第32卷，第12期
2012年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 32，No. 12，pp3208-3212
December，2012

# 变量筛选方法结合局部线性嵌入理论用于近红外光谱定量模型优化

郝 勇[1]，孙旭东[1]，杨 强[2]

1. 华东交通大学机电工程学院，江西 南昌 330013
2. 日照职业技术学院机电工程学院，山东 日照 276826

**摘 要** 变量筛选策略结合局部线性嵌入(local linear embedding，LLE)理论用于近红外光谱(near infrared spectroscopy，NIRS)定量模型优化。蒙特卡罗无信息变量消除方法(monte carlo uninformation variable elimination，MCUVE)和连续投影算法(successive projections algorithm，SPA)以及两者结合的变量筛选策略用于NIRS冗余变量的剔除；偏最小二乘回归(partial least squares regression，PLSR)和LLE-PLSR用于复杂样品光谱定量模型的构建。结果表明：MCUVE方法既能有效的提取信息变量，同时可以提高模型的预测精度；LLE-PLSR可以得到比PLSR方法更加准确的定量分析模型；MCUVE结合LLE-PLSR是一种有效的光谱定量分析方法。

**关键词** 近红外光谱；蒙特卡罗无信息变量消除；连续投影算法；局部线性嵌入

**中图分类号**：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2012)12-3208-05

## 引 言

近红外光谱分析技术广泛用于复杂样品的定性和定量分析，具有分析效率高、速度快、成本低、非破坏性和易于在线分析等特点[1-4]。然而，NIRS定性和定量分析方法必须通过建立校正模型来实现，因此，建模方法及模型优化对于提高NIRS的分析能力具有很重要的意义[5,6]。同时，NIRS不仅反映物质的化学组成和含量，同时也包含了由被测物的温度、表面纹理以及密度等因素引起的光谱响应。因此，在建立分析模型时，从光谱图中提取与组成相关的信息，消除与被测目标物无关的其它影响因素的干扰，对建立稳定可靠的分析模型，提高分析精度非常关键。光谱预处理方法(如标准正态变量校正、多元散射校正、中心化、Savitzky-Golay求导以及小波求导等)和变量筛选策略(如连续投影算法和蒙特卡罗无信息变量消除方法等)可以有效地消除背景噪声及其它物理因素的干扰，提高谱图与化学成分间的相关性[7]。

在定量分析中，常用的主成分回归(principal component regression，PCR)、偏最小二乘回归都属于线性回归方法，当被测目标物含量与光谱响应间存在非线性关系时，采用线性的回归方法无法实现光谱信息的充分提取，影响模型精度。针对光谱的非线性特性，一些非线性机器学习算法如人工神经网络(artificial neural network，ANN)和支持向量回归(support vector regression，SVR)等算法也被引入NIRS光谱处理中，在这些非线性学习方法中，模型的优化过程存在耗时长或多次优化结果不一致等不足，因此建模效率较低[8,9]。为此，探讨光谱的非线性降维结合PLSR方法用于NIRS的定量分析非常必要。

本工作以NIRS为研究手段，采用局部线性嵌入变换方法对光谱进行非线性降维，结合PLSR方法，对谷物样品的淀粉含量和药片活性成分含量进行定量分析，探讨变量筛选策略结合LLE-PLSR方法在复杂样品NIRS定量分析中的应用。

## 1 原理与算法

### 1.1 蒙特卡罗无信息变量消除算法和连续投影算法

蒙特卡罗(monte carlo，MC)方法，是一种基于“随机数”和概率统计来考察问题的计算方法。MC方法在分析复杂多变量问题时非常有效，在统计检验、优化过程、系统分析和信号探测等诸多领域已得到广泛应用。该方法首先采用随机采样技术，从大量的样本中选取部分样品建立PLSR模型，并得到模型的系数矩阵 $\boldsymbol{\beta}=[\beta_1,\cdots,\beta_j]$；$\beta_j$ 代表了第 $j$ 个波长点所相应变量对模型的贡献，每个变量的重要性可以通过变

**收稿日期**：2012-06-26，**修订日期**：2012-09-10

**基金项目**：国家自然科学基金项目(21265006)，江西省青年科学基金项目(20114BAB2011010)，华东交通大学博士启动基金项目(01309021)和江苏省企业博士集聚计划项目(2011)资助

**作者简介**：郝 勇，1978年生，华东交通大学机电工程学院讲师 e-mail：haonm@163.com

量的稳定性值来衡量。稳定性可以定义为

$$s_j = \text{mean}(\beta_j)/\text{std}(\beta_j) \quad j = 1, \cdots, 2p \tag{1}$$

其中 $s_j$ 为第 $j$ 个波长点所相应变量的稳定性值，mean($\beta_j$)和 std($\beta_j$)分别为矩阵 $\boldsymbol{\beta}$ 的平均值和标准偏差。从式(1)可以看出，$\beta_j$ 的平均值越大，标准偏差越小时，该变量的稳定性值越大，相应的变量越重要。因此，可以通过设定一定的阈值将不重要的变量去掉，并利用保留的光谱变量建立优化模型[10,11]。

连续投影算法能够有效剔除众多波长变量间的共线性影响，并使向量间的共线性达到最小，降低模型的复杂度，提高建模的速度和效率，以其简便、快速的特点在多种样品波长选取中得到了很好的应用效果。其原理是假设校正集的样本数为 $M$，光谱包含的变量数为 $K$，组成光谱矩阵 $\boldsymbol{X}_{M\times K}$，SPA 是一种前向循环选择方法，以初始迭代向量为基础，计算其在未选入的波长上的投影，将投影向量最大的波长引入到波长组合，直到循环 $N$ 次($N<M-1$)。每一次新选入的波长，都与前一个线性关系最小。最后得到 $N\times K$ 对波长组合，对每一对波长组合分别建立定标模型，以计算的均方误差值(RMSE)来判断所建模型的优劣。最小的 RMSE 值所对应的波长组合为最佳的波长组合[12,13]。

### 1.2　局部线性嵌入(LLE)理论

LLE 的主要思想是利用样本空间中局部的线性来逼近全局的非线性，通过将 LLE 与 PLSR 结合，可以解决光谱建模中存在的非线性问题[14]。其算法的具体描述为：设 $X=\{x_1, x_2, \cdots, x_N\}\in R^D$ 为输入样本，其低维映射为 $Y=\{y_1, y_2, \cdots, y_N\}\in R^d$，对于 NIRS，$x_i$ 为原始光谱，$y_i$ 为特征向量，高维空间任意一点 $x_i$ 均可表示为其 $K$ 领域内样本点的线性组合

$$x_i \approx \sum_{j=1}^{K} \omega_{ij} x_{ij} \tag{2}$$

其中，$x_{ij}$ 为高维空间中距离 $x_i$ 最近的 $k$ 个样本点，$\omega_{ij}$ 为线性重构系数。LLE 通过局部线性关系实现高维空间样本在低维空间的映射，具体算法流程如下：

(1)计算样本 $X$ 中任意两点 $x_i$ 和 $x_j$ 间的欧氏距离 $d_x(i, j)$，则距离矩阵为 $\boldsymbol{D}_{ij}=d_x(i,j)$；

(2)根据 $\boldsymbol{D}_{ij}$ 找出样本集 $X$ 中距离 $x_i$ 最近的 $k$ 个点 $\{x_{ij}\}_{j=1}^{k}$；

(3)以表达式 $\min\sum_{i=1}^{N} \| x_i - \sum_{j=1}^{k} \omega_{ij} x_{ij} \|^2$ 为目标函数(其中 $\sum_{j=1}^{k} \omega_{ij} = 1$)，计算各点 $x_i$ 和其近邻点 $\{x_{ij}\}_{j=1}^{k}$ 的线性重构系数；

(4)已知 $\omega_{ij}$，以表达式 $\min\sum_{i=1}^{N} \| y_i - \sum_{j=1}^{k} \omega_{ij} y_{ij} \|^2$ 为目标函数，计算低维映射，即求解矩阵 $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})$ 的第 2 个到 $d+1$ 个最小特征值对应的特征向量 $Y$，其中 $\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{W}$ 为 $N\times N$ 的方阵，若 $x_i$ 和 $x_j$ 为近邻点，$\boldsymbol{W}_{ij}=\boldsymbol{\omega}_{ij}$，否则 $\boldsymbol{W}_{ij}=0$。从而实现 LLE 降维变换。

## 2　实验部分

### 2.1　数据验证

实验中采用两组数据进行验证。数据集 1 为谷物样品的 NIRS(下载地址：http://software.eigenvector.com/Data/Corn/corn.mat)，其中包含了 80 个谷物样品的 NIRS 以及样品中水分、含油率、蛋白以及淀粉的含量，分别采用三台不同的近红外光谱仪(M5，MP5 和 MP6)进行光谱采集。以 MP6 光谱仪器采集的 NIRS 以及谷物中的淀粉含量为研究对象[15]。数据集 2 为 University of Copenhagen 提供的药片 NIRS 数据以及相应的活性成分含量值(单位：Wt%)[16]。样本数量为 310 个，样品在粉末状进行透射光谱的采集，波数范围为：7 398.3～10 507.0 $\text{cm}^{-1}$，采样间隔为 8 $\text{cm}^{-1}$。选用 KS(Kennard-Stone)算法分别将两个数据集样品按照 2∶1 的比例划分为校正集和验证集。两组数据集样品信息如表 1 所示。

**Table 1　Statistical composition of sample sets**

| 分析物 | 校正集 | | | | 验证集 | | | |
|---|---|---|---|---|---|---|---|---|
| | $n^a$ | 含量范围 | 平均值 | S. D.$^b$ | $n^a$ | 含量范围 | 平均值 | S. D.$^b$ |
| 淀粉 | 53 | 62.826～66.472 | 64.711 | 0.838 | 27 | 63.247～65.903 | 64.666 | 0.801 |
| 活性成分 | 206 | 4.610～9.382 | 7.463 | 1.279 | 104 | 4.715～9.768 | 7.359 | 1.330 |

$^a n$=样本数；$^b$S. D.=标准偏差

### 2.2　光谱数据预处理

采用 PLSR 方法建立回归模型，采用校正集建立模型，验证集用于考察模型的预测能力。PLSR 模型的因子数通过蒙特卡罗交叉验证(MCCV)结合 $F$ 检验确定。

在 NIRS 分析中，原始光谱含有与样品组成无关的信息，如环境温度、器件噪声以及样品背景干扰等因素，需采用预处理方法进行消除。分别采用中心化(centering)、矢量归一化(vector normalization，SNV)、多元散射校正(multiplicative scatter correction，MSC)、中心化结合 Savitzky-Golay(SG)求导以及中心化结合小波求导(wavelet derivative，WD)[17]的预处理方法探讨预处理方法对模型精度的影响。

### 2.3　模型的评价指标

模型的评价指标包括交叉验证均方根误差(root mean square error of cross-validation，RMSECV)、校正相关系数(correlation coefficient of calibration，$R_c$)、预测均方根误差(root mean square error of prediction，RMSEP)和预测相关系数(correlation coefficient of prediction，$R_p$)。在校正模型的选择过程中，采用 RMSECV 和 $R_c$ 作为评价指标；在模型

预测能力的验证时，采用 RMSEP 和 $R_p$ 作为评价指标。

## 3 结果与讨论

### 3.1 光谱数据预处理方法的选择

预处理方法的选取对光谱模型的预测精度和稳定性具有重要意义。SNV 和 MSC 的光谱预处理方法可以校正基线，减少样品散射对光谱的影响，将光谱间的差异以样品间成分差异的方式表征出来；光谱 centering 处理可以消除仪器每次测量的能量差异；一阶导数可以去除部分线性或接近线性的背景和噪声对目标光谱的影响；小波求导不仅可以消除背景和噪声的影响，而且可以抑制由于求导引入的光谱噪声。表 2 所示为不同光谱预处理方法对模型精度的影响。从表 2 中可知，对于两组数据，采用中心化结合小波导数的预处理方法，都得到了最好的模型结果。与原始光谱相比，模型的 RMSECV 都得到了降低，分别从 0.239 和 0.516 降低为 0.086 和 0.325，$R_c$ 分别由 0.879 和 0.919 提高到 0.984 和 0.967。对于固体样品，中心化结合小波求导是一种较好的预处理方法。

**Table 2 Comparison of the results obtained by different preprocessing methods**

| 预处理方法 | 谷物(淀粉) | | | 药片(活性成分) | | |
|---|---|---|---|---|---|---|
| | PLS 因子数 | RMSECV | $R_c$ | PLS 因子数 | RMSECV | $R_c$ |
| 原始光谱 | 9 | 0.239 | 0.879 | 8 | 0.516 | 0.919 |
| 中心化 | 15 | 0.109 | 0.974 | 6 | 0.352 | 0.961 |
| 矢量归一化 | 13 | 0.119 | 0.969 | 5 | 0.337 | 0.965 |
| 多元散射校正 | 13 | 0.118 | 0.969 | 5 | 0.337 | 9.964 |
| 中心化+一阶导数 | 7 | 0.138 | 0.958 | 5 | 0.356 | 0.960 |
| 中心化+小波导数 | **15** | **0.086** | **0.984** | **8** | **0.325** | **0.967** |

### 3.2 变量筛选方法的选择

NIRS 波长数目较多，而且为了构建稳健的模型，需要收集大量的样本。NIRS 分析中，引起模型不稳定和预测结果较差的主要因素之一是冗余光谱信息干扰。因此，在建立多元校正模型时，消除不相关信息变量不仅可以简化校正模型，而且可以从准确度和稳健性两方面改善模型的预测能力。通过选择包含样品性质特征或其所含成份性质特征的波长(变量)，代替全谱去建立模型，会得到更准确的校正模型。分别采用 MCUVE 方法、SPA 方法以及 MCUVE-SPA 方法对两组数据集进行变量筛选，模型的预测结果如表 3 所示。

由表 3 可知，经变量优选后，对于两组数据，采用 MCUVE 方法都得到了最好的预测结果，RMSEP 分别由 0.145 和 0.327 降低为 0.089 和 0.324，$R_p$ 分别由 0.983 和 0.969 提高为 0.994 和 0.970，建模变量数分别由 700 和 404 减小为 200 和 340；而 SPA 方法及其与 MCUVE 相结合的方法结果稍差，主要原因是 MCUVE 方法在计算稳定性的时候采用 PLSR 方法，而 SPA 方法采用线性回归方法，与线性回归方法相比，PLSR 方法本身具有一定的消噪作用。

**Table 3 Comparison of the prediction results by PLSR models with different variable selection methods**

| 变量筛选方法 | 谷物(淀粉) | | | 药片(活性成分) | | |
|---|---|---|---|---|---|---|
| | $n$[a] | RMSEP | $R_p$ | $n$[a] | RMSEP | $R_p$ |
| None | 700 | 0.145 | 0.983 | 404 | 0.327 | 0.969 |
| MCUVE | **200** | **0.089** | **0.994** | **340** | **0.324** | **0.970** |
| SPA | 36 | 0.171 | 0.977 | 19 | 0.337 | 0.967 |
| MCUVE-SPA | 33 | 0.216 | 0.962 | 32 | 0.339 | 0.967 |

[a] $n$=number of variables

### 3.3 定量方法的比较及参数的确定

在 NIRS 分析中，由于体系中各组分的相互作用、仪器的噪声及基线漂移等原因，会引起光谱响应与样品组分含量的化学测定值之间的非线性。因此，在 NIRS 的定量分析中，需要引入非线性的校正方法。尽管 PLSR 在一定程度上可以校正非线性因素，但如果非线性非常严重，线性校正方法不能得到理想的分析模型。必须针对分析体系特有的非线性特征，建立非线性校正模型。

分别采用 LLE-PLSR 和 MCUVE-LLE-PLSR 方法建立非线性校正模型。模型的计算结果如表 4 所示。其中在 LLE 降维过程中，需要对参数 $k$ 和 $d$ 进行优化，其中 $k$ 为邻域参数，$d$ 为样本本征维数，$k$ 和 $d$ 均取整数，采用网格搜索法分别对两组数据集的 $k$ 和 $d$ 进行计算，优化示意如图 1 所示。其中图 1(a)为淀粉 LLE-PLSR 模型中参数 $k$ 和 $d$ 优化过程示意图，图 1(b)为淀粉 MCUVE-LLE-PLSR 模型中参数 $k$ 和 $d$ 优化过程示意图，图 1(c)为活性成分 LLE-PLSR 模型中参数 $k$ 和 $d$ 优化过程示意图，图 1(d)为活性成分 MCUVE-LLE-PLSR 模型中参数 $k$ 和 $d$ 优化过程，计算优化过程中的 RMSECV 值，RMSECV 取最小值时对应的 $k$ 和 $d$ 即为优化结果。图 1 中(●)对应最小的 RMSECV 值。

**Table 4 Comparison of the predicted results for the PLSR models**

| Modeling method | 谷物(淀粉) | | 药片(活性成分) | |
|---|---|---|---|---|
| | RMSEP | $R_p$ | RMSEP | $R_p$ |
| LLE-PLSR | 0.117 | 0.989 | 0.280 | 0.978 |
| MCUVE-LLE-PLSR | **0.082** | **0.995** | **0.266** | **0.980** |

表 4 分别计算了验证集的 RMSEP 和 $R_p$，从表中可知，对于淀粉和活性成分的定量分析，采用 MCUVE-LLE-PLSR 得到了比 LLE-PLSR 更好的预测结果；与 MCUVE-PLSR 相比(表 3 所示)，RMSEP 分别由 0.089 和 0.324 降低为 0.082 和 0.266，预测相关系数分别由 0.994 和 0.970 提高到 0.995 和 0.980。

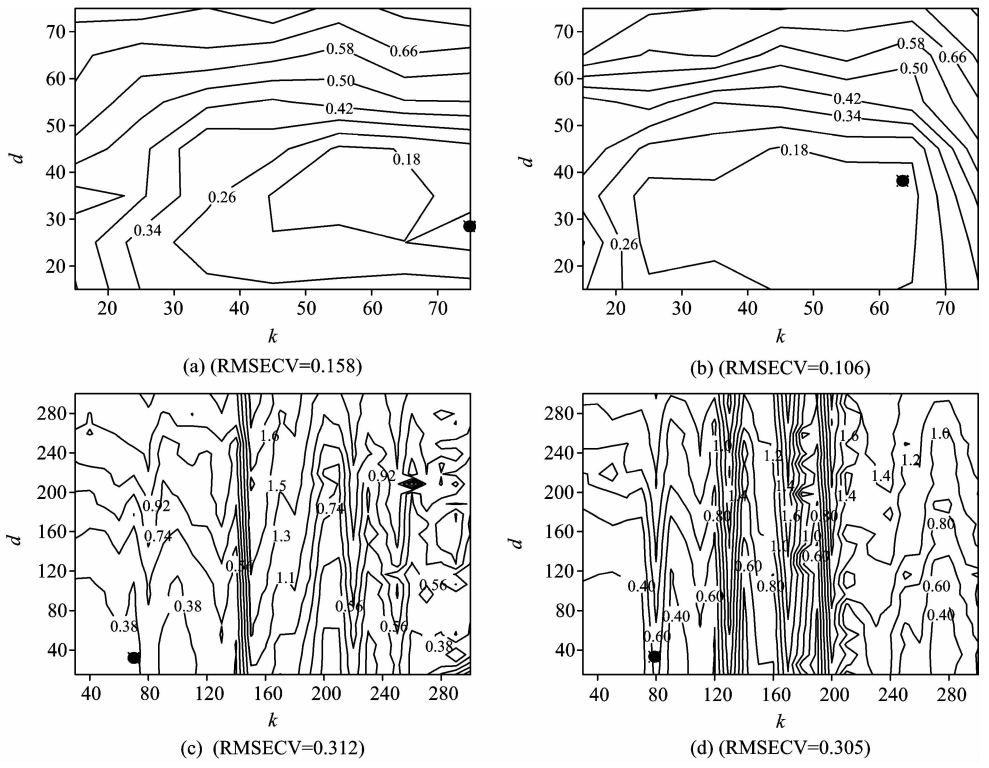


**Fig. 1　The optimization process of *k* and *d* for LLE dimension reduction in different PLSR models**

(a): LLE-PLSR model for starch; (b): MCUVE-LLE-PLSR model for starch;
(c): LLE-PLSR model for active component; (d): MCUVE-LLE-PLSR model for active component

## 4　结　论

分别采用 MCUVE, SPA 和 MCUVE-SPA 方法对谷物和药片的近红外光谱进行变量筛选，优选后的变量分别采用 PLSR 和 LLE-PLSR 进行建模分析。通过优化计算可知，对于两组固体样品数据，MCUVE 方法可以在提高预测精度的前提下，有效的减少建模变量数；LLE-PLSR 建模方法得到了比 PLSR 更好的预测结果。MCUVE-LLE-PLSR 方法是一种简单、准确的非线性光谱校正方法。

## References


[1] Gaydou V, Kister J, Dupuy N. Chemometrics and Intelligent Laboratory Systems, 2011, 106(2): 190.
[2] Yan Hui, Han Bangxing, Wu Qiongying, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2011, 79(1): 179.
[3] Lidia Esteve Agelet, David D Ellis, Susan Duvick, et al. Journal of Cereal Science, 2012, 55(2): 160.
[4] Liao Yitao, Fan Yuxia, Cheng Fang. Journal of Food Engineering, 2012, 109(4): 668.
[5] Shen Fei, Ying Yibing, Li Bobin, et al. Food Research International, 2011, 44(5): 1521.
[6] Sinelli N, Casiraghi E, Barzaghi S, et al. Food Research International, 2011, 44(5): 1427.
[7] Asmund Rinnan, Frans van den Berg, Soren Balling Engelsen. Trends in Analytical Chemistry, 2009, 28(10): 1201.
[8] Cai Chenbo, Yang Hongwei, Wang Bo, et al. Vibrational Spectroscopy, 2011, 56(2), 202.
[9] Mohammadreza Khanmohammadi, Amir Bagheri Garmarudi, Nafiseh Khoddami, et al. Mirochemical Journal, 2010, 95(2): 337.
[10] Han Qingjuan, Wu Hailong, Cai Chenbo, et al. Analytica Chimica Acta, 2008, 612(2): 121.
[11] Cai Wensheng, Li Yankun, Shao Xueguang. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 188.
[12] Soares M C U, Galvao R K H, Araujo M C U, et al. Analytica Chimica Acta, 2011, 689(1): 22.
[13] Liu Fei, He Yong. Food Chemistry, 2009, 115(4): 1430.
[14] Hou Chenping, Wang Jing, Wu Yi, et al. Neurocomputing, 2009, 72(12): 2368.
[15] Kalivas J H. Chemometrics and Intelligent Laboratory Systems, 1997, 37(2): 255.
[16] Dyrby M, Engelsen S B, Nørgaard L, et al. Applied Spectroscopy, 2002, 56(5): 579.
[17] Shao Xueguang, Ma Chaoxiong. Chemometrics and Intelligent Laboratory Systems, 2003, 69(1): 157.

# Variable Selection Methods Combined with Local Linear Embedding Theory Used for Optimization of Near Infrared Spectral Quantitative Models

HAO Yong[1], SUN Xu-dong[1], YANG Qiang[2]
1. College of Machanical and Electronic Engineering, East China Jiaotong University, Nanchang 330013, China
2. College of Machanical and Electronic Engineering, Rizhao Polytechnic, Rizhao 276826, China

**Abstract** Variables selection strategy combined with local linear embedding (LLE) was introduced for the analysis of complex samples by using near infrared spectroscopy (NIRS). Three methods include monte carlo uninformation variable elimination (MCUVE), successive projections algorithm (SPA) and MCUVE connected with SPA were used for eliminating redundancy spectral variables. Partial least squares regression (PLSR) and LLE-PLSR were used for modeling complex samples. The results shown that MCUVE can both extract effective informative variables and improve the precision of models. Compared with PLSR models, LLE-PLSR models can achieve more accurate analysis results. MCUVE combined with LLE-PLSR is an effective modeling method for NIRS quantitative analysis.

**Keywords** Near infrared spectroscopy (NIRS); Monte carlo uninformation variable elimination (MCUVE); Successive projections algorithm (SPA); Local linear embedding (LLE)

(Received Jun. 26, 2012; accepted Sep. 10, 2012)

---

## 关于《光谱学与光谱分析》收取审稿费的通知

尊敬的《光谱学与光谱分析》广大作者、读者同志们，本刊自 2006 年底采用由“北京玛格泰克科技发展有限公司”开发的投稿系统实现网络采编以来，进一步扩展了审稿专家队伍。本刊参考同类期刊的现行做法，决定自 2010 年 12 月 1 日以后登记的稿件向投稿作者收取审稿费 100 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

**1. 没有创新的一般性稿件，请您不要投稿。**

**2. 没有国家级基金资助的稿件，请您不要投稿。**

**3. 不是光谱专业的稿件，请您不要投稿。**

**4. 与其他文章重合率超过 10%的稿件，请您不要投稿。**

作者在投稿后，将会收到缴纳审稿费的通知。请作者及时从我刊网站(http://www.gpxygpfx.com)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，本刊不再受理。汇款时，请写明详细通信地址、邮政编码、收件人姓名等信息，以便准确寄回发票。

汇款方式(在附言里写明审稿费)：

邮局汇款：北京市海淀区学院南路 76 号，《光谱学与光谱分析》期刊社(收)

邮政编码：100081 联系电话：010-62181070，62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2010 年 12 月 1 日