第42卷，第1期
2022年1月

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

Vol. 42, No. 1, pp282-291
January, 2022

一种基于无监督主动学习的苹果品质光谱无损检测模型构建方法

赵小康，赵 鑫，朱启兵*，黄 敏

江南大学轻工过程先进控制教育部重点实验室，江苏 无锡 214122

摘 要 利用光谱技术实现农产品、食品品质无损检测的实质是建立样本光谱信息与样本品质参数之间的机器学习模型。为了获得具有良好泛化性能的机器学习模型，通常需要大量的标记样本，然而，获取样本的光谱信息相对容易，但标注样本品质参数的过程往往涉及到大量的时间和经济成本，并且具有破坏性。主动学习是一种减少训练集有标记样本数量的方法，通过选择最有价值的样本进行标记，而不是随机选择。因此，主动学习能够控制向训练集添加哪些样本，模型不再是被动地接受用于建模的样本。在分类任务中已经提出较多关于主动学习的算法，但回归任务中的研究却相对较少，且现有的用于回归任务的主动学习算法大多是有监督的，即需要少量有标记样本训练初始模型。本文提出了一种基于无监督主动学习方法的训练样本选择策略。该方法首先通过层次凝聚聚类对无标记(标准值)光谱数据集进行多样性划分，获得不同的聚类簇；然后通过局部线性重建算法在每个聚类簇中选择最具代表性的样本构成训练样本集，最后基于训练集构建模型。利用两个年份三个品种苹果的近红外光谱数据，构建了其可溶性固形物含量和硬度的偏最小二乘预测模型，用于验证所提出方法的有效性。实验结果表明：所提出的方法要优于已有的样本选择策略，可以有效地提高模型精度，减少在模型训练中的破坏性理化实验。同时，与随机采样(RS)、Kennard-Stone算法(KS)、光谱-理化值共生距离算法(SPXY)这三种光谱领域常用的样本选择算法相比，该研究所提出的方法表现出了最佳的性能，基于所提出的无监督主动学习算法选取200个样本作为训练集所建立的可溶性固形物含量预测模型的预测均方根误差相对于其他三种算法降低了2.0%～13.2%，硬度预测模型的预测均方根误差相对降低了1.2%～15.7%。

关键词 光谱；品质检测；主动学习；训练样本选择

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)01-0282-10

引 言

光谱检测技术因其快速、无损等特点而广泛用于农产品、食品品质检测领域[1-5]。在利用光谱检测技术进行农产品、食品品质无损检测时，通常都需要一定数量的训练样本(包含光谱特征和理化品质指标)来构建预测模型。目前，已有多种建模方法被用于构建预测模型，例如：偏最小二乘回归模型(partial least square regression，PLSR)、支持向量回归模型(support vector regression，SVR)。在实际应用中，无论用何种建模方法构建光谱预测模型，预测模型的性能都严重依赖于训练样本的多样性和代表性。为了保证训练样本的多样性和代表性，人们往往需要获得大量的训练样本；但训练样本的品质指标(标签)多是通过破坏性理化实验获得，需要较高的时间和人力成本。相比于理化指标检验，样本的光谱信息获取较为容易。如果可以从大量的无标签样本(仅有光谱信息)中选取最有价值的样本进行标注，将有助于减少训练样本标注的盲目性，达到利用少量训练样本获得良好预测模型的目的。Kennard-Stone算法(KS)和光谱-理化值共生距离算法(SPXY)是光谱领域两种较为常见的样本选择方法。KS算法首先选择欧式距离最大的一组样本加入到训练集，然后依次选择一个样本，使已选样本与剩余样本的欧式距离最大，由于样本间的相似性通过欧式距离计算，其选择样本的空间分布易受离散点的影响，样本的代表性难以保证。而SPXY算法[2]在KS算法的基础上增加了对样本输出空间距离的考虑，因此需要获得样本的真实标签值。SPXY算法是一种有监督样本选择方法，在实际应用中仍然需要大量的理化分析，以获得样本标签值。

收稿日期：2020-12-18，**修订日期**：2021-03-22

基金项目：国家自然科学基金项目(61772240，61775086)资助

作者简介：赵小康，1995年生，江南大学物联网工程学院硕士研究生 e-mail：zhaoxk0211@163.com

*通讯作者 e-mail：zhuqib@163.com

主动学习是近年来提出的，综合考虑样本代表性、信息性或多样性的样本选择策略，已被广泛地运用于构建有监督分类模型。例如：王立国等[6]将主动学习算法用于高光谱图像分类任务中；唐金亚等[3]利用主动学习算法研究了玉米种子纯度分类模型的更新。但目前，主动学习在农产品、食品品质预测模型中的应用还鲜有报道。本文将结合农产品、食品品质无损检测的需要，提出了一种融合层次凝聚聚类(hierarchical agglomerative clustering，HAC)和局部线性重建算法(locally linear reconstruction，LLR)的无监督主动学习方法(HAC-LLR)。HAC-LLR 利用 HAC 聚类算法对原始光谱样本集进行聚类操作，以获得具有多样性的多个样本簇；针对不同的样本簇，通过 LLR 选取最具代表性的样本；最后基于选取的代表性样本及其理化指标，构建训练模型。实验结果表明，相比于已有算法，HAC-LLR 方法在训练样本数量相同的前提下，可以显著提高光谱模型的预测性能。

1　基于 HAC-LLR 的无监督主动学习方法

根据统计学习理论，要获得一个具有良好泛化性能的预测模型，用于构建预测模型的训练样本应该能够充分刻画整体样本的概率分布，即训练样本应该具有良好的代表性和多样性。代表性是指训练样本的概率分布应该能够代表整体样本的概率分布状态；而多样性是指训练样本应该尽可能地分布在整体样本空间，以实现整体样本空间的充分表达。多样性和代表性通常会存在一定的矛盾，为了解决这一矛盾，本文提出了 HAC-LLR 无监督主动学习方法，该方法首先对待选样本集进行聚类分析，获得多个样本簇；在不同簇中通过局部线性重建算法选出最具代表性的样本，从而使选择的样本兼具多样性和代表性。

1.1　基于层次凝聚聚类的样本集划分

聚类算法将数据集划分到不同子集中，使得子集内的数据相似度最大，子集间的数据相似度最小，从而可以发现数据中隐藏的模式和规律。本文利用无需预先设定聚类簇数的层次凝聚聚类方法对数据集进行聚类分析。层次凝聚聚类首先对数据集进行初始化，即将每个样本初始化为单独的簇，并计算两两簇之间的距离，然后寻找相距最近的两个簇进行归并，删除合并前的簇，保留新生成的簇，重复该过程，直到所有簇都归为一个大类[7]。整个聚类过程其实是建立一棵树，聚类结果可以根据最终生成的聚类树设置距离阈值，簇间距离大于设定值的不同簇即为期望得到的聚类结果。本文中，根据光谱数据特性，簇间距离采用相似性计算，簇间聚合方式为未加权平均距离法，根据生成的聚类树及聚类结果评价指标，距离阈值设定为 0.8。

1.2　基于局部线性重建算法的代表性样本选择策略

光谱数据多是高维数据，一个高维数据通常是由其低维潜在变量按照某种规则重建获得的。假设 $X=[X_1, \cdots, X_m]^{\mathrm{T}}$ 是已知的原始高维数据集，$Q=[q_1, \cdots, q_m]^{\mathrm{T}}$ 是与 X 同维的由低维潜在变量重建的数据集。LLR 算法认为已知数据集 X 应该与重建数据集 Q 具有相同的邻域表示关系。即对于任意一个样本 X_i，若其可以由其邻域 $N_p(X_i)$ 内(相邻数据点)的点线性表示为

$$X_i = \sum_{j \in N_p(X_i)} W_{ij} X_j + \varepsilon_i^X$$

其中，ε_i^X 为邻域关系表示误差，W_{ij} 为第 j 个样本点对第 i 个样本点的表示系数)；则对于与 X_i 对应的重建数据 q_i，存在相同的邻域表达关系 $q_i = \sum_{j \in N_p(X_i)} W_{ij} q_j + \varepsilon_i^q$。$W_{ij}$ 可以由原始数据集 X 的表示误差最小化获得

$$\begin{aligned} &\min \sum_{i=1}^{m} \| X_i - \sum_{j=1}^{m} W_{ij} X_j \|^2 \\ &\text{s. t.} \quad \sum_{j=1}^{m} W_{ij} = 1,\ i = 1, \cdots, m \\ &W_{ij} = 0 \quad \text{if} \quad X_j \notin N_p(X_i) \end{aligned} \tag{1}$$

根据经验值，将邻域 $N_p(X_i)$ 的样本个数设置为 20。利用式(1)获得 W_{ij} 后，在原始数据集 X 中选择 k 个最具代表性的样本点 $\{x_{s_1}, x_{s_2}, \cdots, x_{s_k}\} \subseteq X$，意味着不仅要使选择的样本点自身要有小的重建误差，而且要使重构样本集 Q 具有较小的邻域关系表示误差 ε_i^X。即具有如下的最小化损失函数

$$\varepsilon(q_1, \cdots, q_m) = \sum_{i=1}^{k} \| q_{s_i} - x_{s_i} \|^2 + \mu \sum_{i=1}^{m} \| q_i - \sum_{j=1}^{m} W_{ij} q_j \|^2 \tag{2}$$

式(2)中，μ 是惩罚系数，用于调节重建误差和重构样本 Q 的邻域关系表示误差。本文中设置为 0.1。

定义 Λ 为 $m \times m$ 的对角矩阵，如果 $i \in \{s_1, \cdots, s_k\}$，则对角元素为 $\Lambda_{ii}=1$，否则 $\Lambda_{ii}=0$。则目标函数(2)可以重新被写成如式(3)矩阵形式

$$\varepsilon(Q) = Tr((Q-X)^{\mathrm{T}} \Lambda (Q-X)) + \mu Tr(Q^{\mathrm{T}} M Q) \tag{3}$$

式(3)中，$M=(I-W)^{\mathrm{T}}(I-W)$，$I$ 为单位对角阵，Tr 为矩阵求迹运算。式(3)最小化，则重建结果可以表示为

$$Q = (\mu M + \Lambda)^{-1} \Lambda X \tag{4}$$

对于原始样本点 $x_1, \cdots, x_m$ 和样本点重建结果 $q_1, \cdots, q_m$，重建误差可以表示如式(5)

$$\| X - Q \|_F^2 = \| (\mu M + \Lambda)^{-1} \mu M X \|_F^2 \tag{5}$$

式(5)中，重建误差只与所选择的点 $\{s_1, \cdots, s_k\}$ 有关，因此，最具代表性的点可以定义为那些能够最小化重建误差的点，即如果所选样本点确定，可以更准确地重建整个原始数据集。式(5)可以通过迭代求解策略获得，其详细计算过程见参考文献[8]。

1.3　基于 HAC-LLR 训练样本选择策略的光谱检测方法流程

基于 HAC-LLR 训练样本选择策略的光谱检测方法流程主要包括：(1)利用层次凝聚聚类对大量的无标记光谱数据集进行聚类分析，根据生成的聚类树和设定的簇间距离阈值划分出不同的数据簇；(2)针对每个数据簇，利用局部线性重建算法，选取一定数量的待标记样本(该簇样本数量占样本总数的比例乘以期望选出样本的总数 k 即为每个簇应选出的样本数)，从所有的簇中总共选出设定的 k 个样本；(3)对选出的样本根据具体检测指标，进行理化分析，获得其标签值 Y，构建训练集样本对 $(X_i, Y_i)_{i=1, \cdots, k}$；(4)利用训练集样本，训练输出模型；(5)利用模型对预测集样本进行预测。图

1 给出了算法的流程示意图。

图 1 基于 HAC-LLR 训练样本选择策略的光谱检测方法流程图

Fig. 1 Flow chart of spectral detecting method based on HAC-LLR training samples selecting strategy

2 实验部分

实验样本是美国密歇根州立大学克拉克斯维尔园艺实验站果园提供的 Golden Delicious(GD)，Jonagold(JG)和 Red Delicious(RD)三个品种的苹果，采收于 2009 年和 2010 年连续两个年份。样本的光谱数据通过微型 Vis-SWNIR 光谱仪(S400，Ocean Optics，Dunedin，FL)采集。Vis-SWNIR 光谱仪的光谱范围为 460～1 100 nm，光谱分辨率为 1 nm，每个光谱样本有 641 个变量。获得光谱数据之后，使用质地分析仪(型号 TA. XT2i，Stable Micro Systems，Inc.，Surrey，UK)和数字折射仪(型号 PR-101，Atago Co.，Tokyo，Japan)在光谱仪测量的位置对苹果的硬度和可溶性固形物(soluble solid content，SSC)进行测量。实验设备和数据的更详细信息参见文献[9]。

表 1 给出了实验样本的 SSC 和硬度统计数据表。由表 1 可以看出，SSC 和硬度的分布范围较大，可以充分验证模型的性能。图 2 为不同年份、不同种类苹果样本的平均光谱。从图中可以看出，不同年份、不同种类的苹果光谱存在着较大差异，难以用一个单一模型进行建模，需要对不同年份、不同种类的苹果构建多个模型。

表 1 苹果样本的品质参数统计信息

Table 1 Statistics of quality reference for apple samples

收获年份	品种	数量	SSC/%				硬度/N			
			均值	方差	最小值	最大值	均值	方差	最小值	最大值
2009	GD	1 070	14.814	1.484	9.700	18.800	55.549	14.800	30.431	98.893
	JG	874	12.719	1.059	9.400	16.400	62.951	22.173	29.319	110.187
	RD	1 078	11.558	1.160	8.100	15.400	58.911	15.099	29.497	89.980
2010	GD	1131	13.232	1.270	9.600	17.500	63.280	16.698	29.434	100.994
	JG	1 087	12.545	1.532	8.700	17.300	63.405	19.015	27.700	105.340
	RD	1 035	12.252	1.454	8.800	15.500	69.068	16.654	29.978	107.428

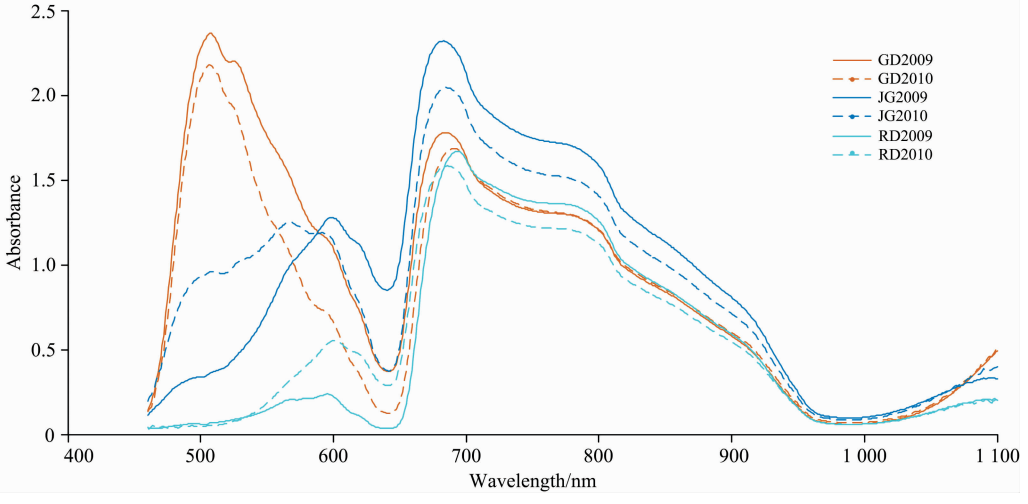


图 2 连续两年采收的三种苹果的平均光谱

Fig. 2 The average spectra of three cultivars apple samples harvestee from two years

3　结果与讨论

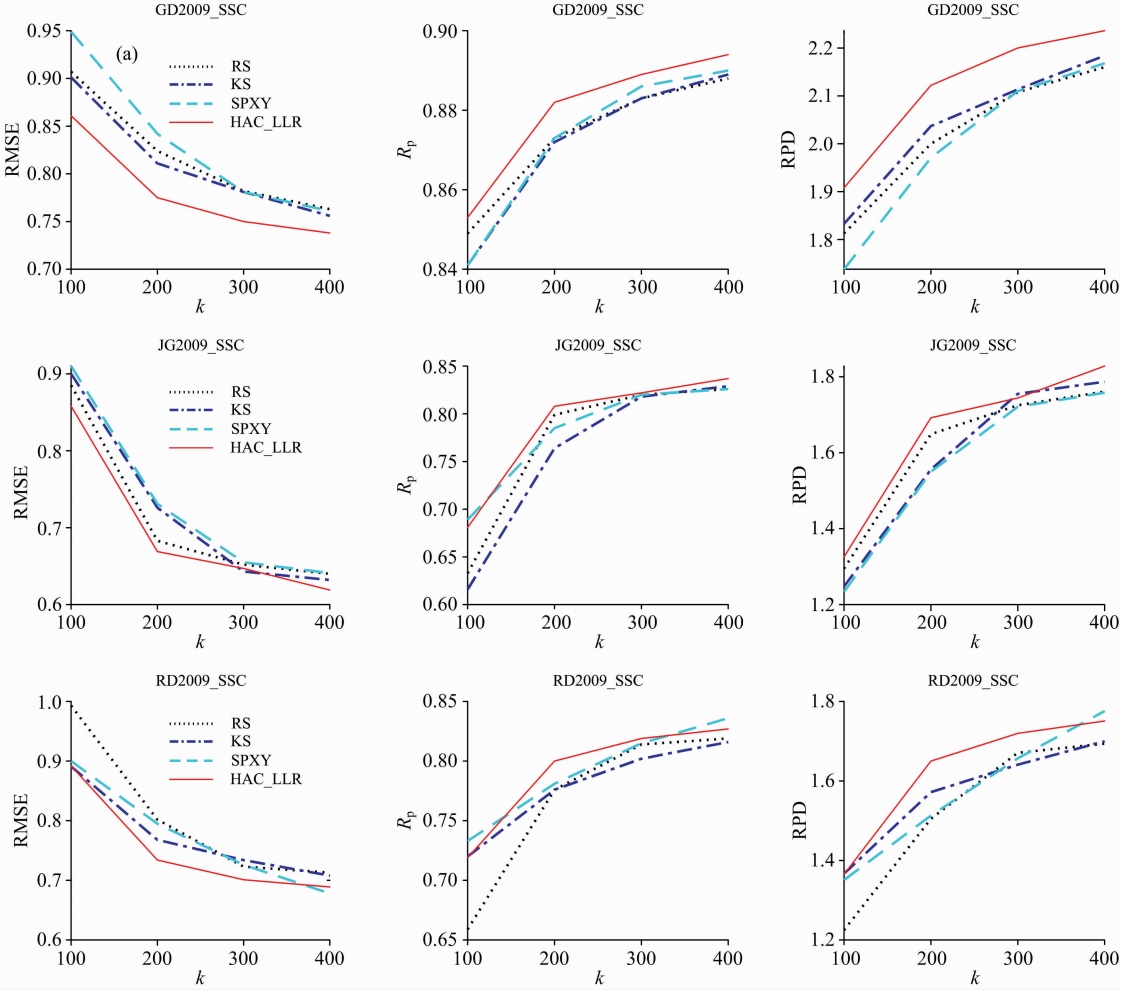
3.1　基于 HAC-LLR 训练样本选择策略的苹果品质检测模型的建立

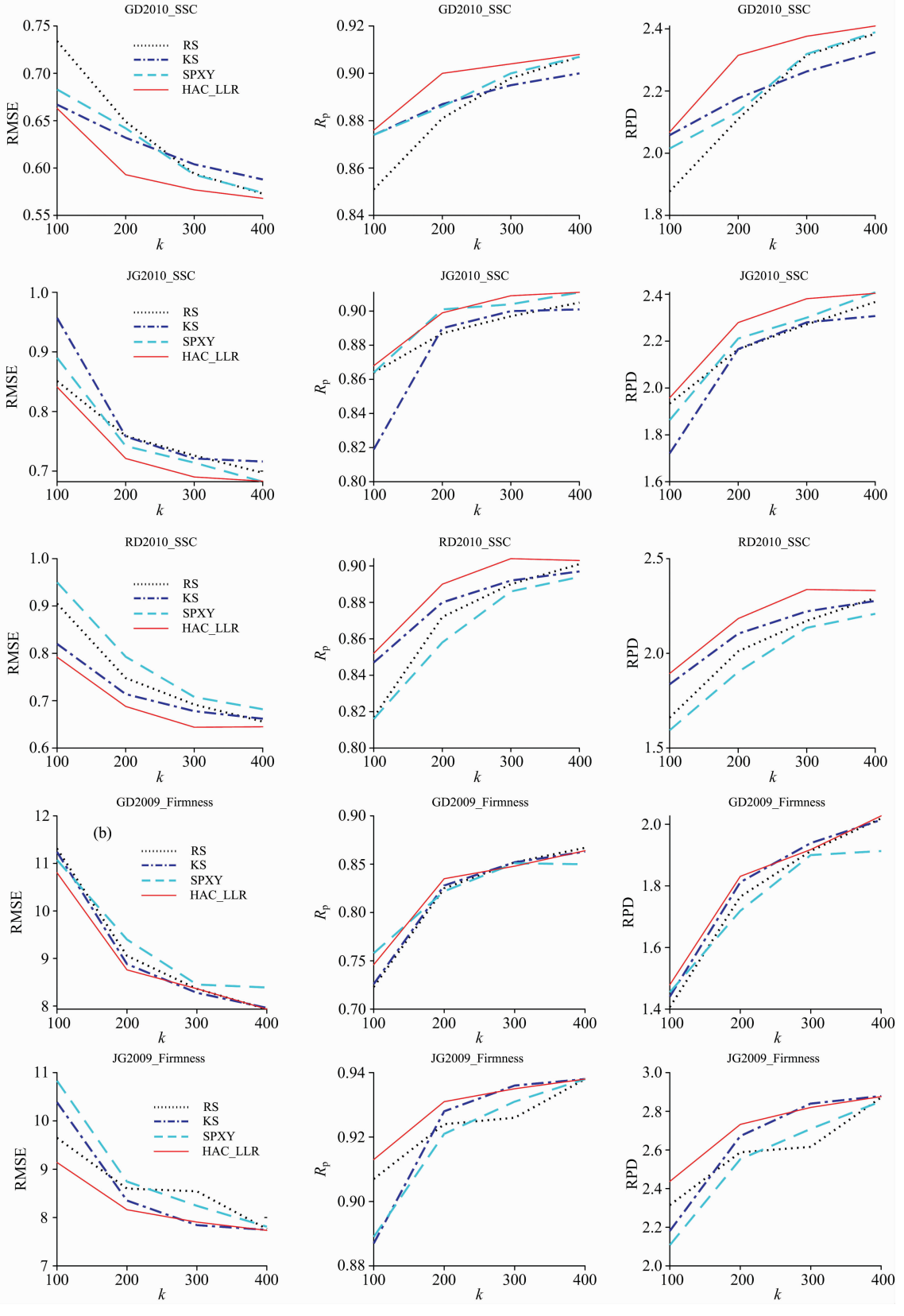
基于无监督主动学习算法选取一定数量的样本用于建立苹果品质检测模型。为充分验证基于无监督主动学习算法的模型性能，针对每个数据集，首先随机选取 100 个未标记样本作为预测集，其余未标记样本作为样本选择池。基于该样本选择池，分别利用随机采样(RS)、Kennard-Stone 算法(KS)、光谱-理化值共生距离算法(SPXY)和本文提出的 HAC-LLR 样本选择策略，选出一定数量的样本作为训练集，用于训练 PLSR 模型。利用预测集均方根误差(RMSE)、相关系数(R_p)和残留预测偏差(residual prediction deviation，RPD)评估最终的模型性能。为了减少预测集样本随机选取对实验结果的影响，每次实验过程随机重复 5 次，5 次随机实验的平均值作为最终结果。考虑到每个光谱样本有 641 个变量，为了避免模型的过拟合，利用竞争自适应重加权采样算法[10](competitive adaptive reweighted sampling，CARS)对原始特征进行筛选，其中，105，120，82，94，131，106，125，90，96，112，103 和 120 个特征变量分别作为 GD2009，GD2010，JG2009，JG2010，RD2009 和 RD2010 的 SSC 和硬度 PLSR 模型的输入。PLSR 模型的最佳主元数量通过 10 折交叉验证确定。

PLSR 建模和光谱数据分析软件分别是 PLS 工具箱(Eigenvector Research，Inc.，Wenatchee，WA，USA)和 MATLAB R2014a(The MathWorks，Inc.，Natick，MA，USA)。

3.2　基于不同样本选择算法的建模结果比较

对于不同数据集，按照与预测集 1∶1，2∶1，3∶1 和 4∶1 的比例划分，四种算法分别选取 100，200，300 和 400 个样本作为训练集，用于建立 PLSR 模型。图 3 给出了不同数据集下 PLSR 模型的预测结果。从图 3 中可以看出，随着训练集样本数量的增加，四种样本选择算法建立的模型性能都有所提高(RMSE 值降低、R_p 和 RPD 值增高)。相比于其他三种算法，本文提出的无监督主动学习算法表现出了最佳的预测性能，特别是在建模集样本数量较少的情况下。当建模集样本数量较多时，不同样本选择算法选出的样本共性较大，模型也趋于稳定，主动学习方法的优势也会逐渐减弱。同一品种不同年份的苹果样本所对应的模型性能也表现出了一定差异，进一步验证了需要对不同年份、不同品种的苹果构建多个模型的设想。另外，四种算法分别选出 200 个样本所建立模型的预测性能如表 2、表 3 所示，基于 HAC-LLR 的 SSC 模型相对于基于 RS，KS 和 SPXY 的 SSC 模型预测结




GD2010_SSC
JG2010_SSC
RD2010_SSC
GD2009_Firmness
JG2009_Firmness
(b)
RS
KS
SPXY
HAC_LLR
RMSE
R_p
RPD
k

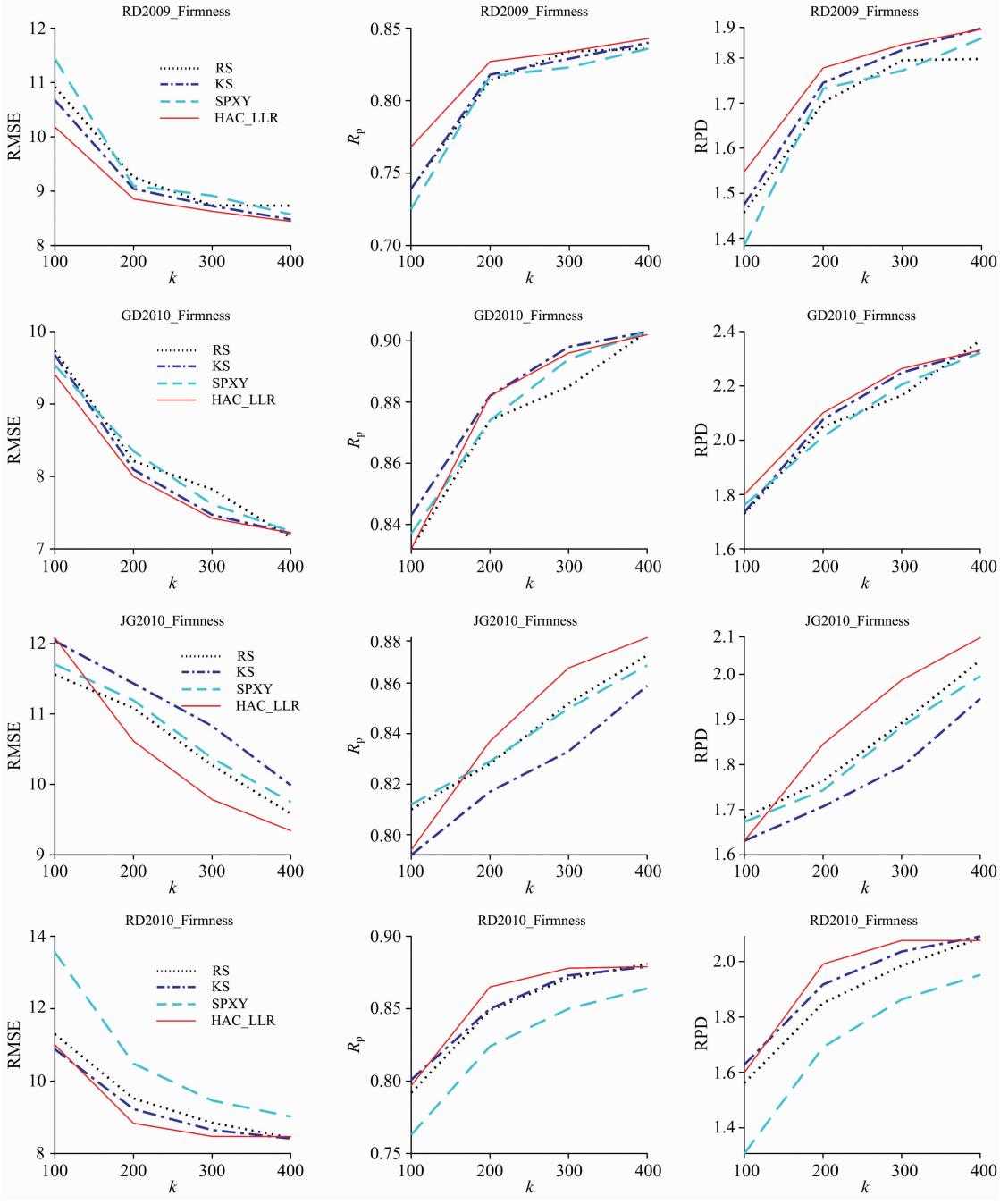


图 3　不同数据集下基于不同样本选择算法的 SSC（a）和硬度(b)的 PLSR 模型预测结果

Fig. 3　PLSR mdoel prediction results of SSC（a）and firmness（b）based on different sample selection algorithms under different datsets

果的 RMSE 值分别降低了 2.0%～8.6%，3.6%～7.9%和 2.8%～13.2%，对于硬度模型，RMSE 值相应地分别降低了 2.6%～7.2%，1.2%～7.2%和 2.6%～15.7%。

为了比较不同算法性能的统计学意义，本文进一步利用参考文献[11]定义的曲线下面积(area under curve，AUC)作为综合性能度量指标对模型的 RMSE，R_p 和 RPD 进行分析(图 4 所示)。本文使用 RS 算法的 AUC 值对其他三种算法进行标准化，因此 RS 算法的 AUC 值始终为 1。对于 RMSE 值而言，较小的 AUC 值代表较高的模型性能，对于 R_p 值和 RPD 值而言，较高的 AUC 值代表较高的模型性能。从图 4 可以看出，基于本文提出的 HAC-LLR 训练样本选择策略所建立的模型，预测无标记样本的 AUC-RMSE 值更低，AUC-R_p 值和 AUC-RPD 值更高。

表 2 四种算法分别选出 200 个 2009 年的样本所建立 PLSR 模型的预测结果

Table 2 The prediction results of PLSR models based on 200 samples from 2009 selected by four algorithms respectively

品种	年份	算法	预测集							
			SSC				硬度			
			RMSE	R_p	RPD	Dif. /%	RMSE	R_p	RPD	Dif. /%
GD	2009	RS	0.824	0.873	2.00	5.9	9.057	0.825	1.765	3.3
		KS	0.811	0.872	2.037	4.4	8.883	0.828	1.812	1.4
		SPXY	0.842	0.873	1.971	8.0	9.399	0.822	1.719	6.8
		HAC-LLR	0.775	0.882	2.122		8.760	0.835	1.831	
JG	2009	RS	0.683	0.799	1.649	2.0	8.601	0.924	2.588	5.1
		KS	0.726	0.764	1.556	7.9	8.353	0.928	2.672	2.3
		SPXY	0.731	0.785	1.550	8.5	8.746	0.921	2.552	6.7
		HAC-LLR	0.669	0.808	1.692		8.164	0.931	2.732	
RD	2009	RS	0.802	0.776	1.505	8.5	9.257	0.814	1.702	4.3
		KS	0.768	0.776	1.572	4.4	9.043	0.818	1.745	2.1
		SPXY	0.795	0.781	1.513	7.7	9.097	0.817	1.732	2.6
		HAC-LLR	0.734	0.800	1.650		8.856	0.827	1.778	

Dif. /%：基于 HAC-LLR 算法的 PLSR 模型和基于其他三种算法的 PLSR 模型的 RMSE 值百分比差异。

表 3 四种算法分别选出 200 个 2010 年的样本所建立 PLSR 模型的预测结果

Table 3 The prediction results of PLSR models based on 200 samples from 2010 selected by four algorithms respectively

品种	年份	算法	预测集							
			SSC				硬度			
			RMSE	R_p	RPD	Dif. /%	RMSE	R_p	RPD	Dif. /%
GD	2010	RS	0.649	0.881	2.110	8.6	8.214	0.874	2.050	2.6
		KS	0.632	0.887	2.177	6.2	8.096	0.882	2.076	1.2
		SPXY	0.642	0.886	2.133	7.6	8.349	0.874	2.013	4.2
		HAC-LLR	0.593	0.900	2.315		8.000	0.882	2.101	
JG	2010	RS	0.759	0.887	2.165	5.0	11.083	0.828	1.764	4.2
		KS	0.758	0.890	2.166	4.9	11.433	0.817	1.707	7.2
		SPXY	0.742	0.901	2.211	2.8	11.194	0.829	1.743	5.2
		HAC-LLR	0.721	0.899	2.279		10.615	0.837	1.845	
RD	2010	RS	0.748	0.872	2.011	8.0	9.524	0.849	1.850	7.2
		KS	0.714	0.880	2.104	3.6	9.229	0.850	1.917	4.2
		SPXY	0.793	0.858	1.902	13.2	10.486	0.824	1.691	15.7
		HAC-LLR	0.688	0.890	2.184		8.837	0.865	1.991	

Dif. /%：基于 HAC-LLR 算法的 PLSR 模型和基于其他三种算法的 PLSR 模型的 RMSE 值百分比差异。

RS 算法选择的样本具有较强的随机性，相应的模型性能有很强的不确定性。KS 算法考虑到了样本光谱信息的欧氏距离，由于光谱数据的高维性，欧氏距离不能很好地表征样本间的真实距离和相似性[8, 12]，但整体性能优于 RS 算法和 SPXY 算法。SPXY 算法基于 KS 算法，虽然增加了对输出空间距离的考虑，即需要使用到样本真实理化标签值，属于有监督的样本选择算法，但是对输出空间的度量仅仅基于不同真实标签的差值，因此整体性能上没有表现出优势，甚至在很多数据集上不及 KS 算法。而本文提出的无监督主动学习方法由于综合考虑了样本的多样性和代表性，因此表现出了最佳性能。综合多个评价指标以及实验结果，验证了本文提出的无监督主动学习方法的有效性。

4 结 论

建立一个精确的且具有良好泛化能力的回归模型通常需要大量的带标记的训练集样本。然而，在样本制备过程中，采集样本的光谱数据是相对容易的，获得样本的真实标记却是费时费力且具有破坏性的。常规的光谱学实验设计中无法充分利用已知样本的信息，使得基于不同训练集的模型的性能相差较大。主动学习是一种选择最有价值的未标记样本进行标记的方法，以少量标记样本建立更好的回归模型。本文提出了一种无监督的主动学习方法，该方法融合了样本多样性和代表性两种选择标准，在连续两年采收的三个品种苹果

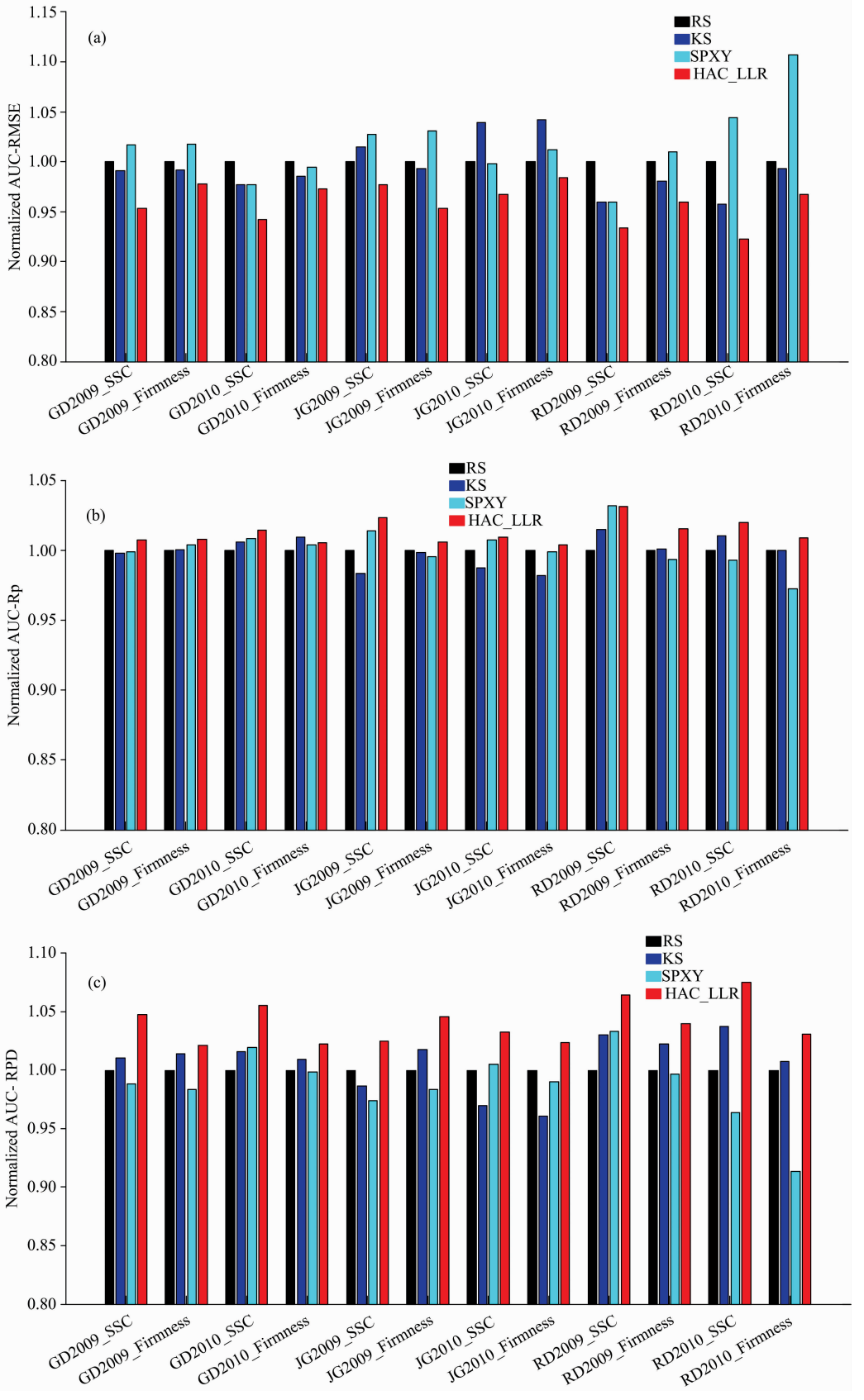


图 4　不同数据集上归一化的 AUC-RMSE(a)，AUC-R_p(b)和 AUC-RPD(c)

Fig. 4　Normalized AUCs of the RMSE (a), the R_p (b) and the RPD (c) on different datasets

的光谱数据集上进行了大量的实验，实验结果验证了所提出的无监督主动学习方法的有效性，为有效减少训练集样本数量、降低破坏性理化实验所带来的成本消耗、提高模型精度提供了一种解决方案。由于本文所提方法考虑的是模型构建中的训练样本选择，因此，同样适用于构建非线性模型。此外，迁移学习和主动学习都可以用于处理标记样本不足的问题，今后我们还将研究如何融合主动学习和迁移学习的思想用于减少光谱分析领域训练集样本的制备。

References

[1] Li X N, Huang J C, Xiong Y J, et al. Computers and Electronics in Agriculture, 2018, 155: 23.

[2] MAO Bo-hui, SUN Hong, LIU Hao-jie, et al(毛博慧，孙 红，刘豪杰，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2017, 48(S1): 160.

[3] TANG Jin-ya, HUANG Min, ZHU Qi-bing(唐金亚，黄 敏，朱启兵). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2015, 35(8): 2136.

[4] GUO Wen-chuan, ZHU De-kuan, ZHANG Qian, et al(郭文川，朱德宽，张 乾，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2020, 51(9): 350.

[5] MA Wen-qiang, ZHANG Man, LI Yuan, et al(马文强，张 漫，李 源，等). Chinese Journal of Analytical Chemistry(分析化学), 2020, 48(12): 1737.

[6] WANG Li-guo, SHANG Hui, SHI Yao(王立国，商 卉，石 瑶). Journal of Harbin Engineering University(哈尔滨工程大学学报), 2020, 41(5): 731.

[7] DAI Xiang, HUANG Xi-feng, TANG Rui, et al(代 翔，黄细凤，唐 瑞，等). Journal of South China University of Technology · Natural Science Edition(华南理工大学学报·自然科学版), 2019, 47(8): 84.

[8] Zhang L J, Chen C, Bu J J, et al. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 2026.

[9] Mendoza F, Lu R F, Cen H Y. Postharvest Biology and Technology, 2012, 73: 89.

[10] Li H D, Liang Y Z, Xu Q S, et al. Analytica Chimica Acta, 2009, 648(1): 77.

[11] LIU Zi-ang, JIANG Xue, WU Dong-rui(刘子昂，蒋 雪，伍冬睿). Aata Automatica Sinica(自动化学报), https://doi.org/10.16383/j.aas.c200071.

[12] YAN Yue, ZHANG Hong-guang, LU Jian-gang, et al(鄢 悦，张红光，卢建刚，等). Computers and Applied Chemistry(计算机与应用化学), 2017, 34(5): 351.

A Model Construction Method of Spectral Nondestructive Detection for Apple Quality Based on Unsupervised Active Learning

ZHAO Xiao-kang, ZHAO Xin, ZHU Qi-bing*, HUANG Min

Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China

Abstract The essence of using near-infrared spectroscopy to realize non-destructive detection of agricultural products and food quality is to establish a machine learning model between sample spectral information and sample quality parameters. In order to obtain a machine learning model with good generalization performance, a large number of labeled samples are usually required. However, it is relatively easy to obtain spectral information of samples, but labeling samples quality parameters often involves a large amount of time and economic costs and is destructive. Active learning is a method to reduce the number of labeled samples in training set by selecting the most valuable samples for labeling instead of random selection. Therefore, active learning can control which samples are added to the training set, and the model no longer passively accepts samples for modeling. There have been many active learning algorithms in classification tasks. There are relatively few researches in regression tasks. Moreover, most of the existing active learning algorithms for regression tasks are supervised. That is, a small number of labeled samples are needed to train the initial model. In this paper, a training sample selection strategy based on unsupervised active learning is proposed. Firstly, the method divides the diversity of unlabeled (standard value) spectral datasets through hierarchical agglomerative clustering to obtain different clustering clusters. Then, the locally linear reconstruction method selects the most representative samples in each clustering cluster to form a training sample set and establish the partial least squares regression model based on the training set to predict the unlabeled samples. In this paper, partial least squares prediction models for soluble solids content and firmness prediction were constructed to evaluate the proposed method's performance, using the near infrared

spectrum data of three varieties of apples from two years. The experimental results show that the method proposed in this paper is superior to the existing sample selection strategy, which can effectively improve the model accuracy and reduce destructive physical and chemical experiments in model training. Meanwhile, compared with random sampling (RS), traditional Kennard-Stone (KS) and joint x-y distances (SPXY), the proposed method achieved the optimal performance. The root mean square error of the soluble solid content prediction models based on the unsupervised active learning algorithm proposed in this paper, which selects 200 samples as the training set, is reduced by 2.0%～13.2% compared with the other three algorithms, and the root means square error of the firmness prediction models is reduced by 1.2%～15.7%.

Keywords Spectroscopy; Quality detection; Active learning; Training sample selection

(Received Dec. 18, 2020; accepted Mar. 22, 2021)

* Corresponding author