干贝水分检测的建模及分级方法

黄 慧[1,2*]，张德钧[1]，詹舒越[1]，沈 晔[1]，王杭州[1]，宋 宏[1]，徐 敬[1]，何 勇[3]

1. 浙江大学海洋学院，浙江 舟山 316021
2. 农业部渔业装备与工程技术重点实验室，上海 200092
3. 浙江大学生物系统工程与食品科学学院，浙江 杭州 310058

摘 要 高光谱成像已被应用于建立干贝水分含量预测模型，其模型性能受样本划分方法及建模方法影响。样本划分方法决定着所选样本是否具有代表性，而建模方法决定着如何利用样本建立模型，但样本划分方法与建模方法的内在联系却鲜有研究报道。在方法优选上，将样本划分方法与建模方法进行组合，探究不同方法组合对干贝水分含量预测模型性能的影响，对干贝水分检测建模及分级方法的优选具有重要意义，同时也能为其他样本的光谱建模提供参考。采集380～1 030 nm波段下270个干贝样本的高光谱图像，提取干贝样本的光谱数据，通过RS，KS，SPXY和CG四种常用的方法划分样本，并以PLSR和LS-SVM两种常用的建模方法建立多个干贝水分含量预测模型，计算和比较各模型的性能指标。结果表明：PLSR模型使用RS法划分干贝水分含量样本最为适宜(其RPD为4.079 6)，LS-SVM模型使用SPXY划分法最为适宜(其RPD为4.175 6)，划分样本方法的优劣与建模方法有关，其优选需要结合特定的建模方法进行。在常用的四种样本划分方法和两种建模方法中，采用SPXY法划分干贝水分含量样本并结合LS-SVM法建模的效果和精度最好。

关键词 高光谱数据；干贝；样本划分；建模方法

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)01-0185-08

引 言

水分含量是决定干贝品质的重要指标。水分含量较低，有利于延长干贝的贮存期，但水分含量过低，则会严重影响干贝的硬度、弹性和咀嚼性等品质指标。高光谱成像包含丰富的光谱图像信息，具有无损、准确、快速等优点，可用于检测干制品的水分含量。本课题组已采用高光谱成像系统对380～1 030 nm波段范围内的6个不同干燥时期共90个干贝样本高光谱图像建立了一种干贝水分含量预测模型[1]。由于高光谱图像包含的数据量大，因而需要采用合适的样本划分法使各划分集的样本数据具备代表性。样本划分方法对模型性能有着重要的影响，其优选也是众多研究的热点。

样本划分方法用于将样本总体划分为建立预测模型所需要的建模集和验证预测模型性能所需要的预测集，并保证各样本划分集的数量能满足建模的数理统计要求。应用较为广泛的样本划分方法有随机划分法(random sampling，RS)[2]、Kennard-Stone法(KS)[3]、光谱-理化值共生距离法(sample set partitioning based on joint X-Y distances，SPXY)[4]和浓度梯度法(concentration gradient，CG)[5]。利用划分好的样本建立水分含量预测模型是高光谱检测干贝水分含量的核心环节，不同建模方法建立的模型具有不同的预测效果。偏最小二乘回归法(partial least squares regression，PLSR)和最小二乘支持向量机法(least squares support vector machine，LS-SVM)拟合效果良好，且使用较为广泛的回归分析法。刘善梅等[6]比较RS，KS，SPXY和CG等不同样本划分法下的土猪肉含水率PLSR模型效果，认为PLSR模型下，CG法最适合划分土猪肉含水率样本。刘雪梅等对RS，KS，SPXY等样本划分法进行分析比较后，选择SPXY法划分样本并基于LS-SVM建模，获得了较为理想的水产养殖水体化学需氧量预测模型。建模方法决定着如何利用样本数据，但鲜有研究将样本划分方法与建模方法组合起来进行优选。

本研究将分别在RS，KS，SPXY和CG这四种常用的样本划分方法下结合偏最小二乘回归法、最小二乘支持向量机

收稿日期：2017-08-12，**修订日期**：2017-12-25

基金项目：农业部渔业装备与工程技术重点实验室开放基金项目(N20150117)，浙江省教育厅项目(N20140264)和中央高校基本科研业务费专项(172210161)，国家自然科学基金项目(41606214)资助

作者简介：黄 慧，女，1986年生，浙江大学海洋学院讲师 e-mail：huih@zju.edu.cn *通讯联系人

法这两种建模方法，建立 8 个干贝高光谱水分含量检测模型，并采用校正集相关系数 R_c(correlation coefficients of calibration)，预测集相关系数 R_p(correlation coefficients of prediction)，建模集均方根误差 RMSEC(root mean square error of calibration)，预测均方根误差 RMSEP(root mean square error of prediction)和剩余预测偏差 RPD(residual predictive deviation)五个性能指标对所建模型进行评估以得到用于干贝水分含量检测的样本划分方法与建模方法最优组合。

1 实验部分

1.1 材料

实验室组建的高光谱成像系统主要包括高精度 CCD 相机(Hamamatsu，Japan)、分辨率为 672×512 的成像光谱仪(ImSpector V10E，Spectral imaging Ltd.，Oulu，Finland)、对称分布的两个 150 W 线光源(Schott Fostec-A08912)、电控位移平台(Isuzuoptics，Taiwan，China)和计算机。成像光谱仪的光谱范围为 380～1 030 nm，光谱分辨率为 2.8 nm。整套系统置于一个室温下的暗箱中，以防止环境光的影响。

实验所用的新鲜海湾扇贝肉购于杭州水产品市场，用保鲜袋包装并编号、冷藏，运至实验室，所有样品均用去离子水洗净后，晾干置于－20 ℃的冷冻柜中存储待用。

1.2 样本数据获取

在冷冻柜中取出存储的所有样品，沥干水分后，置于(105±5)℃的烘箱中干燥 10 min，取出放入干燥器中直到温度冷却至 20 ℃时转到称重仪再次称重，记为 M_i。接着采用高光谱成像系统采集光谱图像。循环烘干、称重、采集图像这三步骤直到扇贝烘干至不再有弹性为止，称重记为 M_f，最后得到 270 个扇贝所对应的高光谱图像。水分含量(MC)通过式(1)计算得到。

$$MC = (M_i - M_f)/M_i \times 100\% \quad (1)$$

然后，将高光谱图像中的干贝肉部分划为感兴趣区域(region of interest，ROI)，提取 ROI 内所有像素点的平均光谱反射率曲线，并将每个样本的平均光谱数据保存。

1.3 样本划分法

分别采用随机划分法(RS)、Kennard-Stone 法(KS)、光谱-理化值共生距离法(SPXY)和浓度梯度法(CG)这四种常用的样本划分方法将总体样本划分为建模集和预测集。各划分方法的原理如下：

(1)随机划分法(RS)

随机选取一定数量的样本构成建模集和预测集，无规律可循。RS 法能保证划分的随机性，但可能导致选取的样本之间存在差异化从而无法保证样本的代表性。

(2) Kennard-Stone 法(KS)

KS 法是按照样本光谱空间中的欧式距离进行样本挑选。首先选择相互间有着最大的欧氏距离的两个样本加入建模集，然后在剩余样本中选出与已挑选出的建模样本中欧氏距离最大的样品加入到建模样品集，循环进行计算，直至挑选出足够样本数量的建模集，此时将剩余样本作为预测集样本。欧氏距离通过式(2)计算得到。

$$d_x(a,b) = \sqrt{\sum_{i=1}^{m}[x_a(i) - x_b(i)]^2} \quad a,b \in [1,n] \quad (2)$$

式(2)中，$d_x(a,b)$为样本 a 和样本 b 间的欧式距离，$x_a(i)$和 $x_b(i)$分别为样本 a 和 b 在第 i 个波长处的光谱反射率，样本光谱波长总数用 m 表示，样本总数用 n 表示。

(3)光谱-理化值共生距离法(SPXY)

在 KS 法的基础上，Galvão 等提出了能兼顾样本光谱信息与被测指标参量的光谱-理化值共生距离法(SPXY)。SPXY 法与 KS 法的原理和步骤相似，不同之处在于二者样本欧氏距离的计算方法不同。SPXY 法计算样本欧氏距离的公式如式(3)所示。

$$d_{xy}(a,b) = \frac{d_x(a,b)}{\max\limits_{a,b\in[1,n]} d_x(a,b)} + \frac{d_y(a,b)}{\max\limits_{a,b\in[1,n]} d_y(a,b)} \quad a,b \in [1,n] \quad (3)$$

式(3)中，$d_x(a,b) = \sqrt{\sum_{i=1}^{m}[x_a(i) - x_b(i)]^2}$，$x_a(i)$ 和 $x_b(i)$分别为样本 a 和样本 b 的第 i 个波段的光谱信息值，$d_y(a,b) = |y_a - y_b|$，y_a 和 y_b 分别为样本 a 和样本 b 的被测指标参量，$d_x(a,b)$ 为样本 a 和样本 b 间在光谱空间的欧式距离，$\max\limits_{a,b\in[1,n]} d_x(a,b)$ 为光谱空间两两样本欧氏距离的最大值，$\max\limits_{a,b\in[1,n]} d_y(a,b)$ 为被测指标参量空间两两样本欧氏距离的最大值，$d_{xy}(a,b)$ 为兼顾样本光谱信息与被测指标参量的样本 a 和样本 b 之间的距离，样本总个数用 n 表示。

(4)浓度梯度法(CG)

浓度梯度法(CG)是充分考虑被测指标参量的代表性的方法。首先将样本按照被测指标参量递增或递减的顺序依次排列，然后按照建模集与预测集的比例，间隔性地选择样本加入预测集，并不将最后一个样本选入预测集(如果选中，则从建模集中除了第一个外随机选择一个样本与之交换)。

1.4 预测模型的建立方法

将通过 RS，KS，SPXY 和 CG 四种划分方法得到的四个建模集分别采用偏最小二乘回归法(PLSR)与最小二乘支持向量机法(LS-SVM)建立模型，得到 8 个干贝高光谱水分含量检测模型。各建模方法的原理如下：

(1)偏最小二乘回归法(PLSR)

偏最小二乘回归是一种在自然、经济和社会等众多科学领域应用广泛的一种新型多元统计数据分析方法[7]。该算法建立的模型是多个或单个因变量 Y 对多自变量 X 的回归模型，在建模的过程中，既包含主成分分析(principal component analysis，PCA)的尽量提取 Y 和 X 中的主成分的思想，又考虑使分别从 X 和 Y 提取出的主成分之间的相关性最大化这种典型关联分析(canonical correlation analysis，CCA)的思想。因此，偏最小二乘回归是 PCA 法、CCA 法和多元线性回归分析法这三种分析方法组合而成的算法。

(2)最小二乘支持向量机法(LS-SVM)

支持向量机(support vector machines，SVM)回归是分类支持向量机推广应用到函数回归问题上的一种方法[8]。该方法是以结构风险最小原理和非线性映射为基本思想，将低维空间非线性问题映射为高维空间线性问题进行求解，其核

心为引入的核函数。

Suykens 等提出的最小二乘支持向量机(least squares support vector machines，LS-SVM)从机器学习损失函数着手，将二范数应用到目标函数中，并且在支持向量机标准算法中把约束条件的不等式以等式替换，将 LS-SVM 方法的优化问题求解转为 Kuhn-Tucker 条件[9]下得到的一组线性方程组的求解，简化了 SVM 算法，并提高了计算效率。

在 LS-SVM 的实际应用中，常常选择径向基核函数，其主要参数是对 LS-SVM 的学习能力及泛化能力有着决定性作用的正则化参数以及核函数参数[10]。本文选择功能良好且应用广泛的基于贝叶斯框架优选参数的 LS-SVM 建模方法[11]进行干贝水分含量的 LS-SVM 模型建立。

1.5　模型的建立与评价

基于 RS，KS，SPXY 和 CG 四种划分方法得到四个建模集与预测集。样本高光谱图像数据由样本的光谱信息量和相应的被测指标参量组成，样本划分的代表性要求建模集一般包含被测指标参量最小与最大的样本，常见的建模集和预测集的样本的数量比例为 2∶1～3∶1[12]，本研究采用的划分比例为 2∶1。

对不同方法划分的建模集，采用 PLSR 和 LS-SVM 分别建立干贝的高光谱水分含量检测模型，通过计算相关系数 r[式(4)]，均方根误差 RMSE(root mean square error)[式(5)]和相对分析误差 RPD[式(6)]评估预测模型的性能。

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

式(4)中，r 在校正集中表示为 R_c，在预测集中表示为 R_p。n 为对应样本集中的样本数，x_i 为对应样本集中的样本 i 实测值，$\bar{x}$ 为 x_i 的平均值；y_i 为对应样本 i 的预测值，$\bar{y}$ 为 y_i 的平均值。模型的拟合效果与相关系数 r 接近 1 的程度成正比。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2} \tag{5}$$

式(5)中，RMSE 在校正集中表示为 RMSEC，在预测集中表示为 RMSEP。n 为对应样本集中的样本数，x_i 为对应样本集中的样本 i 实测值，y_i 为对应样本 i 的预测值。模型的预测精度与均方根误差 RMSE 成反比。

$$\mathrm{RPD}=\frac{\mathrm{std}(y)}{\mathrm{RMSE}} \tag{6}$$

式(6)中，y 为所有样本的被测指标参量的集合，std(y)为被测指标参量的标准差。模型的预测效果与剩余预测偏差 RPD 成正比。

2　结果与讨论

2.1　干贝样本水分含量统计结果

经计算，270 个海湾干贝样本的水分含量范围为 17.99%～66.73%，均值 39.93%，标准差为 12.25%。分别采用 RS，KS，SPXY 和 CG 四种划分方法，将 270 个干贝样本按 2∶1 的比例划分成建模集(180 个样本)和预测集(90 个样本)，其划分后的水分含量统计结果见表 1。

从表 1 可以发现，以上四种划分方法中，只有 KS 法在划分样本时，水分含量的最大值 66.73%没有出现在建模集，没有满足建模集一般应包含最大值及最小值的要求，故而划分结果欠缺代表性。

另一方面，由表 1 可计算得：RS 法所得的建模集和预测集的均值与样本集总体均值 39.93%分别相差 0.77%和 0.54%，标准差与样本集总体标准差 12.25%相差分别为 0.63%和 0.97%。SPXY 法所得的建模集和预测集的均值分别与样本集总体均值相差 1.73%和 3.45%，标准差则分别相差 0.32%和 1.41%。CG 法所得的建模集和预测集的均值与样本集总体均值分别相差 0.03%和 0.07%，标准差与样本集总体标准差分别相差 0.09%和 0.12%。

表 1　RS，KS，SPXY 和 CG 法划分样本后的干贝水分含量统计结果

Table 1　The statistical results of moisture content in dried scallop with different division methods

划分方法	建模集				预测集			
	样本数	浓度范围/%	平均值/%	标准差/%	样本数	浓度范围/%	平均值/%	标准差/%
RS	180	17.99～66.73	40.70	11.62	90	20.80～66.13	39.39	13.22
KS	180	17.99～66.13	41.40	12.68	90	21.02～66.73	37.00	10.81
SPXY	180	17.99～66.73	41.66	12.57	90	19.98～62.28	36.48	10.84
CG	180	17.99～66.73	39.90	12.34	90	20.07～65.65	40.00	12.13

由此可见，单从水分含量数据这一被测指标参量而言，CG 法得到的建模集和预测集的水分含量这一被测指标参量的数据分布比较均匀，这是因为 CG 法是根据被测指标参量按梯度排序后进行划分的方法。然而，划分方法的总体效果不仅与被测指标参量有关，还与光谱信息量及建模方法有关，需要根据模型性能进一步比较和分析。

2.2　模型的比较与分析

为进一步对比分析 KS，CG，SPXY 和 RS 这四种样本划分方法对水分含量预测模型性能的影响。将 RS，KS，SPXY 和 CG 这四种划分方法得到的建模集样本数据分别建立 PLSR 和 LS-SVM 模型。

2.2.1　干贝水分检测 PLSR 模型的样本划分法优选

分别使用 KS 法、CG 法、SPXY 法和 RS 法划分的样本数据进行建模得到了对应的 4 个 PLSR 模型的建模集。4 个 PLSR 模型性能指标见表 2。

表 2　RS，KS，SPXY 和 CG 法划分样本后的干贝水分含量

PLSR 模型结果

Table 2 The PLSR modeling results of moisture content in dried scallop with different division methods

划分方法	R_c	R_p	RMSEC/%	RMSEP/%	RPD
RS	0.935 7	0.973 2	4.161 6	3.244 9	4.076 9
KS	0.944 0	0.932 2	4.174 1	3.906 8	2.768 1
SPXY	0.941 1	0.944 6	4.238 7	3.792 5	2.858 6
CG	0.962 8	0.956 4	3.325 6	3.534 4	3.430 6

如表 2 所示，四种方法均获得了较好的校正和预测结果。其中，基于 RS 法建立的水分含量预测模型效果最优，其预测集相关系数 R_P 在所有模型中最高，达到了 97.32%，预测均方根误差 RMSEP 以及剩余预测偏差 RPD 都是所有模型中的最优值，分别为 3.244 9%和 4.076 9。

为了直观观察四种样本划分方法在 PLSR 模型的建模集中的预测效果，绘制 PLSR 模型对建模集样本的预测水分含量散点分布图。

如图 1 所示，四种样本划分方法均在对建模集的水分含量预测中取得了不错的拟合效果。使用 4 种样本划分方法对应的 PLSR 模型对各自的预测集中的高光谱数据进行干贝水分含量预测，得到的直观的预测效果见图 2。

综合考虑以上分析指标，基于 RS 法建立的水分含量预测模型效果最优。说明在干贝水分含量 PLSR 预测模型的样本划分上，应该选择更为注重划分的随机性的 RS 法。

2.2.2 干贝水分检测 LS-SVM 模型的样本划分法优选

在 LS-SVM 的建模方法下，分别使用 KS 法、CG 法、

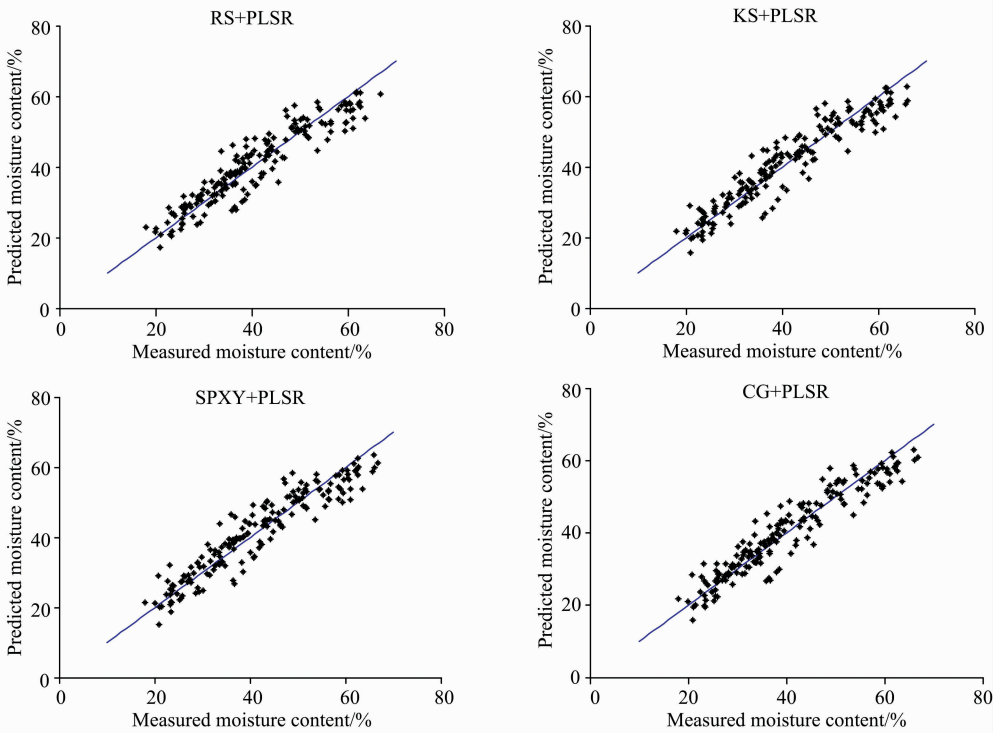
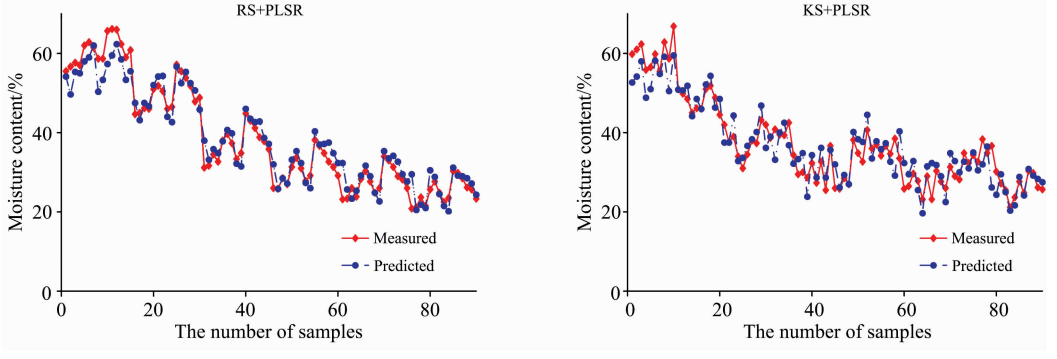


图 1 PLSR 模型下建模集水分含量实测值与预测值的拟合效果图

Fig. 1 The fitting effect of predicted value and measured value in PLSR model



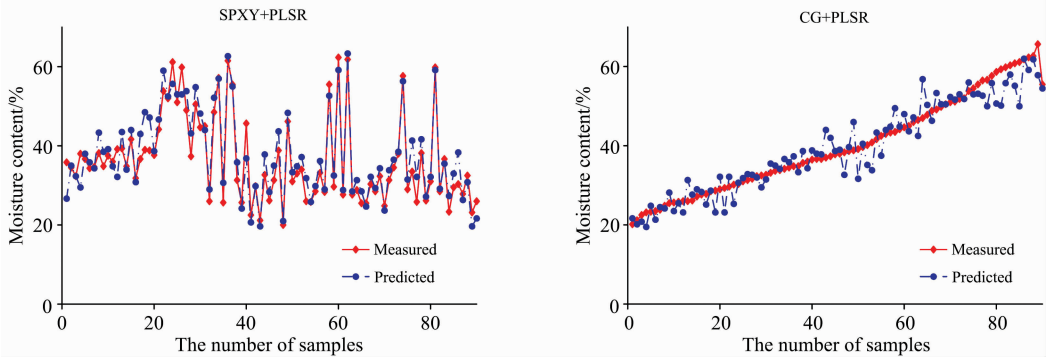


图 2　PLSR 模型下预测集水分含量实测值与预测值的对比曲线图

Fig. 2　The comparion between predicted value and measured value in PLSR model

SPXY 法和 RS 法进行样本划分，得到对应的 4 个 LS-SVM 模型的性能指标见表 3。

对于 LS-SVM 模型，基于 RS，KS 和 SPXY 法划分样本而建立的干贝水分含量预测模型的性能均有较好的表现。

图 3 直观反映了 LS-SVM 模型下，四种样本划分方法在建模集中的拟合效果。其中，基于 RS，KS 和 SPXY 样本划分法的 LS-SVM 模型在建模集中都有不错的拟合效果，基于 CG 样本划分法的 LS-SVM 模型在建模集中的拟合效果明显次于其他三种样本划分法。这四个 LS-SVM 模型对预测集中样本水分含量的直观预测效果见图 4。

由表 3 及图 3 和图 4 可得，基于 CG 划分法的 LS-SVM 模型表现不如其他三个模型，其各项性能指标以及模型预测效果都较差。与其他方法相比，基于 SPXY 划分法的 LS-SVM 模型的预测效果最佳，其预测相关系数 R_P 和剩余预测偏差 RPD 最优，分别为 0.971 5 和 4.175 6，同时也保证了预测均方根误差 RMSEP 较小。

表 3　RS，KS，SPXY 和 CG 法划分样本后的干贝水分含量 LS-SVM 模型结果

Table 3　The LS-SVM modeling results of moisture content in dried scallop with different division methods

划分方法	R_c	R_p	RMSEC/%	RMSEP/%	RPD
RS	0.959 9	0.945 4	3.304 3	4.355 7	3.038 8
KS	0.964 6	0.967 3	3.380 1	2.770 0	3.886 6
SPXY	0.962 1	0.971 5	3.459 6	2.862 4	4.175 6
CG	0.697 6	0.835 1	11.576 1	11.072 5	1.089 0

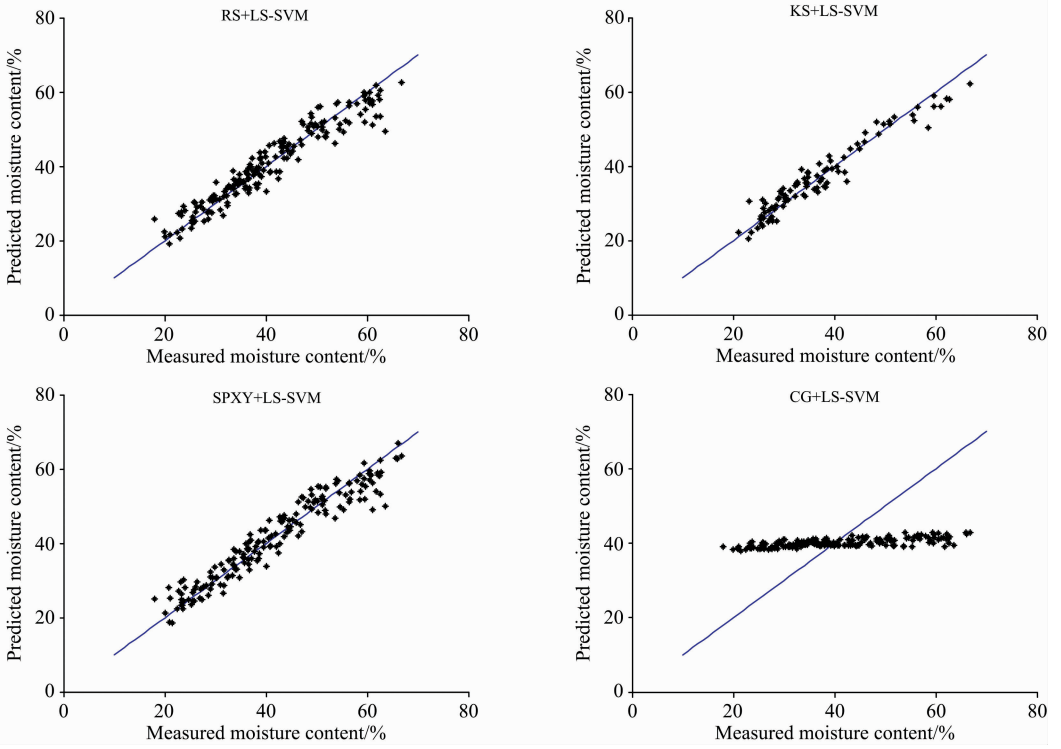


图 3　LS-SVM 模型下建模集水分含量实测值与预测值的拟合效果图

Fig. 3　The fitting effect of predicted value and measured value in LS-SVM model

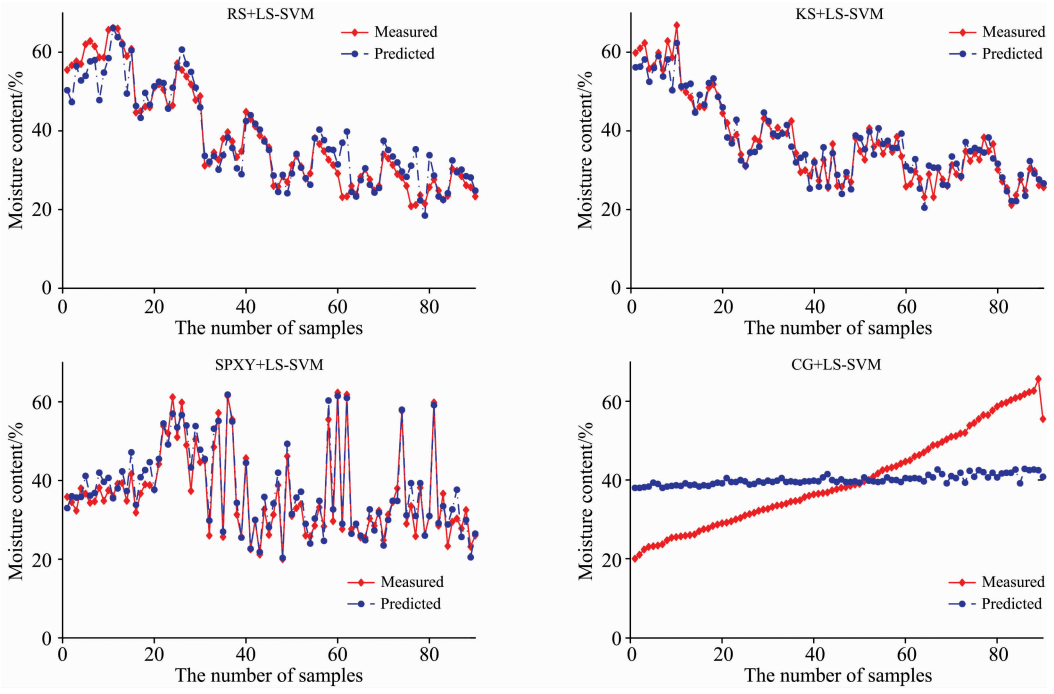


图 4 LS-SVM 模型下预测集水分含量实测值与预测值的对比曲线图

Fig. 4 The comparion between predicted value and measured value in LS-SVM model

因此，在使用 LS-SVM 法建模时，SPXY 法、KS 法和 RS 划分法均适用于划分干贝水分含量样本。SPXY 划分相对 KS，RS 和 CG 划分，与 LS-SVM 模型结合具有更高的预测精度，说明同时兼顾样本光谱数据以及水分含量的代表性的 SPXY 样本划分法最适合应用于干贝水分检测的 LS-SVM 模型。

为了进一步分析 CG 样本划分法在 LS-SVM 模型中表现较差的原因，求取在 LS-SVM 建模中基于贝叶斯框架优选法得到的参数组合(γ，δ^2)，结果见表 4。

表 4 RS，KS，SPXY 和 CG 法划分样本的干贝水分含量 LS-SVM 模型参数

Table 4 The LS-SVM modeling parameters of moisture content in dried scallop with different division methods

划分方法	正则化参数 γ	核函数参数 δ^2
RS	54.598 2	1 172.074 8
KS	54.598 2	1 248.417 7
SPXY	54.598 2	1 232.559 4
CG	0.018 3	1 144.967 1

如表 4 所示，CG 法划分样本的正则化参数和其他三种划分法的差距很大，而其核函数参数与其他三种方法的差距则较小。参数的选择直接决定了 LS-SVM 模型的性能，CG 样本划分法在贝叶斯框架优选下没有得到理想的正则化参数，所以 LS-SVM 建模的效果并不好。

2.2.3 干贝水分检测建模及其样本划分法优选

比较四种划分方法在两种不同建模方法下的水分预测的表现，当选择 LS-SVM 建模时，样本划分建议选择 SPXY 法划分样本法；当选择 PLSR 建模时，样本划分建议选择 RS 法。具体而言，在使用 SPXY 法＋LS-SVM 法和 RS 法＋PLSR 法所建立模型的性能指标中：SPXY 法＋LS-SVM 法建立的干贝水分含量预测模型的校正相关系数 R_c 为 0.962 1，略高于 RS 法＋PLSR 法(R_c＝0.935 7)，而其预测相关系数 R_p 为 0.971 5，与 RS 法＋PLSR 法的(R_p＝0.973 5)相近，说明两者具有相似的拟合效果；在预测精度指标上，SPXY 法＋LS-SVM 法的校正均方根误差 RMSEC 为 3.459 6%，预测均方根误差 RMSEP 为 2.862 4%，均较 RS 法＋PLSR 法(其 RMSEC＝4.161 6%，RMSEP＝3.244 9%)小，其剩余预测偏差 RPD 为 4.175 6 较 RS＋PLSR 法的(RPD＝4.076 9)大，说明 SPXY 法＋LS-SVM 法的模型具有更为理想的预测精度。因此，对于干贝水分含量检测，建议使用 SPXY 法划分干贝水分含量样本同时配合使用 LS-SVM 法建立干贝水分含量预测模型。

此外，如表 2 及图 2 所示，CG 法、SPXY 法、KS 法划分干贝水分样本后，使用 PLSR 法建立的干贝水分含量预测模型都能得到较好的预测效果。然而，在使用 LS-SVM 建立干贝水分含量预测模型时，基于 CG 法的模型预测效果较差(见表 3 及图 4)，显示了建模方法与样本划分方法的相互作用。同一种样本划分方法不一定适用于所有的建模方法。因此，在建立干贝水分含量预测模型中，划分样本方法的优选需要结合建模方法进行讨论。

3 结 论

采用 RS，KS，SPXY 和 CG 四种常用的方法划分干贝水

分含量样本集，并以 PLSR，LS-SVM 两种常用的建模方法建立干贝水分含量预测模型，比较各模型的拟合效果和预测精度，得到以下结论：

(1)在建立干贝水分含量预测模型时，对于 PLSR 模型使用 RS 法划分干贝水分含量样本最为适宜(其 RPD 为 4.079 6)，对于 LS-SVM 模型使用 SPXY 划分法最为适宜(其 RPD 为 4.175 6)。划分样本方法的优劣与建模方法有关，其优选需要结合特定的建模方法进行。

(2)在 RS，KS，SPXY 和 CG 这四种常用的样本划分方法以及 PLSR，LS-SVM 这两种常用的建模方法中，使用 SPXY 法划分干贝水分含量样本配合 LS-SVM 或者 PLSR 建模方法所建立的干贝水分含量预测模型，与其他组合建立的预测模型相比，其预测效果和精度都较优。

References

[1] Huang H, Shen Y, Guo Y L, et al. Journal of Food Engineering, 2017, 205: 47.

[2] Emerson R W. Journal of Visual Impairment & Blindness, 2015, 109(2): 164.

[3] Claeys D D, Verstraelen T, Pauwels E, et al. The Journal of Physical Chemistry A, 2010, 114(25): 6879.

[4] Gani W, Limam M. Journal of Statistical Computation & Simulation, 2016, 86(1): 135.

[5] FU Miao-miao, LIU Mei-ying, NIU Zhi-you, et al(付苗苗，刘梅英，牛智有，等). Journal of Huazhong Agricultural University(华中农业大学学报), 2016, 2: 115.

[6] LIU Shan-mei, LI Xiao-yu, ZHONG Xiong-bin, et al(刘善梅，李小昱，钟雄斌，等). Transactions of The Chinese Society of Agricultural Machinery(农业机械学报), 2013, 44(s1): 165.

[7] Abdi H. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(1): 97.

[8] Shawe-Taylor J, Sun S. Neurocomputing, 2011, 74(17): 3609.

[9] Jian L, Shen S, Li J, et al. IEEE Transactions on Neural Networks and Learning Systems, 2016, (99): 1.

[10] Jiang L, Fei L, Yong H. Sensors, 2012, 12(3): 3498.

[11] Aydogdu M, Firat M. Water Resources Management, 2015, 29(5): 1575.

[12] LIU Jie, LI Xiao-yu, LI Pei-wu, et al(刘　洁，李小昱，李培武，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2010, 26(2): 338.

Research on Sample Division and Modeling Method of Spectrum Detection of Moisture Content in Dehydrated Scallops

HUANG Hui[1,2*], ZHANG De-jun[1], ZHAN Shu-yue[1], SHEN Ye[1], WANG Hang-zhou[1], SONG Hong[1], XU Jing[1], HE Yong[3]

1. Ocean College, Zhejiang University, Zhoushan 316021, China
2. Key Laboratory of Fishery Equipment and Engineering, Ministry of Agriculture, Shanghai 200092, China
3. College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China

Abstract Hyperspectral imaging technology has been used to establish the prediction model of moisture content in dehydrated scallops, and the model performance is affected by sample division method and modeling method. The method of sample division determines whether the selected sample is representative, and the modeling method determines how to use the sample to build the model, but the internal relationship between the sample division method and the modeling method has been rarely reported. It is important to explore the effects of different sample division methods and modeling methods on the prediction of the moisture content of scallops, and it can also provide reference for the study of spectral modeling of other samples. In this paper, the hyperspectral data of 270 scallops were extracted from spectral images captured by a hyperspectral imaging system in the 380～1 030 nm range. The samples were divided by RS, KS, SPXY and CG. The prediction models were established by PLSR and LS-SVM. The performance indexes of each model were calculated and compared. The results showed that the best sample division method is RS when using PLSR building prediction model (the RPD is 4.079 6) and SPXY is most suitable for LS-SVM model (the RPD is 4.175 6). The advantages and disadvantages of the division of the sample set are related to the modeling method, and the best choice should take modeling method into account. In this commonly used four sample division methods and two modeling methods, SPXY method is used to classify the sample set of moisture content and combine with LS-SVM method to optimize the effect and precision.

Keywords Hyperspectral data; Scallop; Sample division; Modeling method

(Received Aug. 12, 2017; accepued Dec. 25, 2017)

* Corresponding author