

分级评价指标优化玉米近红外光谱定性分析模型研究

李佳¹ 常晓莲² 王雅倩³ 刘欢³ 安冬^{1,4} 严衍禄¹

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 北京市农业机械试验鉴定推广站, 北京 100029;
3. 山东科技大学电气与自动化工程学院, 青岛 266590; 4. 农业部农业信息获取技术重点实验室, 北京 100083)

摘要: 验证了种间的相对距离这一评价指标优化玉米近红外光谱定性分析性能的有效性。首先,对 A、B 两组实验数据使用相关性、欧氏距离和熵 3 种常用方法计算原始光谱、预处理后的光谱和特征提取后的光谱数据的种间相对距离,并与每一步得到的正确识别率进行对照分析,得到欧氏距离是一种有效的计算方法。最后,对实验数据 C 采用欧氏距离方法计算数据处理每一过程的种间相对距离,经过对预处理算法的调整,种间相对距离由 0.658 2 增大到了 1.297 2,正确识别率由 40.86% 提高到了 70.08%;通过对特征提取算法的优化,种间相对距离由 1.310 2 增大到了 2.491 0,正确识别率由 68.32% 提高到了 93.27%。通过该评价指标对数据分析过程的评价结果可以看出,正确识别率显著提高,使模型得到了优化。

关键词: 近红外光谱; 相对距离; 正确识别率; 模型优化

中图分类号: O657.33 **文献标识码:** A **文章编号:** 1000-1298(2017)S0-0412-05

Effect of Evaluation Index on Optimizing the Near-infrared Spectral Qualitative Analysis of Corn

LI Jia¹ CHANG Xiaolian² WANG Yaqian³ LIU Huan³ AN Dong^{1,4} YAN Yanlu¹

(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

2. Beijing Agricultural Machinery Test and Appraisal Station, Beijing 100029, China

3. College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China

4. Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, Beijing 100083, China)

Abstract: Near-infrared spectrum analysis as a rapidly developing technique has been applied in recognition analysis because of their simplicity, promptness and low cost. It was used to build an effective model to qualitatively analyze the corn. To evaluate the analysis results, an innovative grading evaluation index, defined with the relative distance of inter-species, was proposed for optimizing the near-infrared spectrum analysis process. It was applied to analyze the effect on optimizing the performance of the near-infrared spectrum qualitative analysis of corn. Firstly, two group spectral data were measured including the transmittance of 6 corn species sampled in Beijing (group A) and the reflectance of 6 corn species sampled in Hainan province (group B). The sampling data were processed involving original spectral data, the spectral data after pre-processing, and the spectral data after feature extraction from the group A and B experimental data. The relative distances of inter-species were calculated by using correlation, Euclidean distance, and entropy respectively. The result of contrast analysis showed that Euclidean distance was an effective calculation method for varieties recognition with good performance both in group A and B. Secondly, the reflectance of 6 corn species sampled in Henan province (group C) was measured. The Euclidean distance method was used to calculate the inter-specific relative distance between process steps as mentioned above. As a result, after the adjustment of the pretreatment algorithm, the relative distance between species increased from 0.658 2 to 1.297 2, and the correct recognition rate increased from 40.86% to 70.08%. By optimizing the feature extraction algorithm, the relative distance between species increased from 1.310 2 to 2.491 0, and the correct recognition rate increased from 68.32% to 93.27%. It was indicated that the correct recognition rate could be improved by the evaluation of the data analysis process.

Key words: near-infrared spectrum; relative distance; correct recognition rate; model optimized

收稿日期: 2017-07-03 修回日期: 2017-11-20

基金项目: 北京市财政项目(PXM2015_036231_000023)

作者简介: 李佳(1994—),女,硕士生,主要从事模式识别等研究,E-mail: 18811128922@163.com

通信作者: 安冬(1977—),女,教授,博士生导师,主要从事信号处理与模式识别等研究,E-mail: andong@cau.edu.cn

引言

近红外光谱分析包括定量和定性两大类,定量分析主要是测定样品中一种或几种成分的含量,目前对其模型的评价指标较成熟,常用的指标有决定系数 R^2 、相对分析误差 RPD、定标集标准偏差 SEC 和验证集标准偏差 SEP 等^[1-5]。定性分析主要用于物质的定性判别分析,即通过比较未知样本和已知样本或标准样本的光谱来确定未知样本的归属,通常使用正确识别率进行性能评价^[6]。AMBROSE 等^[7]通过预测正确率和精度对使用傅里叶变换近红外光谱(FT-NIR)检测玉米种子活性的定性分析性能进行了评价。GHASEMI-VARNAMKHASTI 等^[8]使用分类能力和精度对近红外光谱定性分析应用于不同时期的啤酒类型进行了评价。李天昕等^[9]使用正确识别率、相对距离和相似度对利用近红外反射和透射光谱法鉴定玉米杂交种纯度进行评价。JIA 等^[10]使用中心距离比和识别率评价包衣玉米品种间的可分性和模型性能。

以上工作都通过正确识别率等评价指标对定性分析的性能进行了评价,其作用在于评价分析结果的好坏,而在定性分析过程中评价结果的好坏固然重要,但通过对分析过程的有效优化得到好的分析结果更为关键。以定性分析常用指标正确识别率为例,它是整个分析过程的累积结果,原始光谱数据经过预处理、特征提取和建立模型多个步骤处理,最后得到分析结果,统计出正确识别率。这样的终级评价指标只能说明定性分析性能的优劣,如正确识别率较低,只能说明分析性能不佳,却判断不出问题出现在了预处理、特征提取和建立模型哪一环节,所以以正确识别率为代表的终级评价指标难以直接用于分析过程的有效优化。如果有分级评价指标能够对定性分析过程中每个步骤的效果给出数值化的评价,就能够对每一步骤的效果有一个评判,以便在该步骤进行优化,改善最终的分析结果。

因此,本文提出用于近红外光谱定性分析过程优化的分级评价指标这一概念,并具体针对基于近红外光谱分析的玉米品种定性分析,确定种间、种内差异的比值(相对距离)作为分级评价指标,种间差异越大,种内差异越小,品种的可分性就越好,所以其比值的大小更好地体现了可分性的好坏。本文以玉米种子为实验对象,对分级评价指标用于优化最终分析结果的效果进行研究分析。

1 材料与方 法

1.1 分级评价指标

使用近红外光谱进行玉米品种定性分析时,关键在于减少品种内的样本差异性,同时增大品种间的样本差异性,使异类样本区分度提高,从而提高定性分析性能。所以,本文使用品种内的差异性(种内差异)和品种间的差异性(种间差异)对近红外光谱定性分析效果进行评价。

1.2 计算方法

经过前人的不断探索,目前在光谱差异性度量计算时多使用相关性、欧氏距离和熵 3 种计算方法^[13]。

1.2.1 相关性分析法

相关系数是由英国统计学家皮尔·科尔逊提出的,其数学表达式为

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

式中 x_i ——样本 x 第 i 个属性

y_i ——样本 y 第 i 个属性

\bar{x} ——样本 x 属性均值

\bar{y} ——样本 y 属性均值

(1) 种内相关性

假设每一类样品共采集 m 次信息,即进行 m 次独立实验,对每两次所采集的数据分别进行相关性分析,求得相关系数,然后取其平均值作为该类样品种内相关系数,即

$$R_{zn} = \frac{\sum_{i=1}^{c_m^2} |r_i|}{c_m^2} \quad (2)$$

式中 r_i ——两样本间相关系数

c_m^2 ——样本数量平方

(2) 种间相关性

假设共有 n 类样品,首先求得每类样品的平均数据,然后进行两两相关性分析,求得相关系数,取其平均值作为种间相关系数,即

$$R_{jn} = \frac{\sum_{i=1}^{c_n^2} |r_i|}{c_n^2} \quad (3)$$

式中 c_n^2 ——类别数量的平方

本文使用差异性对数据之间的差异进行评价。

1.2.2 欧氏距离分析法

欧氏距离是一种反映样品间亲疏程度的有效指标。欧氏距离越大,样品间差异性越大,反之则越

小^[14],其数学表达式为

$$d_{ij} = \sqrt{\sum_{p=1}^n (x_{ip} - x_{jp})^2} \quad (4)$$

式中 x_{ip} ——样本 i 的第 p 个属性

x_{jp} ——样本 j 的第 p 个属性

(1) 种内欧氏距离

假设每一类样品共采集 m 次信息,即进行 m 次独立实验,对每两次所采集的数据分别求得欧氏距离,然后取其平均值作为该类样品种内欧氏距离,即

$$D_{zn} = \frac{\sum_{i=1}^{c_n^2} |d_i|}{c_n^2} \quad (5)$$

式中 d_i ——两样本间距离

(2) 种间欧氏距离

假设共有 n 类样品,首先求得每类样品的平均数据,然后计算两两之间的欧氏距离,取其平均值作为种间欧氏距离,即

$$D_{sj} = \frac{\sum_{i=1}^{c_n^2} |d_i|}{c_n^2} \quad (6)$$

1.2.3 熵分析法

信息熵的概念是 1948 年由美国数学家 SHANNON 提出的,是信息论中用于度量信息量的一个概念。熵是系统稳定性评价的重要指标。熵越大,无序性越大^[15],即数据差异性越大。其数学表达式为

$$H = E[-\log_2 p_i] = -\sum_{i=1}^n p_i \log_2 p_i \quad (7)$$

式中 p_i ——样本 i 的概率

类内熵为同一品种多条光谱数据间的熵,类间熵则是多个品种平均光谱间的熵。

本文将使用相关系数计算得到的差异性、欧氏距离计算得到的距离、熵计算得到的熵都做差异性,相对距离为种间差异与种内差异的比值,相对距离越大,可分性越好。

1.3 仪器设备与样品

使用 3 组比较典型的玉米种子近红外光谱数据进行实验(部分原始光谱如图 1 所示)。为避免偶然性,增加实验结果的可信度,首先使用两组数据对上述评价指标的计算方法进行挑选,下文将其简称为实验数据 A 和实验数据 B。最后使用实验数据 C 对评价指标优化模型效果进行验证。数据分析软件为 Matlab R2014a。

(1) 实验数据 A

使用美国 JDSU 公司的 MicroNIR 1700 型近红外微型光谱仪采集单籽粒种子的透射光谱。谱区范

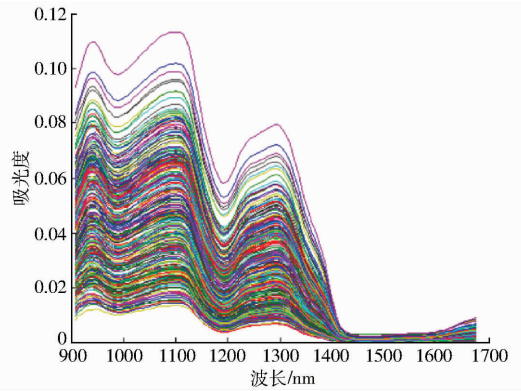


图 1 玉米种子原始光谱

Fig. 1 Original spectrum of maize seed

围为 908.1 ~ 1 677.2 nm,相邻波段间隔 6.194 4 nm,共 125 个波长点。使用外置近红外光源,玉米籽粒放置在转盘小孔上,光源从上方照射玉米籽粒,穿透籽粒后进入光谱仪检测器,检测结束后转盘转动进行下一次检测。小孔直径 5.2 mm,从进样到检测结束一次时间约为 3 s,每隔 20 min 做一次参比。

实验使用产自北京的 6 个品种的玉米种子,收获年份为 2015 年,平均每个品种采集 100 条光谱。

(2) 实验数据 B

使用德国 Bruker 公司的 MPA 型傅里叶变换近红外光谱仪采集单籽粒漫反射光谱。光谱波数范围 4 000 ~ 12 000 cm^{-1} ,分辨率 16 cm^{-1} (共 1 037 个数据点),扫描次数为 75 次(本实验使用到的光谱波数范围是 4 000 ~ 12 000 cm^{-1} ,相当于波长范围 1 111 ~ 2 500 nm,共 649 个数据点)。单籽粒样品测量附件为 $\Phi 22$ mm 小型样品池。使用仪器配套的 OPUS 6.5 软件对光谱进行存储和格式转换。

实验使用产自海南的 6 个品种的玉米种子,收获年份为 2013 年,每个品种采集 75 条光谱。

(2) 实验数据 C

该数据与实验数据 B 为同一仪器测量所得,6 个品种的玉米种子均产自河南省浚县,收获年份为 2015 年,平均每个品种采集 100 条光谱。

1.4 数据处理方法

根据以往的数据处理经验,并参考他人在玉米品种分类时使用的方法^[16-18],本次实验中每组实验数据采取的处理方法如表 1 所示(A、B 两组数据的数据处理方法使用识别率最高)。

2 结果与讨论

2.1 评价结果统计与分析

2.1.1 评价指标计算结果

使用上述 3 种计算方法得到的分级指标对 A、B 两组数据的处理效果进行评价。

表 1 各组实验数据的处理方法

Tab. 1 Processing methods of the experimental data sets

	预处理方法	优化的预处理方法	特征提取方法	优化的特征处理方法	识别方法
试验数据 A	数值中心化 + db 小波		PCA(前 16 维) + LDA(前 6 维)		BPRI
试验数据 B	数值中心化 + db 小波		PCA(前 18 维) + LDA(前 6 维)		BPRI
试验数据 C	平滑(窗口数 9) + 一阶导数	平滑(窗口数 9) + 一阶 导数 + 矢量归一化	PCA(前 20 维)	PCA(前 20 维) + LDA(前 6 维)	BPRI

注: PCA 是主成分分析; LDA 是线性判别分析; 识别算法 BPRI 是基于 2002 年王守觉院士提出的仿生模式识别(BPR)^[19], 结合最近邻方法得到的改进仿生模式识别算法^[20]。

由表 2、3 可以看出, 对 A、B 两组数据使用 3 种方法计算得到的评价结果具有一致性。使用相关性分析法计算得到的结果显示, 原始光谱数据预处理后的种间相对距离明显增大, 说明预处理的效果较好, 而再经过特征提取后相对距离提高不明显, 说明特征提取效果一般; 使用欧氏距离计算得到的结果显示, 预处理后的种间相对距离略有变化, 但特征提取后明显增大, 可见预处理的效果一般, 而特征提取的效果明显; 使用熵计算得到的结果显示, 经过预处理和特征提取 2 个步骤的数据处理都没有对分析性能起到改善作用。

表 2 实验数据 A 分级指标评价结果

Tab. 2 Evaluation results of A based on the grading index

	原始光谱	预处理	特征提取
相关性相对距离	0.622 8	2.076 0	2.262 6
欧氏相对距离	0.794 7	0.793 4	2.003 9
熵相对距离	0.978 2	0.798 9	0.940 6

表 3 实验数据 B 分级指标评价结果

Tab. 3 Evaluation results of B based on the grading index

	原始光谱	预处理	特征提取
相关性相对距离	0.937 2	1.580 0	1.935 1
欧氏相对距离	1.187 7	1.727 0	3.434 6
熵相对距离	1.002 0	0.681 2	0.827 1

2.1.2 识别率与评价指标的对照分析

正确识别率是一种已经广泛应用于对模型性能进行评价的指标。在得到了 3 种计算方法对 A、B 两组实验数据分别进行评价的结果后, 对照每个数据处理步骤所得到的正确识别率(表 4)对评价效果进行对比分析。为了使对照结果量化, 计算相对距离与正确识别率的相关性(表 5)。

表 4 A、B 两组数据正确识别率统计结果

Tab. 4 Statistics of two sets of data's correct acceptance rate

	原始光谱	预处理	特征提取
实验数据 A	31.67	31.27	90.52
实验数据 B	55.33	56.67	91.33

从表 5 可以看出, 对于实验数据 A, 使用欧氏距

表 5 评价结果与正确识别率的相关性统计结果

Tab. 5 Statistics of correlations between evaluation results and correct acceptance rates

	相关性	欧氏距离	熵
实验数据 A	0.582 5	1.000 0	0.323 4
实验数据 B	0.792 6	0.980 2	-0.084 9

离进行计算其评价结果与识别效果完全契合, 而使用相关性进行计算时评价效果较差, 熵最差; 对于实验数据 B, 使用欧氏距离进行计算得到的评价结果与模型识别率契合度同样最高, 相关性次之, 而使用熵值进行计算得到的评价结果与识别率相关性非常低, 且为负相关。所以选定欧氏距离作为种间相对距离的计算方法, 进行接下来的验证实验。

2.2 评价指标在优化模型算法中的应用

上述实验中统计的结果是经过本文提出的分级评价指标优化后的效果, 为验证该指标的有效性, 使用另外一组实验数据 C 进行实验。对实验数据 C 中的 6 个品种玉米近红外光谱进行定性分析, 使用欧氏距离方法计算数据处理每一过程的种间相对距离, 结果统计如表 6 所示。

表 6 评价结果与正确识别率的统计结果

Tab. 6 Statistics of evaluation results and correct acceptance rates

	原始光谱	原预处理	优化的预处理	原特征提取	优化特征提取
相对距离	0.661 9	0.658 2	1.297 2	1.310 2	2.491 0
正确识别率/%	33.89	40.86	70.08	68.32	93.27

由表 6 前两列可以看出, 原预处理方法并没有增大种间相对距离, 正确识别率没有得到有效提高, 经过对预处理算法进行调整, 种间相对距离增大了 0.96 倍, 正确识别率提高了 1.07 倍; 同理, 经过优化特征提取算法处理, 使得相对距离增加至原特征提取的 1.9 倍, 正确识别率提高了 24.95 个百分点。

由于数据的局限性, 本文只验证了分级评价指标对玉米种子近红外光谱定性分析模型优化的有效性。后续研究中可以探索其对其他近红外光谱数据的效果。

3 结论

(1)研究了近红外光谱玉米品种定性分析的性能分级评价指标,使用种间的相对距离进行评价,并与已经广泛应用的正确识别率进行对照分析。实验结果显示,使用欧氏距离法计算得到的相对距离与

正确识别率的评价结果相关性最高,相关系数均达到0.9以上。

(2)使用上述评价指标对近红外光谱玉米品种定性分析的性能进行优化,通过相对距离给出的结果进行优化,提高了正确识别率,验证了其有效性。

参 考 文 献

- 1 申艳,张晓平,梁爱珍,等.近红外光谱分析法测定东北黑土有机碳和全氮含量[J].应用生态学报,2010,21(1):109-114. SHEN Yan, ZHANG Xiaoping, LIANG Aizhen, et al. Determination of organic carbon and total nitrogen in black soil of Northeast China by near infrared spectroscopy[J]. Journal of Applied Ecology, 2010, 21(1): 109-114. (in Chinese)
- 2 邓云,陈志燕,曾德芬.近红外快速定量分析烟草化学成分模型的建立[J].广西烟草,2008(2):36-38. DENG Yun, CHEN Zhiyan, ZENG Defen. Establishment of rapid quantitative analysis model of tobacco chemical components by near infrared spectroscopy [J]. Guangxi Tobacco, 2008(2): 36-38. (in Chinese)
- 3 陈美丽,张俊,龚淑英,等.茉莉花茶主要品质成分定量近红外光谱分析模型的建立[J].茶叶科学,2013,33(1):21-26. CHEN Meili, ZHANG Jun, GONG Shuying, et al. Establishment of quantitative near infrared spectroscopy analysis model for main quality components of jasmine tea [J]. Tea Science, 2013, 33(1): 21-26. (in Chinese)
- 4 郝勇,陈斌.茶叶中低含量氨基酸近红外光谱定量分析模型研究[J/OL].农业机械学报,2014,45(6):216-220. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20140633&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2014.06.033. HAO Yong, CHEN Bin. Quantitative determination of low amino acid contents in tea by using near-infrared spectroscopy[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(6): 216-220. (in Chinese)
- 5 闫润,王新忠,邱白晶,等.基于特征光谱的草莓品种快速鉴别[J/OL].农业机械学报,2013,44(9):182-186. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20130932&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2013.09.032. YAN Run, WANG Xinzong, QIU Baijing, et al. Discrimination of strawberries varieties based on characteristic spectrum [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2013, 44(9): 182-186. (in Chinese)
- 6 严衍禄,赵龙莲,韩东海,等.近红外光谱分析基础与应用[M].北京:中国轻工业出版社,2005.
- 7 AMBROSE A, LOHUMI S, LEE W H, et al. Comparative nondestructive measurement of corn seed viability using Fourier transform near-infrared (FT-NIR) and Raman spectroscopy[J]. Sensors and Actuators, 2016, B224(3): 500-506.
- 8 GHASEMI-VARNAMKHAZI M, FORINA M. Nir spectroscopy coupled with multivariate computational tools for qualitative characterization of the aging of beer[J]. Computers and Electronics in Agriculture, 2014, 100(1): 34-40.
- 9 李天听,贾仕强,刘旭,等.应用近红外反射和透射光谱法鉴定玉米杂交种的纯度[J].光谱学与光谱分析,2015,35(12):3388-3392. LI Tianxi, JIA Shiqiang, LIU Xu, et al. Identification of purity of maize hybrids by near infrared reflectance and transmission spectroscopy [J]. Spectroscopy and Spectral Analysis, 2015, 35 (12): 3388-3392. (in Chinese)
- 10 JIA Shiqiang, AN Dong, LIU Zhe, et al. Variety identification method of coated maize seeds based on near-infrared spectroscopy and chemometrics[J]. Journal of Cereal Science, 2015, 63: 21-26.
- 11 张浚哲,朱文泉,董燕生,等.一种基于变权重组合的光谱相似性测度[J].测绘学报,2013,42(3):418-424,432. ZHANG Junzhe, ZHU Wenquan, DONG Yansheng, et al. A spectral similarity measure based on variable weight combination [J]. Journal of Surveying and Mapping, 2013, 42(3): 418-424, 432. (in Chinese)
- 12 王力宾,顾光同.多元统计分析:模型、案例及PASS应用[M].北京:北京经济科学出版社,2010:135-137. WANG Libin, GU Guangtong. Multivariate statistical analysis: models, cases and PASS applications [M]. Beijing: Economic Science Press, 2010: 135-137. (in Chinese)
- 13 禾子文.关于熵的统计讨论[J].涪陵师专学报,1997,13(3):97-104. HE Ziwen. Statistical discussion on entropy [J]. Journal of Fuling Teachers College, 1997, 13(3): 97-104. (in Chinese)
- 14 邬文锦,王红武,陈绍江,等.基于近红外光谱的商品玉米品种快速鉴别方法[J].光谱学与光谱分析,2010,30(3):1248-1251. WU Wenjin, WANG Hongwu, CHEN Shaojiang, et al. Rapid identification of commercial corn varieties based on near infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2010, 30 (3): 1248-1251. (in Chinese)
- 15 郭婷婷,邬文锦,苏谦,等.近红外玉米品种鉴别系统预处理和波长选择方法[J].农业机械学报,2009,40(增刊1):87-92. GUO Tingting, WU Wenjin, SU Qian, et al. Effects of spectral pretreatment and wavelength selection on discrimination of maize seed varieties by NIR spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(Supp. 1): 87-92. (in Chinese)
- 16 褚小立,袁洪福,陆婉珍.近红外分析中光谱预处理及波长选用方法进展与应用[J].化学进展,2004,16(4):528-542. CHU Xiaoli, YUAN Hongfu, LU Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004, 16(4): 528-542. (in Chinese)
- 17 王守觉.仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用[J].电子学报,2002,30(10):1417-1420. WANG Shoujue. Biomimetic pattern recognition (topological pattern recognition)—theory and application of a new pattern recognition model [J]. Acta, 2002, 30 (10): 1417-1420. (in Chinese)
- 18 靳召晰,张秀娟,罗付义,等.近红外光谱建模样本选择方法研究[J].光谱学与光谱分析,2016,36(12):3920-3925. JIN Zhaoxi, ZHANG Xiujuan, LUO Fuyi, et al. Sample selection method for near infrared spectroscopy modeling [J]. Spectroscopy and Spectral Analysis, 2016, 36 (12): 3920-3925. (in Chinese)