

课题类别：广东省普通高校青年创新人才项目

课题名称：基于网络学习的偏最小二乘加核优化方法

课题批准号：2019KQNCX213

项目负责人：洪绍勇

所在单位：广州华商学院数据科学学院

主要成员：石金诚、李远飞、颜远海、杨春

一、项目研究的目的和意义

光谱分析是利用测量样品得到的光谱信息，结合样品成分含量来对样品进行定量分析的快速检测技术。与传统化学检测相比，NIR 光谱分析技术具有无试剂、微破坏、耗时短、成本低等特点，并且能够实现多种成分同时检测，在应用上有很大的优势，目前已经在农业、食品、环境、生物医学、石油化工等领域得到广泛的应用。由于采集的近红外光谱数据是高维复杂样本数据，要对高维数据建立有效的光谱分析模型，实现光谱定标预测，需要研究有效的化学计量学方法，在降低数据维度的同时能够消除（或减少）噪音干扰，这其中存在很多具有挑战性的数学问题。通过基于网络学习的偏最小二乘加核优化方法的研究，可以达到以下目的：

1. 通过坐标变换提取得分矩阵和载荷矩阵过程中采用迭代优化模式，找到能够更好地确定 PLS 因子数 F 的最优化方法
2. 研究加核 PLS 算法中的不同核函数对算法机制的影响，通过网络学习模式提取光谱数据中的特征信息，使得改进 PLS 算法能够应用于非连续非线性光谱数据建模分析。
3. 通过对得分矩阵与载荷矩阵的推导、计算和关联运用，研究基于网络学习模式的核函数参数优化机制，使改进的 PLS 算法在面向实际数据的应用中能够提高运算效率。

二、研究成果的主要内容和创新之处

（一）主要内容

1. 利用 PCA 算法思想提取潜变量，并将潜变量提取过程与 SVM 和 ELM 进行联合优化，形成操作方便的新型数据优化分析模型。

建立 LV-SVM 模型进行训练，调试前 30 个潜变量，结合 SVM 学习过程进行联合优化，设置正则化参数的调整范围为 $\gamma = 1, 2 \dots 20$ ，将每一个参数组合所对应的模型应用于 45 个验证集样本蛋白质含量的预测，通过比较不同潜变量个数 (LV)、不同正则化参数 (γ) 取值，依据模型评价指标 (RMSEV) 确定建模优化参数。

利用 LV-ELM 模型针对定标集样本进行训练，基于全谱数据提取潜变量 LX，调试潜变量数量为 1, 2... 30，结合 ELM 的学习优化过程，设置 SLFN 网络的隐含层节点数量可变，调试取值为 $k \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ ，通过反馈式迭代确定各个隐含节点的参数，利用最小二乘回归计算 SLFN 隐含层至输出层的权值 β ，进而完成对验证集样本的蛋白质含量预测。通过比较不同潜变量个数 (LV)、不同隐含层节点个数 (k) 的取值，依据 RMSEV 确定建模优化参数。

2. 运用加核偏最小二乘回归建立了柚子果实中 OA 含量的 FT-NIR 定标模型。原始数据通过一个特定的非线性核函数映射到一个高维的特征空间中，这样原始的线性 PLS 算法就可以用来发现特征数据和目标分析物之间的关系。传统的映射是使用常用的核函数，如线性函数、多项式函数、高斯径向基函数和 sigmoid 函数。这些通用内核需要建立模型，并调整其内置参数。这种局限性使现代智能机器学习分析中模型优化的灵活性大大降低。在这种情况下，网络结构能够提高模型的优化性能，因为网络连接权重可以以数据驱动的方式进行训练。带神经节点调谐的全连接分层网络是启动深度学习定标模型的普遍方法。网络由三层构成。隐含层采用可调神经元节点数设计，提取光谱特征变量，优化核 PLS 模型。预先初始化连接权值，然后通过模型预测误差的迭代反馈自动调整连接权值

(二) 创新之处

1. 研究 SVM 和 ELM 两种机器学习方法针对速溶咖啡粉末的蛋白质快速定量检测的 NIR 光谱建模分析，提出利用 SVM 和 ELM 方法结合潜变量技术进行建模。利用 PCA 算法思想提取潜变量，并将潜变量提取过程与 SVM 和 ELM 进行联合优化，形成操作方便的新型数据优化分析模型。

2. 设计了一种以机器学习方式提取光谱特征数据的神经网络结构内核。结合 PLS 潜变量的优化选择，对网络隐节点数进行了调谐。采用误差反馈迭代机制

对网络链接权值进行自适应训练。

三、取得的主要成果

1. Hong shaoyong. Quantitative analysis of organic acids in pomelo fruit using FT-NIR spectroscopy coupled with network kernel PLS regression[J]. Infrared Physics & Technology, 2021, 112 : 103582- (SCI 2 区)

2. 洪绍勇. 潜变量机器学习方法在咖啡 NIR 定量分析中的应用[J]. 光谱学与光谱分析, 2021, 41 (05) :1441-1445. (SCI 3 区)

四、成果的应用价值

我们常用的核函数需要通过建立模型来调整其内置参数,这种局限性使现代智能机器学习分析中模型优化的灵活性大大降低。网络结构优化及其连接权值的自动调整有望成为一种在光谱分析领域引入智能机器学习方法的现代媒介,用于在线检测任何目标成分。将潜变量提取过程与 SVM 和 ELM 进行联合优化,形成操作方便的新型数据优化分析模型,能够为 NIR 快速检测技术提供良好的建模分析手段。这些方法的应用,使得近红外光谱技术有望作为一种智能技术推广应用于食品品质的快速、准确评价。