外光谱也有波长点多、谱峰重叠、吸光强度低、波长点之间共线性严重等特点。采用全光谱数据进行分析和建模,不但会增加模型的运算时间,也会增加模型的复杂度,其中部分波长点还会降低模型的预测质量^[2]。选择出有效的波长点参与建模有着重要的实际意义。

为了有效选出合适的波长点,基于不同原理的波长选择算法先后被提出。主要有蒙特卡罗无信息变量消除法(Monte Carlo uninformative variable elimination, MCUVE)^[3]、遗传算法(genetic algorithms, GA)^[4]、连续投影算法(successive

projections algorithms, SPA)^[5]、协同间隔偏最小二乘法 (synergy interval PLS, SiPLS)^[6]、向后间隔偏最小二乘法 (backward interval partial least square, B-iPLS)^[7]、变量迭代空间收缩算法 (variable iterative space shrinkage approach, VISSA)^[8]、竞争性自适应权重取样法 (competitive adaptive reweighted sampling, CARS)^[9]、随机蛙跳 (random frog, RF)^[10]和最小绝对收缩和选择算法 (least absolute shrinkage and selection operator, LASSO)等。基于 PLS 回归系数为波长重要性指标进行波长筛选是一类常用方法^[11-18],如 MCUVE,VISSA,CARS等。但是,基于回归系数"重要性"进行选择的方法存在两个问题:①由于光谱容易受到噪声和仪器测量误差的影响,回归系数并非总是体现波长"重要性"的真实信息。②利用回归系数选择波长有着主因子个数选取的困难,不同的主因子个数对应的回归系数往往有着较大的

收稿日期: 2021-01-07, 修订日期: 2021-04-11

基金项目: 国家重点研发计划项目(2018YFF01011204)资助

作者简介: 谢林江, 1993 年生, 重庆大学大数据与软件学院硕士研究生 e-mail: 584283862@qq. com

不稳定性。

针对上述问题,提出一种通过滤除"不重要"波长进行选择的新方法。首先用直接正交信号校正(direct orthogonal signal correction, DOSC)提取光谱中与物质浓度阵无关的信息,将其权重向量绝对值归一化后作为波长被筛除的概率,然后用 Monte Carlo 方法进行迭代,逐步剔除不重要的波长。通过玉米和汽油数据集对 MC-DOSC 进行了验证,并与MCUVE,GA和 CARS 三种算法进行了对比,验证 MC-DOSC 的有效性。

1 实验部分

1.1 数据来源

1.1.1 玉米数据集

玉米数据集包含玉米近红外光谱及其植物油(oil)的含量,此数据集是80个玉米样本在m5,mp5和mp6三台不同光谱仪采集得到。实验采用的光谱由mp5光谱仪采样得到。波长采集范围为1100~2498nm,采集间隔为2nm,共700个波长点。使用duplex方法将样本分成60个训练集样本和20个验证集样本。数据集可从http://www.eigenvector.com/data下载。

1.1.2 汽油数据集

汽油(gasoline)数据集 [14] 包含汽油的近红外光谱及其辛烷值(octane)。此数据集含有 60 个样本,波长采集范围为 $900\sim1~700~nm$,采集间隔为 2~nm,由于其前 100~个波长 ($900\sim1~100~nm$)几乎不携带有效信息,故提前将其删除。使用 duplex 方法将样本分成 45~个训练集样本和 15~个验证集样本。

1.2 正交信号校正

正交信号校正方法(OSC)通常用于光谱预处理,核心原理是将光谱阵 X 和浓度阵 Y 进行正交,扣除光谱中与浓度无关的信息。当光谱和浓度的相关性不大或者光谱中背景噪声太大时,使用 OSC 方法可以有效减少 PLS 模型的主因子个数,降低模型的复杂度,增强模型的稳健性以及预测能力。

根据扣除方式的不同,衍生了多种 OSC 算法,包括 Wold-OSC、DO (direct orthogonal)、Fearn-OSC、O-PLS (orthogonal projects to latent structure)、DOSC 等。大部分 OSC 算法,需要选择主因子个数,然后通过迭代逐步扣除。而 DOSC 算法可以只用一个主因子扣除光谱中与浓度无关的信息,避免了主因子个数选择的问题^[15]。因此,采用 DOSC 提取波长不重要的度量信息。

DOSC 首先提取浓度阵 Y 在光谱阵 X 所张开的线性空间的投影 M,如图 1 所示

$$\mathbf{M} = \mathbf{X}^{T} ((\mathbf{X}^{T})^{-1})^{T} \mathbf{Y}$$
 (1)

计算 X 在 M 的正交补空间的投影 Z , 见式(2)

$$\mathbf{Z} = \mathbf{X} - \mathbf{M}\mathbf{M}^{-1}\mathbf{X} \tag{2}$$

再对 ZZ^{T} 进行主成分分析,提取前 k 个需正交处理的主成分得分矩阵 T,再计算权重矩阵 W,见式(3)

$$\mathbf{W} = \mathbf{X}^{-1}\mathbf{T} \tag{3}$$

得到权重矩阵 W。由 DOSC 只需一个主因子可知,仅需取 W 中第一个权重向量 w 来度量波长的"不重要"性,与 PLS 利用回归系数 β 度量波长的重要性有互补关系。将 DOSC 权重向量 w 与 PLS 回归系数 β 作对比,如图 1 所示。可以看出,DOSC 的权重向量和 PLS 的回归系数具有明显的 互补关系。

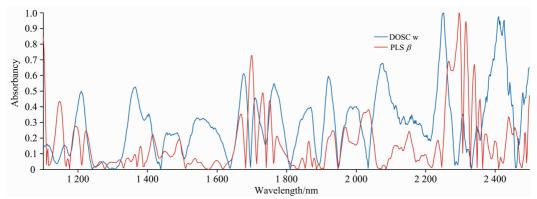


图 1 DOSC 权重向量和 PLS 回归系数对比,其中w和 β 来源于玉米数据集

Fig. 1 The comparison of PLS regression coefficients and DOSC weights, where w and β come from the corn data set

1.3 MC-DOSC 波长选择算法

将 DOSC 的权重向量 w 作为度量波长"不重要"性的依据,并以此选择波长。由于近红外光谱受到吸光度低、谱峰重叠和噪声等因素的影响^[16],严格根据 DOSC 权重向量 w 的绝对值大小进行波长筛选一般不会得到最佳波长点子集。针对这个问题,将权重向量 w 绝对值归一化后作为波长点被滤除的概率,绝对值越小,对应的波长点被保留的概率越

大。选择正态随机分布函数进行大量的蒙特卡罗模拟,得到RMSECV最小时对应的波长点作为备选子集,再对以上过程进行迭代,得到一系列备选子集,然后从中选出RMSECV最小的作为最佳子集。MC-DOSC波长选择算法的主要步骤如下:

Step1:将光谱阵 X 与浓度阵 Y 进行 DOSC,得到权重向量 w;

Step2: 对权重向量 w 求绝对值并归一化得到相应的概率 p_i ;

Step3: 初始化 k=1,生成正态分布随机向量 $R \sim N(0$,1),将 R 求绝对值并归一化得到 r_i ;

Step4: 滤除波长点 $\{j \mid p_j > r_j, j=1, 2\cdots, n\}$, 其余波长点用 PLS 建模, 使用十折交叉验证(10-fold cross-validation) 计算 RMSECV, 并确定主因子数;

Step5: 当 k < N 时,k = k + 1,回到 Step3,当 k = N 时,进入 Step6:

Step6: 在 $N \uparrow$ RMSECV 中选择最小的 RMSECV 所对应的波长点作为备选子集;

Step7:将备选子集作为新的光谱阵 X_{new} 代回 Step1 进行 迭代,直到备选子集波长点个数等于 1,执行 Step8;

Step8: 从所有的备选子集中,选出 RMSECV 最小的备选子集作为最佳波长子集。

在该算法中,Step5 中的蒙特卡罗模拟次数 N 取 500;最大主因子数由全谱的蒙特卡罗交叉验证(MCCV)确定。为了提高算法运行效率,同时尽量保留关键波长,在第一次迭代时,进行波长粗筛,直接剔除掉"不重要"性 p_j 大于 δ 的波长点,为了确定 δ ,以 0.05 为间隔,取值范围为 0 到 1,通过RMSECV确定最优值;在玉米数据集上, δ 取 0.5;在汽油数据集上, δ 取 0.6;之后进行波长精选,用蒙特卡罗方法滤除波长点。

1.4 模型建立及评价

将 MC-DOSC 得到的最佳波长子集建立 PLS 模型,并同

MCUVE, GA 和 CARS 三种方法所建 PLS 模型进行对比,以此验证 MC-DOSC 算法的有效性。采用预测集相关系数 R和预测均方根误差(RMSEP)来评价模型的性能。所有算法均在 Matlab 2015b 软件中实现。

2 结果与讨论

2.1 MC-DOSC 算法的波长选择结果

2.1.1 玉米数据集

以玉米数据集的 MC-DOSC 波长选择过程为例,如图 2 所示。图 2(a)展示了每一次迭代中滤除"不重要"性较大的波长点这一过程,第一次筛选直接剪掉"不重要"性大于 δ=0.5 的波长点,达到粗选的目的,然后用蒙特卡罗方法进行细选,直到波长点不能再减少为止。从图 2(b)可以看出,随着迭代的不断进行,备选子集所建模型的 RMSECV 逐步减小到最小值,然后在 34 次迭代后迅速增大。分析认为 RMSECV 达到最小值前,滤除的是干扰波长点或者对模型贡献很小的波长点;在达到最小值后,滤除了对模型贡献很大的关键波长点。将第 34 次迭代的备选子集作为最佳波长子集。图 2(c)为最佳波长子集所建 PLS 模型与全光谱所建 PLS 模型的 RMSEP 对比图,由图 2(c)可知,经过 MC-DOSC 波长筛选后,模型的预测能力大幅度提高了。

2.1.2 汽油数据集

在汽油数据集中,波长选择过程与玉米数据集类似,如图 3(a)、图 3(b)所示,在第 16 次迭代后,RMSECV 达到最

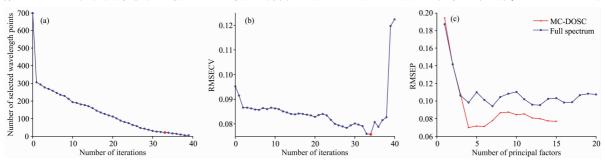


图 2 (a)每一次迭代选择的波长点个数;(b)每一次迭代交叉验证均方根误差;(c)波长选择前后 RMSEP的比较

Fig. 2 (a) The number of wavelength points selected in each iteration; (b) RMSECV in each iteration;

(c) Comparison of RMSEP before and after wavelength selection

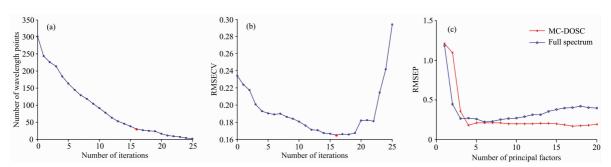


图 3 (a) 每一次迭代选择的波长点个数; (b) 每一次迭代交叉验证均方根误差; (c) 波长选择前后 RMSEP 的比较 Fig. 3 (a) The number of wavelength points selected in each iteration; (b) RMSECV in each iteration;

(c) Comparison of RMSEP before and after wavelength selection

小值,最终选择第16次迭代时对应的波长点建立 PLS模型。 从图 3(c)可以看出,相比于全谱的 PLS模型,经过 MC-DOSC 波长筛选后,模型的预测能力同样提高了。

2.2 不同方法建模结果的对比及分析

为了验证 MC-DOSC 算法的有效性,将 MCUVE, GA, CARS 和 MC-DOSC 4 种不同方法选择的波长点建立 PLS 模型,然后对比不同方法所建模型预测能力的差异,以验证 MC-DOSC 算法的有效性。

2.2.1 玉米数据集

4 种方法在玉米数据集选择的波段如图 4 所示,可以看出,4 种方法均选择了 1 700 和 2 300 nm 附近的波长点,对应着 C-H 键的伸缩振动区域。GA, CARS 和 MC-DOSC 几乎都仅选择了 1 650~2 498 nm 波段的波长点,此波段对应着 C—H 键的一级倍频与合频。说明对于玉米油分含量的预测,1 650~2 498 nm 波段有着关键作用。

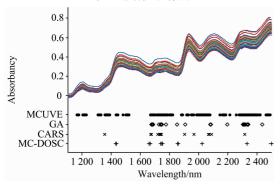


图 4 4 种方法选择的波长点

Fig. 4 Wavelength points selected by four methods

从表 1 可知,相比于全光谱,用 4 种不同方法筛选波长后,模型的预测能力都有一定程度的提升。其中 MCUVE 选择的波长点最多,预测结果也最差。这主要是因为,MCUVE 滤除了噪声波长点,保留了不是噪声但对模型没有贡献的波长点。CARS 选择了较少的波长点,预测结果也较好。MC-DOSC 选择了 15 个波长点,略少于 CARS。MC-DOSC 预测结果是 4 种方法中最好的,这是因为,MC-DOSC 滤除了噪声波长点,也滤除了对模型没有贡献的波长点。相较于全光谱的 PLS 模型,经过 MC-DOSC 算法筛选波长后,PLS 模型的预测能力大幅度提高。验证相关系数 R,从0.828 2提高到 0.931 4, RMSEP 从 0.109 8 减少到 0.071 3。

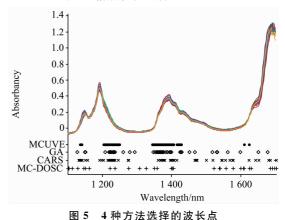
表 1 不同模型在玉米数据集预测能力的对比 Table 1 Comparison of prediction ability of

different models in corn dataset

模型	波长点个数	测试集相关系数 R _p	RMSEP
PLS	700	0.828 2	0.1098
MCUVE	296	0.8624	0.099 2
GA	32	0.865 3	0.098 2
CARS	17	0.916 2	0.078 5
MC-DOSC	15	0.931 4	0.0713

2.2.2 汽油数据集

4 种方法在汽油数据集选择的波段如图 5 所示,可以看出,4 种方法选择的波长点数相差不大,均选择了 1 410 nm附近的波段,这对应着芳烃和 C—H 的合频吸收区域。MCUVE选择的波段较为集中,几乎没有选择 1 450~1 700 nm波段的波长点,而这个区域是各种 C—H 基团以及 N—H 基团、O—H 基团的伸缩振动的一级倍频吸收区域,因此MCUVE对映的预测误差也最大。而 CARS 和 MC-DOSC 都选择了其他方法没有选择的 1 700 nm 附近的波段,这对映着甲基 C—H 的一级倍频吸收区域。



国。 11777亿是14171次代派

Fig. 5 Wavelength points selected by four methods

从表 2 可知,尽管 MCUVE 选择了最多的波长,但预测结果并不理想。MC-DOSC 选择了最少的波长,预测结果也是最好的。相较于全光谱 PLS 模型,经过 MC-DOSC 算法筛选波长后,验证相关系数 $R_{\rm P}$ 从 0.987 5 提高到 0.993 9,RMSEP 从 0.255 5 减小到 0.178 8。

表 2 不同模型在汽油数据集预测能力的对比 Table 2 Comparison of prediction ability of different models in gasoline data sets

模型	波长点个数	测试集相关系数 R _p	RMSEP
PLS	301	0.987 5	0.255 5
MCUVE	68	0.9903	0.225 2
GA	43	0.9918	0.206 6
CARS	44	0.9911	0.215 2
MC-DOSC	31	0.9939	0.1788

3 结 论

针对近红外光谱的特性与光谱分析存在的问题,提出一种新的波长选择算法 MC-DOSC。该算法利用直接正交校正 (DOSC)得到的权重向量 w 度量波长的"不重要"性,再结合蒙特卡罗方法进行波长筛选。用 MCUVE, GA 和 CARS 在两个数据集上的实验进行了对比,验证了 MC-DOSC 算法的有效性。实验结果表明,MC-DOSC 是一种有效的波长选择算法,具有广泛的应用价值。

References

- [1] HAN Jian, LI Xue-zhao, CAO Zhi-min, et al(韩 建,李雪昭,曹志敏,等). Chinese Journal of Scientific Instrument(仪器仪表学报), 2019, 40(6); 78.
- [2] Chen J M, Yang C H, Zhu H Q, et al. Chemometrics and Intelligent Laboratory Systems, 2018, 182: 188.
- [3] Han QJ, Wu HL, Cai CB, et al. Analytica Chimica Acta, 2008, 612: 121.
- [4] Tang G, Wei B, Wu DC, et al. Journal of Applied Spectroscopy, 2018, 85(1): 109.
- [5] Liu Y S, Zhou S B, Liu W X, et al. Journal of Near Infrared Spectroscopy, 2018, 26(1): 34.
- [6] JIANG Wei-wei, LU Chang-hua, ZHANG Yu-jun, et al(蒋薇薇, 鲁昌华, 张玉钧, 等). Journal of Electronic Measurement and Instrument(电子测量与仪器学报), 2017, 31(12): 1960.
- [7] SHI Ji-yong, ZOU Xiao-bo, ZHAO Jie-wen, et al(石吉勇,邹小波,赵杰文,等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2011, 30(10): 458.
- [8] Deng BC, Yun YH, Liang YZ, et al. The Analyst, 2014, 139(19): 4836.
- [9] Li H D, Liang Y Z, Xu Q S, et al. Analytica Chimica Acta, 2009, 648: 77.
- [10] Li H D, Xu Q S, Liang Y Z, et al. Analytica Chimica Acta, 2012, 740: 20.
- [11] ZHANG Feng, TANG Xiao-jun, TONG Ang-xin, et al(张 峰, 汤晓君, 全昂鑫, 等). Chinese Journal of Scientific(仪器仪表学报), 2020, 41(1): 64.
- [12] YU Lei, ZHU Ya-xing, HONG Yong-sheng, et al(于 雷,朱亚星,洪永胜,等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(33): 138.
- [13] ZHAO Huan, HUAN Ke-wei, SHI Xiao-guang, et al(赵 环, 宦克为, 石晓光, 等). Analytical Chemistry(分析化学), 2018, 46 (1): 136.
- [14] CHEN Yuan-yuan, WANG Zhi-bin, WANG Zhao-ba(陈媛媛, 王志斌, 王召巴). Infrared and Laser Engineering(红外与激光工程), 2014, 43(8): 2715.
- [15] Westerhuis J A, de Jong S, Smilde A K. Chemometrics and Intelligent Laboratory Systems, 2001, 56(1): 13.
- [16] Jiang H, Zhang H, Chen Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 149: 1.

A Wavelength Selection Method Combining Direct Orthogonal Signal Correction and Monte Carlo

XIE Lin-jiang, HONG Ming-jian*, YU Zhi-rong

School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China

Abstract In the analysis of near-infrared spectroscopy data, full-spectrum data has the characteristics of multiple wavelength points, large redundancy, and serious collinearity. This leads to some wavelength points that have no positive effect on the establishment of the correction model and even reduce the model's predictive ability. Wavelength selection has proven to be an important method to avoid above problems effectively. Aiming at the characteristics of near-infrared spectroscopy, a wavelength selection algorithm based on the combination of Direct Orthogonal Signal Correction (DOSC) and Monte Carlo (MC) is proposed. Unlike most methods of selecting wavelength according to its "importance", MC-DOSC selects wavelength according to its "unimportance". The "unimportance" of wavelength is measured by the weight W of DOSC. Specifically, first, normalize was the probability of wavelength being filtered to establish the probability model of wavelength selection, and Monte Carlo random sampling is used to obtain the set of N wavelength subsets. The selected wavelength point is used to establish a PLS model in each sampling process, and the corresponding cross-validation root mean square error (RMSECV) is calculated. After N times of random sampling, the wavelength subset corresponding to the PLS model with minimum RMSECV is selected as the candidate subset. The spectral data contained in the candidate subset is used as a new spectral matrix, and the above process is repeated until the RMSECV no longer drops. After the iteration stops, the candidate subset with the smallest RMSECV is taken as the best wavelength subset. And compared with the three algorithms of Monte Carlo Uninformative Variable Elimination (MCUVE), Genetic Algorithm (GA) and Competitive Adaptive Weight Sampling (CARS). Experimental results show that the algorithm can greatly reduce the number of wavelength points, and the prediction ability of the corresponding PLS model is also improved. In the experimental results of the corn data set, the number of wavelength points is reduced from 700 in the full spectrum to 15. The correlation coefficient of the prediction set is increased from 0. 828 2 to 0. 931 4, and the RMSEP is reduced from 0. 109 8 to 0. 071 3. In the experimental results of the gasoline data set, the number of wavelength points was reduced from 301 in the full spectrum to 31. The correlation coefficient of the prediction set was increased from 0. 987 5 to 0. 993 9, and the RMSEP was reduced from 0. 255 to 0. 178 8. The performance of this algorithm in the two data sets is better than the three algorithms compared.

Keywords Near-infrared spectroscopy; Wavelength selection; Direct orthogonal signal correction; Monte Carlo; Partial least squares

(Received Jan. 7, 2021; accepted Apr. 11, 2021)

* Corresponding author

(上接 422 页)

报告形式

为充分提高会议学术交流的效率,会议将采用"口头报告"和"墙报展示"两种方式进行学术交流。无论是口头报告还是墙报展示,均属大会同等学术交流。为尊重个人意见和便于组委会的安排,请大家在会议注册时,提交"口头报告"或"墙报"的题目。为了鼓励博士、硕士研究生积极参与学术交流活动,本次会议将继续设立"优秀青年论文奖"和"优秀墙报奖",表彰那些研究水平高、能突出研究内容要点、条理清晰的"口头报告"和"墙报",大会将给获奖作者颁发优秀论文证书和奖金。同时会议还将邀请国内外知名专家学者就光谱有关学术领域的前沿热点问题作大会报告和主题报告。

主要报告形式有:

- 1. 大会邀请报告:主要邀请国内外知名专家学者报告光谱分析的前沿技术在各个领域的最新研究进展。
- 2. 主题邀请报告:本次会议将选择光谱技术的热点应用领域,开设多个专题论坛,邀请在该领域的知名专家作论坛主题报告。
 - 3. 口头报告: 由参会代表申请、组委会审核方式确定报告人选。
 - 4. 青年论坛报告: 为博士、硕士研究生开设交流平台, 并评选"优秀青年论文奖"。
 - 5. 墙报展示:作为本次会议的主要交流和展示形式。会议统一安排墙报讲解时间,希望作者按时到位讲解。

重要时间

开通会议注册系统: 2022年5月10日

论文截稿日期: 2022 年 6 月 30 日

第二轮会议通知:2022年5月

第三轮会议通知: 2022年10月

会议召开期: 2022年11月11日-14日

会议组织机构、注册费及缴纳方式、宾馆住宿介绍及住房预定等信息将在 2022 年 5 月初在会议主页上发布,请您经常浏览光谱网上会议主页,了解会议筹备情况和会议具体安排。网址: http://www.sinospectroscopy.org.cn

产品展示

会议热忱邀请国内外仪器厂商参会及展示仪器设备,大会组委会将在本次会议的网站和会议现场提供展出场所,希望各 仪器厂商充分利用这次机会展示自己的最新产品。

(下转 453 页)