平均分布差异最小化的NIR标定迁移方法研究

赵煜辉，芦鹏程，罗昱博，单　鹏

东北大学秦皇岛分校，河北 秦皇岛　066000

摘　要　凭借高效、无损和环保的优点，近红外光谱在多个领域广泛用作物质快速分析方法的同时，仍面临着光谱标定模型生命周期短，构建仪器标定迁移方法的标准样品难以获得和保存等问题。在化学计量学文献中，迁移方法通常能够矫正主从仪器之间的光谱差异，但绝大多数方法都需要在两台仪器相同条件下测量一组迁移标准样品。虽然样品数目不必过多，但总体上表明，必须对其进行很好的选择才能保证成功迁移。对于在主从仪器中选择代表性的样本子集，现有Kennard-Stone算法作为样本选择的主要算法。在标准样本的确定问题中，假设主仪器已找到标准样本，选择的样本集需要在从仪器中进行测量，仅当迁移样本足够稳定时才有可能，但现有近红外光谱技术无法保证这一点。如果假设使用从仪器的样本作为标准样本，考虑到新工业应用中光谱光源的变更，主仪器被从仪器代替，因此不再可用。基于目前存在的这些问题，提出了一种平均分布差异最小化的NIR标定迁移方法(MCT)，此方法可以在不考虑从仪器标准样本(即标准样本自由)的情况下，针对近红外光谱数据的多重共线性，首先假设存在一个主从仪器光谱的共同偏最小二乘子空间，并将主从仪器光谱数据分别投影到该公共子空间；然后，引入平均分布差异最小化算法，即分别给出主从光谱数据在子空间的平均分布中心表示函数，在最小化两个光谱平均分布(中心点)的差异的同时，最大化投影后主仪器光谱的协方差，推导求解出最佳子空间；最后，将主光谱样本和从光谱预测样本分别投影到该偏最小二乘子空间中，利用主光谱数据得到回归模型，该模型可用于预测从光谱浓度。通过对玉米数据集和小麦数据集的测试研究，证明的预测效果与SBC，PDS，CCACT，TCR和MSC相比有所改善，该方法可以实现更低的预测误差。

关键词　近红外光谱；标定迁移；平均分布差异；标准样本自由；偏最小二乘回归

中图分类号：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)10-3051-07

引　言

近红外光谱(NIR)分析技术具备操作简单、分析数据速度快、成本较低、不污染样品等优势，已在各领域得到广泛应用，如农产品生产、化工产品生产、食品生产以及环境监测领域[1-4]。近红外光谱技术在定性分析和快速物质成分定量分析以及实现在线检测方面具有独特优势[5]。建立多元校正模型是近红外光谱分析技术的重要内容。即通过一定的数学分析方法，对近红外光谱数据进行分析建模，从而达到对一些指标进行预测的目的，这是一种根据已有样本总结出规律生成模型的方法。但实际的工业生产中，测量仪器、环境和场景通常并不一致，依据已有近红外光谱数据建立的模型往往并不适用新的仪器采集的数据，原有模型失效，并且在测量环境或其他条件变化后，也需更新模型。

标定迁移是指在不同测量仪器或测量状态下的多元标定模型迁移方法，通过将从光谱数据迁移到主光谱数据空间，进而实现主光谱数据模型对从光谱数据模型的预测，避免重复建模[6-7]。

已有标定迁移方法[8-9]，主要是通过一组标准样品构建迁移模型，它需要在主仪器和从仪器上分别测量一组标准样本，通过一组标准样本来纠正主仪器和从仪器之间光谱的差异。分段直接标准化(piecewise direct standardization，PDS)，主仪器的每个波长与从仪器的波长窗口相关，基于每个窗口间回归系数形成带状迁移矩阵。实验结果与假设是一致的，即在各种迁移方法中，主仪器和从仪器之间的频谱相关性被限制在较小的区域。PDS的关键是窗口大小的选择和标准样本数目的确定。在偏差斜率校正(slope and bias correction，SBC)[10]中，假设不同仪器的预测值之间存在线性关系，先计算光谱和响应值之间的回归系数；并用该系数分别计算主

收稿日期：2020-10-08，**修订日期**：2021-02-27

基金项目：国家自然科学基金青年科学基金项目(61601104)资助

作者简介：赵煜辉，1971年生，东北大学秦皇岛分校教授　　e-mail：1000272@neuq.edu.cn

仪器和从仪器的预测值；最后，在预测值之间进行线性拟合。SBC 算法为一种单变量方法，因此在测量仪器和测量条件变化引起系统化的光谱差异的情况下，才能取得较好的效果。现实生活中，光谱差异往往比较复杂，此时它的预测能力是不确定的。Liang 等提出了基于典型相关分析(canonical correlation analysis，CCA)的标定迁移方法成功地校正了不同光谱之间的差异。首先，使用主仪器的标定集构建 PLS 模型；选取主仪器和从仪器的标定集的一部分作为标准样本；通过典型相关分析分别提取特征[11]。

标准样本要求主从仪器在相同的环境及条件下测量同一组样本。工业应用中，由于标样组分的挥发性及可变性，使保持标准样品的完整性很难实现[12]，为此，需建立标准样本自由的标定迁移模型[13]。

Bouveresse 等提出的多元散射校正(multiplicative scatter/signal correction，MSC)[14]是一种信号预处理方法。MSC 计算校准集的平均光谱作为参考光谱，并在每个光谱和参考光谱之间找到线性关系，得到斜率和偏差，利用斜率和偏差来校正从光谱，虽然不需要标准样本，但难以处理复杂情况，且模型性能多数情况较差。

迁移成分回归(transfer component regression，TCR)也是一种无标准的迁移方法[13]，它结合了迁移成分分析(transfer component analysis，TCA)[15]和普通最小二乘法(ordinary least square，OLS)。TCA 的基本思想是在再生希尔伯特空间中投影两个仪器的数据，在这个空间中，主仪器和从仪器的数据分布尽可能的接近，同时保留原始数据的关键属性。TCR 是一个具有良好泛化能力的稳健模型，但无法实现更准确的预测。

针对标准样本难以获得和保存，现有的标准样本自由的标定迁移方法预测能力相对一般的情况，提出了一种标准样本自由的基于最小化平均分布差异的 NIR 偏最小二乘标定迁移方法(minimizing mean distribution discrepancy Calibration Transfer for NIR，MCT)。此方法在不考虑从仪器标准样本的情况下，为去光谱数据的多重共线性，首先假设存在一个适用于主从仪器的偏最小二乘子空间，该子空间通过后续优化主从仪器在此空间中的分布差异获得，接着将主从仪器光谱数据分别投影到该假设的公共子空间；然后引入平均分布差异最小化算法，即分别给出主从光谱数据在子空间的平均分布(中心点)表示函数，最小化两个光谱平均分布(中心点)的差异，并最大化投影后主仪器光谱的协方差，目的是使主仪器投影后的数据具有最大相关性，推导求解出最佳子空间；最后，将主光谱样本和从光谱预测样本分别投影到该子空间中，利用主光谱数据得到回归模型，通过此回归系数计算出从光谱预测浓度。该方法无需标准样本的获取，便能缩小主从仪器数据间的分布差异，同时对比现有标准自由迁移方法，更加简单高效，并具有更好的预测性能。本文使用玉米数据集和小麦数据集，将 MCT 的性能与 SBC，PDS，CCACT，TCR 和 MSC 进行比较。

1 理论知识

1.1 定义符号

源域和目标域将用下标“S”和“T”表示，$X_S=[X_S^1,\cdots,X_S^{N_S}]\in R^{D\times N_S}$ 表示源域训练集，$X_T=[X_T^1,\cdots,X_T^{N_T}]\in R^{D\times N_T}$ 表示目标域训练集。其中 D 表示域中数据的维数，N_S 和 N_T 表示源域和目标域样本个数。设 $P\in R^{D\times d}$ 表示将源数据和目标数据的原始空间映射到维数为 d 的子空间的基变换。$\|\cdot\|_2$ 表示 2 范数。$Tr(\cdot)$ 表示矩阵的迹运算符，$(\cdot)^{\mathrm{T}}$ 表示转置运算符。

1.2 偏最小二乘法

在化学计量学中，偏最小二乘算法(partial least square，PLS)是一种很有效的多元标定方法。PLS 算法结合了多元线性回归、主成分分析、典型相关分析的优点，被广泛用于建立输入空间和响应空间之间的关系。PLS 通过分数向量建立输入空间和响应空间之间的关系。PLS 模型的目的是确保最佳的潜变量数量。潜变量是原始变量的线性组合。它包含了关于 X 和 y 之间关系的最大相关信息。在数学上，由式(1)表示目标函数

$$H=\underset{w}{\operatorname{argmax}}\operatorname{cov}\langle Xw,\ y\rangle \quad \text{subject to } \|w\|_2=1 \tag{1}$$

其中 w 代表权重向量。该目标函数是在一个约束下的最大化问题，可以通过拉格朗日乘数法进行求解。

在这个算法中，第一个权重向量必须是矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}$ 的主要的特征向量。从第二个潜变量开始，它要求接下来的潜变量与前面的潜变量正交(不相关)。因此，接下来的权重向量也是矩阵的主要特征向量，重复这一系列步骤直到收敛。模型被构建通过如下等式

$$\begin{cases}\boldsymbol{X}^{N\times D}=\boldsymbol{T}^{N\times A}(\boldsymbol{P}^{D\times A})+\boldsymbol{E}^{N\times D}\\ \boldsymbol{y}^{N\times 1}=\boldsymbol{T}^{N\times A}(\boldsymbol{Q}^{1\times A})+\boldsymbol{F}^{N\times 1}\end{cases}$$

其中 $\boldsymbol{T}$ 是得分矩阵，$\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 分别代表 $\boldsymbol{X}$ 的载荷矩阵和 $\boldsymbol{y}$ 的载荷矩阵向量；$\boldsymbol{E}$ 和 $\boldsymbol{F}$ 分别表示残差矩阵；A 是 PLS 模型潜变量的最佳数量。

最后，模型的回归系数 β 可写如式(2)

$$\beta=\boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{Q}^{\mathrm{T}} \tag{2}$$

式(2)中，$\boldsymbol{W}=[w_1,w_2,\cdots\cdots,w_A]$为权重矩阵。

1.3 模型建立

该文目的是学习将源数据和目标数据的原始空间映射到某个 PLS 子空间的基变换，在该子空间中，映射的源数据和目标数据之间的得分 $T_S=[t_S^1,\cdots,t_S^{N_S}]\in R^{D\times N_S}$ 和 $T_T=[t_T^1,\cdots,t_T^{N_T}]\in R^{D\times N_T}$ 可以保持相似，因此，有理由认为，T_S 和 T_T 之间的平均分布差异(mean distribution discrepancy，MDD)是最小的。因此，源数据和目标数据在低维子空间中 MDD 最小化被表述为

$$\min\|\mu_S-\mu_T\|_2^2=\min\left\|\frac{1}{N_S}\sum_{i=1}^{N_S}t_S^i-\frac{1}{N_T}\sum_{j=1}^{N_T}t_T^j\right\|_2^2 \tag{3}$$

其中，$\mu_S=\dfrac{1}{N_S}\sum_{i=1}^{N_S}t_S^i$ 和 $\mu_T=\dfrac{1}{N_T}\sum_{j=1}^{N_T}t_T^j$ 表示源域与目标域变换到新的子空间的平均分布(中心点)。

令 $\boldsymbol{T}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}$，最小化问题式(3)可以重新表示为

$$\min_P\left\|\frac{1}{N_S}\sum_{i=1}^{N_S}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_S^i-\frac{1}{N_T}\sum_{j=1}^{N_T}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_T^j\right\|_2^2 \tag{4}$$

为了学习得到这样一个能使式(4)中的平均分布差异最小化的基变换矩阵 $\boldsymbol{P}$，还应确保投影后的源数据 X_S 与源数据浓度 y_S 之间的关系具有最大相关信息。因此，对于源域的数据，合理的做法是将以下项最大化

$$\max_{P} \| \operatorname{cov}(\boldsymbol{P}^{\mathrm{T}} X_S, y_S) \|_2^2 = \max_{P} Tr(\boldsymbol{P}^{\mathrm{T}} X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}} \boldsymbol{P}) \quad (5)$$

在求解式(5)时可以看出，源域数据的协方差在新学习的子空间中已经被最大化，那么在这一过程中就保留了尽可能多的可用信息。

结合式(4)和式(5)，可以得到以下优化目标

$$\max_{P} \frac{Tr(\boldsymbol{P}^{\mathrm{T}} X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}} \boldsymbol{P})}{\left\| \frac{1}{N_S} \sum_{i=1}^{N_S} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{x}_S^i - \frac{1}{N_T} \sum_{j=1}^{N_T} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{x}_T^j \right\|_2^2} \quad (6)$$

令 $\mu_S = \frac{1}{N_S} \sum_{i=1}^{N_S} x_S^i$ 和 $\mu_T = \frac{1}{N_T} \sum_{j=1}^{N_T} x_T^j$ 为源域和目标域投影后的中心，然后式(6)中的最大化问题就可以写成

$$\max_{P} \frac{Tr(P^{\mathrm{T}} X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}} P)}{\| P^{\mathrm{T}} \mu_S - P^{\mathrm{T}} \mu_T \|_2^2} = \max_{P} \frac{Tr(P^{\mathrm{T}} [X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}}] P)}{Tr(P^{\mathrm{T}} (\mu_S - \mu_T)(\mu_S - \mu_T)^{\mathrm{T}} P)} \quad (7)$$

在式(7)的最大化问题中，P 有许多的可能解(即并非唯一解)，为了保证解的唯一性，式(7)施加了一个等式约束，这样就可以写成

$$\max_{P} Tr(P^{\mathrm{T}} [X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}}] P)$$
$$\text{s. t. } Tr(P^{\mathrm{T}} (\mu_S - \mu_T)(\mu_S - \mu_T)^{\mathrm{T}} P) = \eta \quad (8)$$

其中 η 是一个常数。

为了解出式(8)，将其改为拉格朗日函数，见式(9)

$$L(P, \gamma) = Tr(P^{\mathrm{T}} [X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}}] P) - \gamma(Tr(P^{\mathrm{T}} (\mu_S - \mu_T)(\mu_S - \mu_T)^{\mathrm{T}} P) - \eta) \quad (9)$$

其中 γ 表示拉格朗日乘子系数。

接下来，将 $L(P, \gamma)$ 对 P 求偏导，令其偏导数为 0，就得到

$$\boldsymbol{A}\boldsymbol{P} = \gamma \boldsymbol{P} \quad (10)$$

其中 $\boldsymbol{A} = ((\mu_S - \mu_T)(\mu_S - \mu_T)^{T})^{-1}(X_S y_S y_S^{\mathrm{T}} X_S^{\mathrm{T}})$

由此得出，最优子空间 $\boldsymbol{P}^*$ 表示矩阵 $\boldsymbol{A}$ 特征值分解后的前 k 个最大特征值所对应的特征向量，而 γ 表示是一个对角矩阵，对角线上的值分别为前 k 个最大特征值。

为了便于实现，将所提出的 MCT 算法归纳到下列算法描述中。

1.4　MCT 的算法描述

输入：给定主仪器中心化后的数据集($\boldsymbol{X}_{\mathrm{cen}}^{\mathrm{S}}$，$\boldsymbol{y}_{\mathrm{cen}}^{\mathrm{S}}$)，从仪器中心化后的训练数据 $\boldsymbol{X}_{\mathrm{train\_cen}}^{\mathrm{T}}$ 和最大主成分数 pc_num。

输出：回归系数 β。

(1) 分别计算主从仪器数据($\boldsymbol{X}_{\mathrm{center}}^{\mathrm{S}}$，$\boldsymbol{X}_{\mathrm{train\_cen}}^{\mathrm{S}}$)得到源域目标域中心($\mu_s$，$\mu_t$)。

$$\mu_S = \sum_{i=1}^{N_S} X_S^i / N_S, \ \mu_T = \sum_{j=1}^{N_T} X_T^i / N_T$$

(2) 计算矩阵 $\boldsymbol{A}$

$$\boldsymbol{A} = ((\mu_S - \mu_T)(\mu_S - \mu_T)^{\mathrm{T}})^{-1}(\boldsymbol{X}_{\mathrm{cen}}^{\mathrm{S}} \boldsymbol{y}_{\mathrm{cen}}^{\mathrm{S}} \boldsymbol{y}_{\mathrm{cen}}^{\mathrm{S}\ \mathrm{T}} \boldsymbol{X}_{\mathrm{cen}}^{\mathrm{S}\ \mathrm{T}})$$

(3) 根据公式 $\boldsymbol{A}\boldsymbol{P} = \gamma \boldsymbol{P}$ 对 $\boldsymbol{A}$ 进行特征值分解。

(4) 得最优子空间 $P^* = [p_1, p_2, \cdots, p_k]$

(5) 计算投影到子空间后的矩阵

$$\boldsymbol{T}_{\mathrm{new}}^{\mathrm{S}} = \boldsymbol{P}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{cen}}^{\mathrm{S}}, \ \boldsymbol{T}_{\mathrm{new}}^{\mathrm{T}} = \boldsymbol{P}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{train\_cen}}^{\mathrm{T}}$$

(6) 计算回归系数

$$(\boldsymbol{T}_{\mathrm{new}}^{\mathrm{S}}, \boldsymbol{y}_{\mathrm{cen}}^{\mathrm{S}}) \xrightarrow{\text{最小二乘回归}} \beta$$

MCT 算法到此结束。

(7) 预测

$$\boldsymbol{X}_{\mathrm{test\_cen}}^{\mathrm{T}} = \boldsymbol{X}_{\mathrm{test}}^{\mathrm{T}} - \mathrm{mean}(\boldsymbol{X}^{\mathrm{S}})$$
$$\boldsymbol{T}_{\mathrm{test\_new}}^{\mathrm{T}} = \boldsymbol{X}_{\mathrm{test\_cen}}^{\mathrm{T}} \boldsymbol{P}$$
$$\hat{\boldsymbol{y}}_{\mathrm{test}}^{\mathrm{T}} = \boldsymbol{T}_{\mathrm{test\_new}}^{\mathrm{T}} \beta + \boldsymbol{y}_{\mathrm{mean}}^{\mathrm{S}}$$

2　实验部分

为了验证算法的准确性和实用性，使用玉米数据集和小麦数据集作为实验对象，对数据集进行了数据分析，来检验 MCT 方法的性能。

2.1　数据集介绍

第一个数据集是在三个近红外光谱仪(M5，MP5 和 MP6)上测量含有 80 个样本的玉米数据集。每个样品含有四种成分：水分，油，蛋白质和淀粉。波长范围为 1 100～2 498 nm，间隔为 2 nm(700 个通道)。该数据集可以从 http://www.eigenvector.com/Data/Corn/下载。使用这三个近红外光谱仪和玉米数据集成分中的水分进行研究讨论。仪器 M5 和仪器 MP5 之间的光谱差异如图 1(a)所示；仪器 M5 和仪器 MP6 之间的光谱差异如图 1(c)所示；仪器 MP5 和仪器 MP6 之间的光谱差异如图 1(e)所示。其中横轴表示波长，纵轴表示吸光度差异，曲线表示光谱样本。

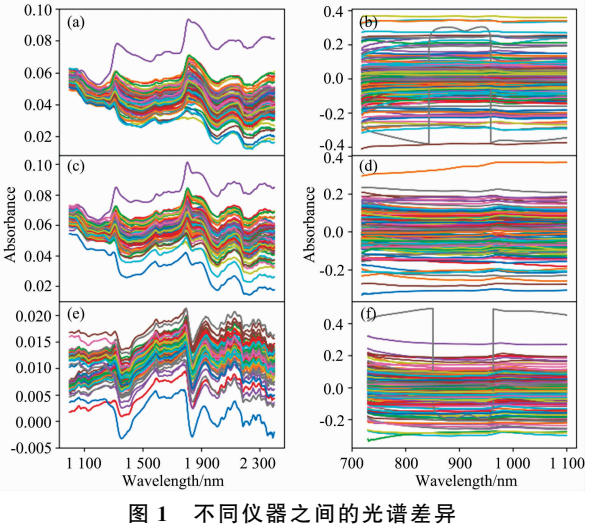

图 1　不同仪器之间的光谱差异

Fig. 1　Spectral difference between different instruments

小麦数据集由制造商 A 的三个仪器(A1，A2 和 A3)测量的 248 个样本组成。数据集只提供蛋白质参考值。波长范围为 730 ～ 1 100 nm，间隔为 0.5 nm。可在 http://www.idrc-chambersburg.org/获取。使用了三个近红外光谱仪和蛋白质含量进行研究讨论。仪器 A1 和仪器 A2 之间的光谱差异如图 1(b)所示；仪器 A1 和仪器 A3 之间的光谱差

异如图 1(d)所示；仪器 A2 和仪器 A3 之间的光谱差异如图 1(f)所示。其中横轴表示波长，纵轴表示吸光度差异，曲线表示光谱样本。

2.2 数据处理

通过 Kennard-Stone 算法将玉米数据集的 80 个样本分成两组：80%用做标定集样本，20%用做测试集样本；将小麦数据集的 248 个样本分成两组：80%用作标定集样本，20%用作测试集样本。对于有迁移标准的迁移方法，使用 Kennard-Stone 算法在标定样本上选择若干个标准样品。

2.3 性能评估

在该实验中，均方根误差(root mean square error，RMSE)被用作参数选择和模型评估的指标。此外，RMSEC 表示标定集的训练误差，RMSEP 表示测试集的预测误差。RMSE 计算方法写为

$$\mathrm{RMSE}=\sqrt{(y-\hat{y})(y-\hat{y})^{\mathrm{T}}/n} \tag{11}$$

式(11)中，$\hat{y}$ 是预测值；y 是测量值；n 是样本数目。

文中 RMSEP 代表从仪器测试集。

3 结果与讨论

选用玉米和小麦光谱数据集检验模型的性能。使用 SBC，PDS，CCACT，MSC 和 TCR 五种方法进行对比实验。对于 SBC，PDS，CCACT 和 MSC 算法均采用 PLS 算法作为主体算法，使用主仪器的光谱数据建立多元标定模型作为参考模型，用于对从仪器的待测样本进行预测。实验结果主要包含两个部分：(1) MCT 和对比方法的 RMSEC 和 RMSEP 比较；(2) MCT 和对比方法预测结果的拟合能力示意图。

MCT 和其他五种标定迁移方法的标定误差和预测误差被展示在表 1 和表 2 中。

玉米数据集实验结果分析如下：

对于仪器 MP5 到仪器 M5 的标定迁移，MCT 的 RMSEP 小于 TCR 和 MSC 这两种标准样本自由的方法，同时也小于 SBC，PDS 和 CCACT 这三种有标样的 RMSEP，并且 MCT 的 RMSEC 也低于其他五种迁移方法。对于仪器 MP6 到仪器 M5 的标定迁移，由 SBC，PDS，CCACT，TCR 和 MSC 获得的最低 RMSEP 分别为 0.36，0.40，0.41，0.47 和 1.92。表 1 中列出的结果清楚地表明 MCT 具有比其他五种方法更低的 RMSEP 和 RMSEC。对于从 MP6 到 MP5 的标定迁移，MCT 再一次达到了最小的 RMSEP 和 RMSEC。

小麦数据集实验结果分析如下：

对于仪器 A1 到 A2 的迁移，当标准样品数为 35，25 和 35 时，SBC，PDS 和 CCACT 分别取到最小值。从表 2 中能够看出方法 MCT 的 RMSEC 和 RMSEP 都小于其他五种方法的最佳结果。对于仪器 A2 到 A3 的迁移，当标准样品数为 35 时，SBC，PDS 和 CCACT 均取到最小值，由表看出 MCT 的 RMSEP 均小于其余五种方法。对于仪器 A3 到 A2 的迁移，MCT 再一次达到了最小 RMSEP 和 RMSEC。

这六组对比实验可以看出，MCT 模型在通常情况下能够取得最优的预测效果，并具有更好的鲁棒性。

图 2—图 4 和图 5，图 6 分别显示了在玉米集和小麦集中，六种不同的标定迁移方法的预测值与测量值的关系图。预测浓度和测量浓度之间的零差异，将会使得样本点在直线上。对于有标准样本的标定迁移方法，选取预测性能最优时的数据用于比较，以便更加充分的体现出 MCT 能够取得良好的预测性能。表 3 是六种迁移方法的预测值与测量值曲线拟合斜率表。

表 1 SBC，PDS，TCR，CCACT，MSC 和 MCT 六种迁移方法在玉米数据集下的 RMSEC，RMSEP

Table 1 RMSEC and RMSEP of corn datasets with SBC，PDS，TCR，CCACT，MSC and MCT

玉米数据集(水分)		m5 主仪器-mp5 从仪器		m5 主仪器-mp6 从仪器		mp5 主仪器-mp6 从仪器	
迁移方法	N	RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP
SBC	15	0.33	0.32	0.34	0.44	0.21	0.23
	25	0.35	0.25	0.37	0.39	0.19	0.23
	35	0.32	0.27	0.37	0.36	0.19	0.24
PDS	15	0.32	0.26(15^{a})	0.54	0.43(7^{a})	0.24	0.35(15^{a})
	25	0.34	0.24(15^{a})	0.60	0.47(9^{a})	0.22	0.39(15^{a})
	35	0.32	0.24(15^{a})	0.50	0.40(5^{a})	0.23	0.35(15^{a})
CCACT	15	0.42	0.24	0.71	0.55	0.43	0.25
	25	0.39	0.26	0.60	0.42	0.41	0.25
	35	0.36	0.27	0.54	0.41	0.37	0.25
TCR	0	1.10	0.43(6^{b})	1.30	0.47(4^{b})	0.83	0.36(10^{b})
MSC	0	1.63	1.79	1.87	1.92	0.92	0.90
MCT	0	0.24	0.20	0.27	0.19	0.16	0.14

注：N：需要标准样本的迁移方法中，标准样本的数目；a：PDS 中最优的窗口大小；b：TCR 中对应的最优子空间的维度，下同

Note：N：the numbers of standard samples required by the transfer method；a：the optimal window size in PDS；b：the dimension of corresponding optimal subspaces in TCR；the same below

表 2　SBC，PDS，TCR，CCACT，MSC 和 MCT 六种迁移方法在小麦数据集下的 RMSEC，RMSEP

Table 2　RMSEC and RMSEP of wheat datasets with SBC，PDS，TCR，CCACT，MSC and MCT

小麦数据集(蛋白质)		A2 主仪器-A1 从仪器		A3 主仪器-A2 从仪器		A2 主仪器-A3 从仪器	
迁移方法	*N*	RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP
SBC	15	0.64	0.59	8.74	8.66	0.44	0.48
	25	0.58	0.56	8.26	7.08	0.54	0.48
	35	0.52	0.55	7.73	6.00	0.54	0.48
PDS	15	2.20	2.68(5[a])	4.19	3.59(15[a])	1.81	1.52(7[a])
	25	1.81	2.11(15[a])	3.19	2.23(3[a])	1.58	1.41(15[a])
	35	2.23	2.39(15[a])	2.42	2.07(3[a])	1.67	1.35(15[a])
CCACT	15	2.67	2.36	2.70	2.25	3.26	2.03
	25	2.54	2.04	2.84	2.25	2.94	2.07
	35	2.53	1.20	2.46	2.06	2.75	1.90
TCR	0	4.33	2.42(18[b])	4.16	2.12(10[b])	4.15	1.92(18[b])
MSC	0	1.63	1.39	3.80	1.25	1.43	1.22
MCT	0	0.38	0.53	2.45	0.63	0.42	0.30

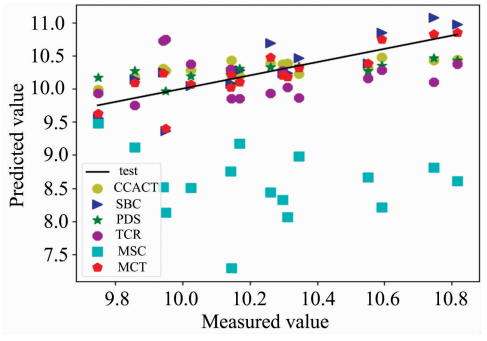

图 2　SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 M5 和仪器 MP5 之间预测结果的散点图

Fig. 2　Scatter plots for prediction between instruments M5 and MP5 in SBC，PDS，CCACT，TCR，MSC and MCT

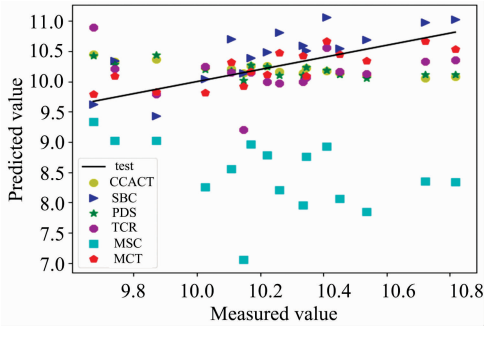

图 3　SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 M5 和仪器 MP6 之间预测结果的散点图

Fig. 3　Scatter plots for prediction between instruments M5 and MP6 in SBC，PDS，CCACT，TCR，MSC and MCT

对于玉米数据集表，图 2—图 4 显示 MCT 方法的预测结果相比其他五种迁移方法具有更好的预测性。根据表 3 中的数据也可证明 MCT 相比其他方法更加接近直线。通过上面的陈述，可以得到结论：MCT 能够在玉米集所有模型中实现最佳的预测性能，同时具有更好的泛化能力。

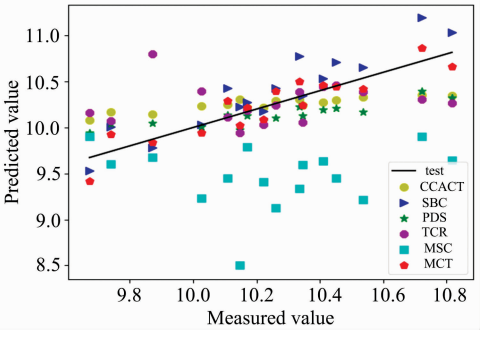

图 4　SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 MP5 和仪器 MP6 之间预测结果的散点图

Fig. 4　Scatter plots for prediction between instruments MP5 and MP6 in SBC，PDS，CCACT，TCR，MSC and MCT

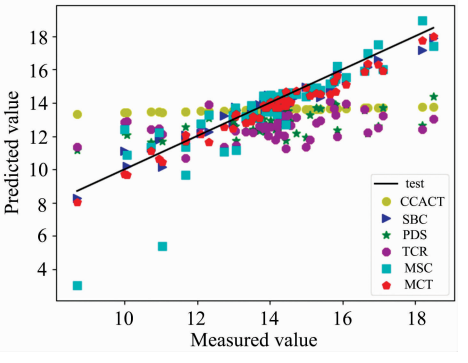

图 5　SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 A2 和仪器 A1 之间预测结果的散点图

Fig. 5　Scatter plots for prediction between instruments A2 and A1 in SBC，PDS，CCACT，TCR，MSC and MCT

对于小麦数据集，图 5—图 7 及表 3 中均可以看出，MCT 的样本点更加接近直线，相比其他五种方法，其能够达到更好的预测效果。通过上述对比，可以很容易地得到结

论：MCT 在小麦集的所有模型中能够实现最佳的预测性能，同时具有更好的泛化能力。

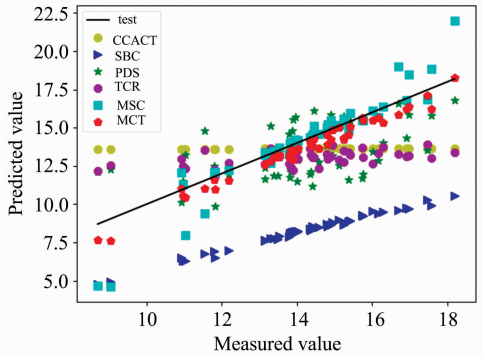

图 6 SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 A3 和仪器 A2 之间预测结果的散点图

Fig. 6 Scatter plots of prediction results between instruments A3 and A2 by SBC，PDS，CCACT，TCR，MSC and MCT

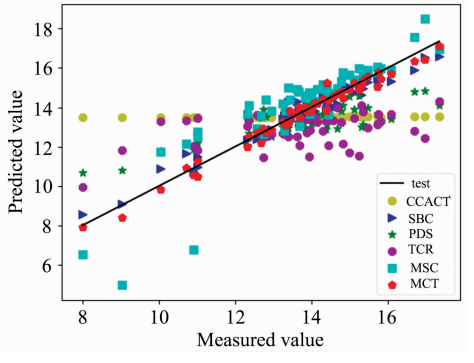

图 7 SBC，PDS，CCACT，TCR，MSC 和 MCT 六种方法在仪器 A2 和仪器 A3 之间预测结果的散点图

Fig. 7 Scatter plots of prediction results between instruments A2 and A3 by SBC，PDS，CCACT，TCR，MSC and MCT

表 3 SBC，PDS，TCR，CCACT，MSC 和 MCT 六种迁移方法预测结果斜率对比表

Table 3 Slope comparison of predicted results of SBC，PDS，TCR，CCACT，MSC and MCT

迁移方法	slope					
	m5-mp5	m5-mp6	mp5-mp6	A2-A1	A3-A2	A2-A3
SBC	0.58	0.55	0.68	1.11	1.65	1.12
PDS	1.66	−1.84	2.30	1.62	0.56	1.72
CCACT	1.82	−2.69	3.84	22.55	164.78	163.26
TCR	0.01	−0.01	0.15	0.91	2.22	1.03
MSC	−0.10	−0.24	−0.01	0.67	0.64	0.69
MCT	1.06	0.74	1.02	0.93	1.02	0.99

3 结 论

提出了一种基于最小化平均分布差异的标准样本自由 NIR 偏最小二乘标定迁移方法。该方法学习了如何找到能够使主从仪器数据投影后，两域平均分布差异最小的同时，还能使主光谱投影后的数据相关性最大的一个公共子空间。在该子空间中，主从仪器的数据分布得到了极大的校正，能够使从仪器共用主仪器模型，实现标定迁移。

在玉米和小麦数据集中，使用 SBC，PDS，CCACT，TCR 和 MSC 作为对比实验来检验 MCT 方法的性能，并且所提出的方法(MCT)通常实现了最佳的 RMSEC 和 RMSEP。结果清楚地表明，MCT 能够成功地用于校正在不同仪器上测量的光谱之间的差异。对于 SBC，PDS 和 CCACT 这三种迁移方法，它们需要标准样品建立迁移模型。在 TCR 中，从仪器样品还需要少量的参考值。这两个条件在实际应用中，都会产生很昂贵的代价，甚至无法满足这一条件。因此，当标准样品在实际应用中不可获得时，同时对比现有标准样本自由方法，MCT 是一种有效的标定迁移方法。

References

[1] Aryal G H，Hunter K W，Huang L. Organic & Biomolecular Chemistry，2018，16(40)：7425.
[2] Rahimpour A，Noubari H A，Kazemian M. Informatics in Medicine Unlocked，2018，11：44.
[3] Fukuda M. Seishin Shinkeigaku Zasshi=Psychiatria et neurologia Japonica，2015，117(2)：79.
[4] Debarpan G，Anirban D. Frontiers in Neuroscience，2016，10：261.
[5] Campbell W，Coller A，Noble S，et al. Waste and Biomass Valorization，2020，11：2959.
[6] Zheng K，Feng T，Zhang W，et al. Analytical Methods，2020，12(11)：1495.
[7] Zimmerman N，Presto A A，Kumar S P N，et al. Atmospheric Measurement Techniques，2018，11(1)：291.
[8] Workman J J Jr. Applied Spectroscopy，2018，72(3)：340.
[9] Malli B，Birlutiu A，Natschläger T. Chemometrics and Intelligent Laboratory Systems，2016，161：49.
[10] Brown S D. Transfer of Multivariate Calibration Models. Elsevier Inc.，2019. 345.
[11] Zheng K，Zhang X，Iqbal J，et al. Journal of Chemometrics，2014，28(10)：773.
[12] Abdelkader M F，Cooper J B，Larkin C M. Chemometrics and Intelligent Laboratory Systems，2012，110(1)：64.
[13] Ouyang G，Cai J，Zhang X，et al. Journal of Separation Science，2015，31(6-7)：1167.
[14] LIU Yan-de，XU Hai，SUN Xu-dong，et al(刘燕德，徐 海，孙旭东，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2020，40(3)：992.
[15] Pan S J，Tsang I W，Kwok J T，et al. IEEE Transactions on Neural Networks，2011，22(2)：199.

NIR Calibration Transfer Method Based on Minimizing Mean Distribution Discrepancy

ZHAO Yu-hui, LU Peng-cheng, LUO Yu-bo, SHAN Peng
Northeastern University Qinhuangdao Campus, Qinhuangdao 066000, China

Abstract With the advantages of high efficiency, non-destructive and environmental protection, NIR is widely used in many fields to rapidly analyse substances. However, it is still faced with the problems of the short life cycle of spectral calibration model and difficulty obtaining and preserving standard samples for instrument calibration transfer method. In the stoichiometric literature, transfer methods usually correct the spectral differences between master and slave instruments. Most methods need to measure a set of transfer standard samples under the same conditions of two instruments. Although the number of samples does not need to be too much, generally speaking, it must be well selected to ensure a successful transfer. The Kennard-Stone algorithm is the main algorithm for selecting representative sample subset in the master-slave instrument. In determining the standard sample, it is assumed that the master instrument has found the standard sample, and the selected sample set needs to be measured in the slave instrument. It is only possible when the transferred sample is sufficiently stable, but this cannot be guaranteed in the near-infrared spectroscopy technology. If it is assumed that the sample of the slave instrument is used as the standard sample, the master instrument is replaced by the slave instrument in consideration of the change of the spectrum light source in the new industrial application, so it is no longer available. Based on these problems, this paper proposes a method of minimizing mean distribution discrepancy calibration transfer for NIR (MCT), without considering the standard sample (standard-free) of the slave instrument, due to the multicollinearity of NIR spectroscopy data, this method first assumes that there is a subspace of the partial least squares of the master-slave instrument, and then the spectral data of the master-slave instrument are projected to the common subspace respectively; then, the mean distribution discrepancy minimization algorithm is introduced, that is, the mean distribution (center point) representation function of the master-slave spectral data in the subspace is given Function to minimize the discrepancy between the mean distribution (center point) of the two spectra, and maximize the covariance of the main instrument spectrum after projection to derive the optimal subspace; finally, the main spectrum samples and the secondary spectrum prediction samples are projected into the partial least squares subspace respectively, and the regression model is obtained by using the main spectral data, and the modified model can be used to predict the secondary spectral concentration. Through the test and research on the corn data set and the wheat data set, it is proved that the prediction effect of this method is improved compared with SBC, PDS, CCACT, TCR and MSC. The experiment shows that MCT can achieve a lower prediction value.

Keywords Near infrared spectroscopy; Calibration transfer; Mean distribution discrepancy; Standard-free; Partial least square regression

(Received Oct. 8, 2020; accepted Feb. 27, 2021)