引用格式:高云飞,付霖宇,瞿军,等. 基于相似性度量的改进KS算法对近红外光谱分析模型的影响研究[J]. 电光与控制,2019,26(6):18-21,26. GAO Y F, FU L Y, QU J, et al. Influence of similarity measure based improved KS algorithm on near-infrared spectroscopy analysis model[J]. Electronics Optics & Control, 2019, 26(6):18-21, 26.

基于相似性度量的改进KS算法对近红外光谱分析模型的影响研究

高云飞[1], 付霖宇[1], 瞿 军[1], 王菊香[1], 邢志娜[1], 翁新华[2]
(1. 海军航空大学,山东 烟台 264001; 2. 中国人民解放军91515部队,海南 三亚 572061)

摘 要: 研究近红外光谱分析模型中的样本有效划分问题,针对经典KS算法依据距离度量描述高维度光谱数据间差异时效果不尽人意甚至失去意义的问题,结合目前相似性度量方法的不足,构造出一种新的相似性度量函数,采用光谱特征和性质特征相结合的方式计算样本间差异,提出一种改进的KS算法以寻求样本差异的最佳表达方式。通过与其他改进方法的对比,从有效性和对近红外光谱分析模型的影响两方面对所提改进算法进行分析,验证了所提算法的合理性和优越性。

关键词: 近红外光谱分析; 相似性度量; 模型传递; 多元校正模型; KS算法; 样本划分
中图分类号: O213.2 **文献标志码**: A **doi**:10.3969/j.issn.1671-637X.2019.06.004

Influence of Similarity Measure Based Improved KS Algorithm on Near-Infrared Spectroscopy Analysis Model

GAO Yun-fei[1], FU Lin-yu[1], QU Jun[1], WANG Ju-xiang[1], XING Zhi-na[1], WENG Xin-hua[2]
(1. Naval Aviation University, Yantai 264001, China; 2. No. 91515 Unit of PLA, Sanya 572061, China)

Abstract: The effective sample partition in the near-infrared spectroscopy model is studied. When classical Kennard Stone (KS) algorithm uses the distance metric to describe the difference between high-dimensional spectral data, the effect is unsatisfactory or even meaningless. To solve the problem, and considering the shortcomings of current similarity measurement methods, we constructed a new similarity measure function. The spectral features and property features were combined to calculate the difference between the samples. An improved KS algorithm was thus proposed to find the best expression of sample difference. The improved algorithm was analyzed from the aspects of effectiveness and the impact on the near-infrared spectroscopy model by comparing with other improved methods, and the rationality and superiority of the proposed algorithm were verified.

Key words: near-infrared spectroscopy analysis; similarity measure; model transfer; multivariate correction model; KS algorithm; sample partition

0 引言

近红外光谱分析技术借助快速高效、成本低廉、破坏性小等突出优势逐步取代了传统分析方法,被广泛应用于定量分析、状态评估和在线监测等领域。其中,样本划分是一个共性的基础问题,与分析模型的性能密切相关:建立多元校正模型时涉及到校正集和验证集样本的划分,文献[1-2]提及GALVAO等研究发现校正集的选择直接影响到近红外光谱定量分析模型的预测性能。同时,转换集的划分选取也是近红外光谱分析模型传递过程中的关键部分,李华等的研究表明,转换集的合理选择能够简化模型的传递过程,提高传递精度[3]。

KS算法是一种有效的、普遍采用的样本划分方法,已有研究证明[4]其在选择样本时具备快速高效、简单直观、代表性强的突出优势,因此被广泛应用于机器学习、聚类分析、分类决策等问题。但是,KS算法因本

收稿日期:2018-07-27 **修回日期**:2018-08-24
基金项目:国家自然科学基金(51605487); 山东省自然科学基金(ZR2016FQ03)
作者简介:高云飞(1993—),男,内蒙古呼和浩特人,硕士,研究方向为近红外光谱和油液分析。

身借助计算欧氏距离实现样本划分常常被学者们诟病[5]，王菊香等提出的用马氏距离代替欧氏距离的改进方法[6]虽然弥补了欧氏距离的固有缺陷，但是仍旧存在夸大微小变量作用和结果不稳定的问题。此外，在近红外光谱模型建立和传递过程中应用 KS 算法进行样本划分时，算法中的传统距离度量方法描述高维度光谱数据差异性还难以达到预期效果[7]。针对以上问题，本文构造了一种新的相似性度量函数，在考虑光谱特征的同时结合性质特征描述高维度光谱数据的差异性，进而提出了一种改进的 KS 算法，以期实现样本的有效划分。通过与其他改进方法的对比，分别从有效性和对近红外光谱分析模型建立与传递过程的影响角度对所提改进算法进行分析，验证了改进 KS 算法的合理性和优越性。

1　KS 算法及距离度量特性分析

经典 KS 算法以两两样本之间的欧氏距离为依据实现代表性强、分布范围均匀的样本划分选择。欧氏距离虽然简单直观，但它将样本属性之间的差别同等对待，而且忽略属性间的联系，有时无法满足实际需要；没有考虑到量纲，容易导致“大数吃小数”的问题。学者们提出用“尺度无关”的马氏距离代替欧氏距离，在排除相关性干扰的同时弥补了欧氏距离的不足。但是马氏距离中协方差矩阵由样本计算得到，稳定性不佳，容易夸大微小变量的作用，而且当总体样本数少于样本维数时协方差矩阵的逆矩阵不存在，马氏距离失效。

通过近红外光谱技术得到的光谱波数范围往往比较广泛，数据维度较大，考虑到试验成本问题，样本数目相比于波数数量更少，很难满足马氏距离的存在条件。此外，高维度数据空间由于变量过多，常常引发诸如“维数灾难”的一系列问题[8]，采用传统距离度量方法描述高维度数据的差异性时，其效果不尽人意甚至毫无意义。一方面是因为高维度数据空间随着维度的增加变得更加稀疏，最近邻点与最远邻点之间的对比被削弱，如果仍旧采用传统距离度量方法，计算出的两两样本间的距离几乎相等[9]；另一方面，随着维度的提升，与样本差异性无关的噪声增加，会淹没大量有用信息。

2　基于相似性度量的改进 KS 算法

2.1　目前相似性度量函数

针对传统距离度量方法应用于 KS 算法时效果大打折扣甚至失效的问题，考虑采用相似性度量函数，其选取直接影响 KS 算法的分类效果。目前较为流行的相似性度量函数[10]为

$$Hsim(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{1+|x_i-y_i|} \tag{1}$$

$$Close(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{d}\sum_{i=1}^{d}\mathrm{e}^{-|x_i-y_i|}\text{。} \tag{2}$$

$Close(\boldsymbol{X},\boldsymbol{Y})$ 函数利用 e^{-x} 单调递减的特性，相较于 $Hsim(\boldsymbol{X},\boldsymbol{Y})$ 函数更具优势，但是二者都只考虑了同一维度下的绝对差值，存在数据数量级对相似度影响较大的问题。

2.2　改进相似性度量函数

针对目前相似性度量函数存在的不足，构造改进的相似性度量函数为

$$Isim(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{d}\omega_i\mathrm{e}^{\frac{-|x_i-y_i|}{\frac{1}{2}|x_i+y_i|+1}} \tag{3}$$

式中：$\boldsymbol{X}=(x_1,x_2,\cdots,x_d)$ 和 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_d)$ 为数据向量，d 为维度；ω_i 是每个维度对应的权重。$Isim(\boldsymbol{X},\boldsymbol{Y})$ 具有以下特点：

1）利用指数函数 e^{-x} 比幂函数 $\frac{1}{1+x}$ 在[0,1]区间下降速度更快、曲线更加陡峭的特点，可以提高同一维度下的区分度；

2）e^{-x} 指数部分加入 $\frac{|x_i+y_i|}{2}$ 分量，在保留同一维度下绝对差值的同时，加入相对差值因素，对高值数据的相似性和低值数据的相似性加以区分，使得同一维度下数据相同或者相近时，削弱数量级的影响；

3）添加维度权重因素，根据实际情况定夺不同维度对整体相似度的贡献程度，具有一定的伸缩性和灵活性。

2.3　改进 KS 算法

目前在近红外光谱分析模型中应用 KS 算法时，往往只考虑样本间的光谱特征，忽略性质特征。因此，本文将性质特征与光谱特征结合，提出一种改进的 KS 算法，将其用于选择模型传递转换集的步骤如下：

1）根据测得的光谱和理化指标数据分别构造光谱样本的光谱特征矩阵 $\boldsymbol{A}_g$ 和性质特征矩阵 $\boldsymbol{A}_x$；

2）按照式(3)分别计算特征矩阵 $\boldsymbol{A}_g$ 和 $\boldsymbol{A}_x$ 中两两光谱样本 $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_j$ 之间的相似性度量函数 $Isim(\boldsymbol{S}_i,\boldsymbol{S}_j)$，取倒数表征样本之间的距离 $d_{ij}=\frac{1}{Isim(\boldsymbol{S}_i,\boldsymbol{S}_j)}$，获得光谱特征距离矩阵和性质特征距离矩阵 $\boldsymbol{D}_g$ 和 $\boldsymbol{D}_x$，得到最终特征距离矩阵 $\boldsymbol{Z}=\boldsymbol{D}_g+\boldsymbol{D}_x$；

3）选择特征距离矩阵 $\boldsymbol{Z}$ 中最大元素对应的两个样本进入转换集；

4）记录剩余样本与被选样本之间的距离，挑选出最小距离，选取这些最小距离中最大距离所对应的样本进入转换集；

5）以此类推，直至满足转换集样本个数要求。

3 有效性及影响分析

3.1 有效性分析

通常采用下式给出的衡量标准判别距离度量函数的有效性，即

$$v = \frac{d_{max} - d_{min}}{d_{min}} \tag{4}$$

式中，d_{max}和 d_{min}分别为数据对象之间的最大和最小距离。传统距离度量方法中，随着数据维度数的增加，会出现样本之间的距离近似相等，最大和最小距离之差趋近于 0 的情况，即比值 v 趋于 0，度量方法失效。为了验证本文构造改进相似性度量函数 $Isim(X,Y)$ 的有效性，通过 Matlab 的 *normrnd* 函数随机生成 1000 个维度数分别为 10，20，50，80，100，150，200，500 和 1000 的样本数据，得到比值 v 随维度数的变化趋势，如图 1 所示。

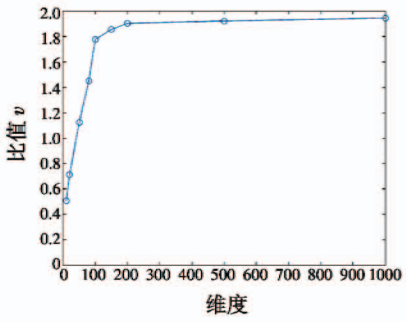

图 1 比值 v 随维度变化趋势

Fig. 1 Change of ratio v with the dimension

可以看出，随着维度数的增加，比值 v 逐渐增加直至趋于稳定，没有出现趋近于 0 的情况，这说明改进的相似性度量函数在区分最近邻点与最远邻点时适应性较好，在一定程度上缓解了“维数灾难”。

为了说明改进相似性度量函数 $Isim(X,Y)$ 的优越性，选取两组数据进行分析，与 $Hsim(X,Y)$ 和 $Close(X,Y)$ 函数的对比结果如表 1 所示。

表 1 *Isim* 函数与 *Hsim*，*Close* 函数对比结果

Table 1 Comparison of *Isim* function with *Hsim* function and *Close* function

		维 1	维 2	*Hsim*	*Close*	*Isim*
数据对象	*A*	2	603	0.200 0	0.018 3	0.721 4
	B	6	607			

数据对象 $A(2,603)$ 和 $B(6,607)$ 在同一维度下的差值均为 4，虽然相比于维 1、维 2 的值大了 2 个数量级，但是直观上看，A 和 B 相似度较高。根据 $Hsim(X,Y)$ 和 $Close(X,Y)$ 函数计算出的相似度结果明显偏低，与实际偏差过大，适用性较差。利用 $Isim(X,Y)$ 函数计算出的相似度更擅于表达数据对象间的接近程度。

数据对象分布如表 2 所示。

表 2 数据对象分布

Table 2 Distribution of data objects

		维 1	维 2
数据对象	*C*	6	606
	D	1002	606
	E	1006	606

对于表 2 中的数据对象，若采用欧氏距离描述差异性，则会得到数据对象 *CD* 之间和 *ED* 之间的距离分别为 996 和 4，相差 2 个数量级，受尺度影响产生的误差较大；若采用马氏距离，得到的协方差矩阵行列式为 0，逆矩阵不存在，无法利用马氏距离公式计算。

3.2 改进 KS 算法对模型建立的影响

为了考察本文设计的改进 KS 算法对近红外光谱多元校正模型建模过程和精度的影响，以某导弹在用航空煤油为研究对象，由于其性质指标数目较多，这里只以密度指标为例。结合按照标准方法测定的 68 个样品的密度值和采集的 2002 个近红外光谱数据点，分别采用经典 KS 算法、文献[7]中的改进 KS 算法和本文设计的改进 KS 算法划分校正集和验证集，进而建立偏最小二乘(PLS)模型进行预测分析。对模型进行评价时，依据的指标为最佳主成分数(F)、校正标准偏差(SEC)、预测标准偏差(SEP)和相关系数(R^2)，对比结果如表 3 所示。

表 3 不同样本划分方法对比

Table 3 Comparison of different sample division methods

	F	SEC	SEP	R^2
经典 KS 算法	8	0.001 40	0.001 70	0.972 6
文献[7]的方法	8	0.000 91	0.000 95	0.991 4
本文方法	6	0.000 70	0.000 72	0.991 7

总体上看，应用 3 种 KS 算法进行样本划分的多元校正模型所表现出的预测性能均令人满意。其中，本文的改进 KS 算法的主成分数更少、偏差更小、相关系数更高，预测结果如图 2 所示。这说明所建立校正模型的预测精度得到提高，建模过程实现了简化，从而验证了本文方法的优越性和合理性。

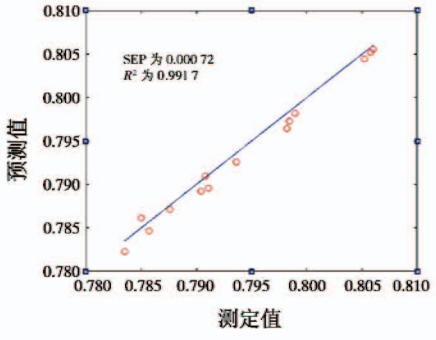

图 2 预测值和测定值相关图

Fig. 2 Correlation graph of predicted and measured data

3.3　改进 KS 算法对模型传递的影响

转换集的确定是近红外光谱有标模型传递的一个基础性问题，为了说明不同样本划分方法对模型传递的影响，利用 3.2 节中主仪器建立的模型对从仪器测得的光谱进行预测，结果如图 3a 所示，SEP 为 0.0023，R^2 为 0.9021，说明直接将主仪器模型应用于从仪器光谱时，仪器差别导致响应函数发生变化，产生较大的预测偏差。

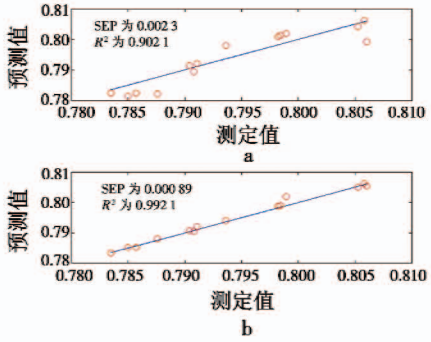

图 3　从仪器光谱在主仪器模型上的预测值和测定值相关图

Fig. 3　Correlation graph of predicted and measured data of slave instrument spectrum using master instrument model

采用分段直接校正(PDS)法校正主从仪器变化带来的误差，其中转换集样本的个数根据 SEP 最小原则确定。考虑到 PDS 是一种局部性质的传递方法，所需转换集个数较少，因而设置转换集个数变化范围为 4～12，分别采用经典 KS 算法、文献[7]中的改进 KS 算法和本文设计的改进 KS 算法时，SEP 随转换集个数的变化趋势如图 4 所示，其中 PDS 窗口宽度同样根据 SEP 最小原则寻优确定。

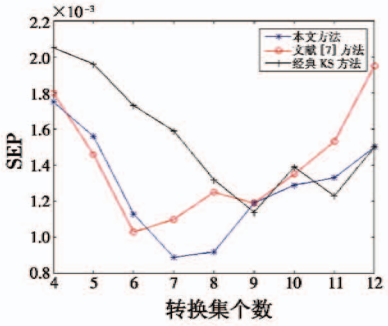

图 4　SEP 随转换集个数变化趋势

Fig. 4　Change of SEP with number of transfer set

从整体上看，随着转换集个数的增加，转换矩阵包含的有用信息更充分，因而 SEP 降低；当转换集个数增加至恰好消除仪器间差异时，SEP 开始逐渐下降。此外，相比于图 3a 中不进行传递的预测结果，3 种划分方法下的 SEP 均始终处于较低水平，说明 KS 算法选取的转换集代表性较好，能够有效消除仪器间差异，实现仪器间的共享。

从模型传递过程上看，本文改进 KS 算法所需的转换集样本最少，传递过程最简洁；从传递精度上看，本文方法的 SEP 最小，当转换集个数为 7 时，SEP 为 0.00089，R^2 为 0.9921，如图 3b 所示，PDS 最佳窗口宽度为 15。说明本文方法显著降低了模型传递过程的复杂度，提高了传递精度。

4　结束语

本文针对传统 KS 算法无法满足高维度光谱数据样本划分的实际需求问题，在深入分析距离度量特性的基础上，结合当前流行的相似性度量方法，构造了一种新的相似性度量函数，进而提出改进的 KS 算法。有效性分析结果表明，本文提出的改进算法能够更好地区分最近邻点与最远邻点，在一定程度上缓解了“维数灾难”引发的“距离度量失效”，并且克服了经典 KS 算法无法描述高维度数据的问题；某导弹在用航空煤油的实例分析表明，本文提出的改进算法不仅能够实现样本的有效划分，而且相较于其他改进方法，应用本文算法划分样本的近红外光谱分析模型的复杂度更低，精度更高。主要结论和创新点有：

1）重构相似性度量函数时，利用指数函数提高同一维度下的区分度，并加入相对差值和维度权重因素，削弱数量级影响的同时提高函数的灵活性和适应性；

2）计算样本间差异时，在考虑样本光谱特征的同时结合性质特征，包含的信息更加充分、全面，有利于选取更有代表性的样本；

3）对改进算法进行验证时，通过与其他改进方法的对比，在有效性分析的基础上，以精度和复杂度为切入点，对近红外光谱模型的建立和传递过程进行详细分析，结论更具说服力。

参 考 文 献

[1] PENG Y F, LUO H P, LUO X N, et al. SPXY sample classification method and successive projections algorithm combined with near-infrared spectroscopy for the determination of total sugar content of southern Xinjiang jujube [J]. Advanced Materials Research, 2014, 1030/1031/1032: 352-356.

[2] 詹雪艳，赵娜，林兆洲，等. 校正集选择方法对于积雪草总苷中积雪草苷 NIR 定量模型的影响[J]. 光谱学与光谱分析，2014, 34(12):3267-3272.

[3] 李华，王菊香，邢志娜，等. 改进的 KS 算法对近红外光谱模型传递影响的研究[J]. 光谱学与光谱分析，2011, 31(2):362-365.

(下转第 26 页)

5 结论

基于雷达目标一维距离像的目标识别方法在雷达目标识别领域中占据着重要的地位，本文以LVQ神经网络作为目标识别的分类器，同时分析比较了主成分分析、核主成分分析以及粒子群优化的核主成分分析用于特征提取。经实验表明：由粒子群算法优化后的核主成分分析应用于特征提取，首先克服了原核主成分分析方法中依靠经验来确定未知参数的缺点，能够准确地得到最优的参数；其次，大大降低了数据的复杂程度，减小了计算量，又确保具有较高的识别率。

参 考 文 献

[1] 刘宏伟，杜兰，袁莉，等. 雷达高分辨距离像目标识别研究进展[J]. 电子与信息学报，2005, 27(8)：1328-1334.

[2] 彭红星，潘梨莉，赵鸿图. 一种改进的KPCA传感器故障识别方法及其应用[J]. 仪表技术与传感器，2016(6)：92-94.

[3] 曾番，黄文龙，夏伟鹏，等. 小波包特征能量算子与多核函数组合KPCA的声目标识别[J]. 电光与控制，2017, 24(4)：5-7.

[4] 王力，周志杰，赵福均. 基于BP神经网络和证据理论的超声检测缺陷识别[J]. 电光与控制，2018, 25(1)：65-69.

[5] 宋娟，邹翔，尹俭芳，等. 基于神经网络的人脸朝向识别研究[J]. 工业控制计算机，2017, 30(4)：111-112.

[6] KOHONEN T. Self-organization and associative memory [M]. 3rd ed. Berlin：Springer, 1989.

[7] 夏飞，罗志疆，张浩，等. 混合神经网络在变压器故障诊断中的应用[J]. 电子测量与仪器学报，2017, 31(1)：118-124.

[8] MUNLIN M, ANANTATHANAVIT M. Hybrid radius particle swarm optimization [C]//Region 10 Conference, IEEE, 2017：1-5.

[9] 袁莉，刘宏伟，保铮. 基于中心矩特征的雷达HRRP自动目标识别[J]. 电子学报，2004, 32(12)：2078-2081.

[10] DU L, LIU H, BAO Z, et al. Radar automatic target recognition using complex high-resolution range profiles [J]. LET Radar, Sonar & Navigation, 2007, 1(1)：18-26.

(上接第21页)

[4] 潘国锋. 基于K-S算法的水质硝酸盐含量光谱检测方法研究[J]. 光谱实验室，2011, 28(5)：2700-2704.

[5] 梁晨. 近红外光谱多元校正模型传递方法的研究[D]. 北京：北京化工大学，2016.

[6] 王菊香，孟凡磊，刘林密，等. 样品选择结合分段直接校正法和偏最小二乘法用于近红外光谱分析模型传递研究[J]. 兵工学报，2016, 37(1)：91-96.

[7] LIANG C, ZHAO Z, CAO Y T, et al. A new study of calibration model transfer method for near-infrared spectral analysis[J]. Spectroscopy and Spectral Analysis, 2017, 37(5)：1587-1594.

[8] 王晓阳，张洪渊，沈良忠，等. 基于相似性度量的高维数据聚类算法研究[J]. 计算机技术与发展，2013, 23(5)：30-33.

[9] 陈海燕，刘晨晖，孙博. 时间序列数据挖掘的相似性度量综述[J]. 控制与决策，2017, 32(1)：1-10.

[10] 李海林，郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述[J]. 计算机应用研究，2013, 30(5)：1285-1291.