变分模态分解展开极限学习机用于复杂样品光谱定量分析

Variational Modal Decomposition Unfolded Extreme Learning Machine for Spectral Quantitative Analysis of Complex Samples

沈亮亮，刘强，吴德云，刘鹏，卞希慧*

（天津工业大学化学工程与技术学院，中国天津 300387）

摘要：构建一个对未知样品准确预测的多元校正模型是至关重要的。传统的多元校正方法在原始光谱和目标值之间建立单一模型，是对整体信息的一个大致反映。数学分解方法将信号从原始数据空间转换到其他空间，揭示原始光谱中可能被掩盖的信息，提高校正过程的准确性和可靠性。本研究基于变分模态分解（Variational mode decomposition, VMD）和极限学习机（Extreme learning machine, ELM）的优势，提出了一种变分模态分解展开极限学习机方法（Variational mode decomposition unfolded extreme learning machine, VMD-UELM）。利用VMD 将每个光谱分解为多个模态分量，将模态分量展开成一个扩展矩阵，ELM 用于建立定量模型。将其应用于血液、燃料油和二元中药数据集的光谱中，并分别对血红蛋白、双环芳烃和三七的含量进行定量分析。与 ELM 和 PLS 相比，该方法有更低的 RMSEP 和更高的相关系数 R，表明了其在定量分析中具有更高的准确性。

关键词：变分模态分解；多元校正；定量分析；集成建模

1 引言

复杂样品由于成分复杂多样，对其进行定量分析是一项具有挑战性的任务。发展快速、准确且可靠的分析技术对于复杂样品的定量分析具有重要意义。光谱分析结合化学计量学方法为已在食品、医药、石油、环境等领域中广泛应用，为复杂样品的定量分析提供了有效的工具[1-4]。其中，建立高质量的多元校正模型，是光谱定量分析的核心所在[5,6]。多元校准方法通过建立原始光谱数据与目标值之间的单一模型来确定无需化学分离的多组分含量。目前提出的多元校准方法主要有多元线性回归（multiple linear regression, MLR）[17]、主成分回归（principle component regression，PCR）[18]、偏最小二乘回归（partial least squares regression, PLSR）[19,20]等线性算法以及人工神经网络（artificial neural network,

* Corresponding author.

ANN）[21]、支持向量回归（support vector regression, SVR）[23]、极限学习机（Extreme learning machine, ELM）等非线性算法。相比与其他的算法，极限学习机（Extreme learning machine，ELM）作为一种高效的多元校正算法，拥有线性及非线性算法的优点，同时以其快速的学习速度、简洁的结构及出色的泛化性能而备受关注。

ELM 是由黄光斌教授提出的一种基于单隐层前馈神经网络的新算法[7]。与传统的学习算法不同，ELM 除了最小化训练误差外，还倾向于实现输出权重的最小范数。前馈神经网络达到最小误差的泛化性能随着输出权重范数的降低而提高。然而，在应用 ELM 进行复杂样品的光谱数据分析时，光谱信号可能会受到背景噪音等干扰信息的影响，因此在对目标组分进行分析之前，对光谱信号进行数学分解是非常必要。数学分解可有效地从复杂的光谱信号中提取所需信息，揭示原始光谱中可能被掩盖的有用信息，提高校正过程的准确性和可靠性。

变分模分解（variational mode decomposition，VMD）是一种基于坚实数学基础的新型信号分解技术[8]。它可以将非线性和非平稳信号分解为具有特定稀疏性特性的模态分量集合。VMD 不仅抑制了模态混叠和端点效应，而且可以很好地将噪声与信号分离，从而获得更准确的预测结果并提高效率[9]。如股票价格预测[10]、风速预测[11]和故障诊断[12]。然而，关于在化学计量学中使用 VMD 算法对复杂样品进行紫外-可见光谱定量分析的报道却很少。卞教授[13]等人首次提出了用于复杂样品测定的变分模态分解加权多尺度支持向量回归法。在这种方法中，对所有子模型的预测结果进行加权平均，从而得到最终的预测结果。虽然加权平均是一种常用的模型组合方法，但权重的确定仍然是一个问题。展开策略是将所有训练子集沿变量方向展开，而不是在此基础上创建多个单独的子模型[14]。它只建立一个模型，而不会引入多个模型的偏差。展开策略不仅充分利用了光谱信息，而且避免了模型权重的确定。

基于 VMD 和 ELM 的优势，提出了变分模态分解展开极限学习机（Variational mode decomposition unfolded extreme learning machine, VMD-UELM）的建模方法。

VMD-UELM 首先对每个光谱进行变分模态分解，获得 K 个模态分量，然后将这些模态分量展开为一个扩展矩阵。在构建模型之前，为了确保模型的性能，需要对 VMD-UELM 的关键参数进行优化。参数包括模态分量数、激励函数以及隐藏节点的数量。利用该方法对血液中的血红蛋白、燃料油中的单环芳烃和二元中药中的三七组分含量进行了预测。为了评估模型的预测性能，计算预测均方根误差（root mean square error of prediction，RMSEP）和相关系数（correlation coefficient，R），并与 PLS、ELM 进行比较。

2 算法原理

2.1 变分模态分解（VMD）

VMD 是一种新的分解技术，具有良好的理论基础，能有效处理复杂信号。它将信号分解为多个带有不同频率$\{\omega_k\}=\{\omega_1,\ldots,\omega_K\}$的离散模态分量$\{u_k\}=\{u_1,\ldots,u_K\}$。整个算法可分为两部分：变分问题的构造和求解。在构造部分，基本原理是将与模态分量有关的变分问题转化为模态函数，该函数寻求最小化估计带宽的总和。此外，各个模态分量的和保持等于原始输入信号 X。变分问题的约束可以表示为：

$$\begin{cases} \min\limits_{\{u_k,\omega_k\}}\{\sum\limits_k \left\|\partial_t[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)]e^{-j\omega_k t}\right\|_2^2\} \\ s.t.\sum\limits_k u_k=\mathbf{X} \end{cases} \tag{1}$$

其中δ是脉冲函数，$e^{-j\omega_k t}$是每个频谱信号的估计中心频率，*是卷积。

在变分问题的求解中，对构造的变分问题模型采用了交替方向乘子法。通过更新模态分量和中心频率来获得满足条件的最优解。Dragomiretskiy 和 Zosso[8]详细描述了该算法。VMD 的流程图如图 1 所示。

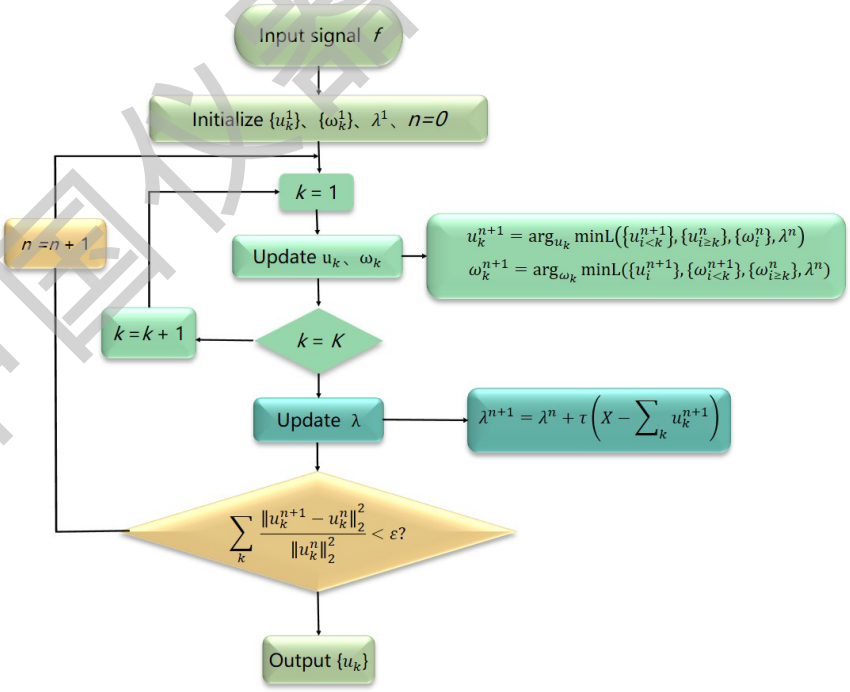

图 1　VMD 流程图

2.2 极限学习机

ELM 作为单隐层前馈神经网络，先根据某种分布随机选取输入层到隐藏层之间的权值，再利用最小二乘法获得隐藏层到输出层的权值。图 2 是 ELM 的示意图。ELM 将输入数据映射到 L 维随机特征空间，并生成以下形式的输出函数:

$$f_L = \sum_{i=1}^{L} \beta_i \, h_i(x) = h(x)\beta \tag{2}$$

其中$\beta = [\beta_1 \dots \beta_L]^T$表示隐藏层节点和输出层节点之间的输出权重矩阵，$h(x) = [G_1(x) \dots G_L(x)]$表示输入数据 **X** 的隐藏节点的输出。

ELM 用于解决以下线性方程：

$$H\beta = T \tag{3}$$

β表示连接隐藏节点和输出节点的权重向量

$$\beta = [\beta_1 \dots \beta_L]^T_{L*m} \tag{4}$$

H 表示隐藏层随机化输出矩阵

$$H = \begin{bmatrix} h(x_1) \\ h(x_2) \\ . \\ . \\ h(x_N) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & h_L(x_1) \\ & . \\ & . \\ h_1(x_N) & h_L(x_N) \end{bmatrix}_{N*L} \tag{5}$$

T 表示目标函数矩阵

$$T = \begin{bmatrix} t(x_1^T) \\ t(x_2^T) \\ . \\ . \\ t(x_N^T) \end{bmatrix} = \begin{bmatrix} t_{11} & t_{1m} \\ & . \\ & . \\ t_{N1} & t_{Nm} \end{bmatrix}_{N*m} \tag{6}$$

ELM 使用一组 N 个不同样本（x_i，t_i），其中 $x_i \in R^m$ 和 $t_i \in R$。具有 L 个隐藏神经元的单隐藏层前馈神经网络显示输出，其形式：

$$t_j = \sum_{i=1}^{L} \beta_i \, g(w_i x_i + b_i); \quad j \epsilon [1, N] \tag{7}$$

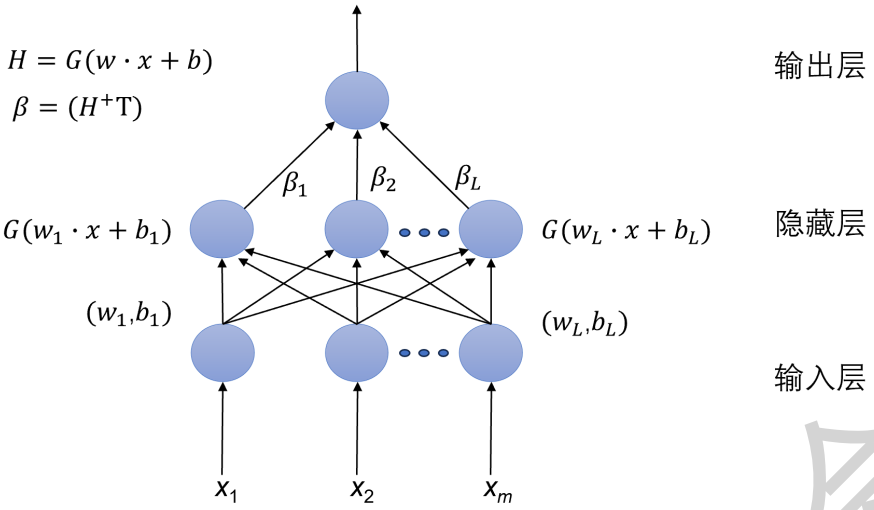

图 2　ELM 的示意图

2.3 变分模态分解展开极限学习机

本研究提出了 VMD-UELM 方法，用于复杂样品的光谱定量分析，充分利用了 VMD 和 ELM 的优点。该方法的原理如图 3 所示，分为校正和预测两个阶段。在校正阶段，首先通过 VMD 将训练集中的每个光谱分解为 K 个带有不同频率的模态分量 uk（k=1，2，…，K）。然后，将模态分量沿变量方向展开为一个扩展矩阵。最后，在扩展矩阵和训练集中的目标值之间建立 ELM 模型。这种方法的主要优点在于只需建立一个 ELM 模型，而无需为每个模态分量建立多个 ELM 子模型进行训练和组合预测结果。因此，引入的展开策略不仅能有效利用数据集的信息，还能避免权重子模型的繁琐确定过程。在预测阶段，以与训练集相同的方式，将预测集中的光谱通过 VMD 分解并展开为扩展矩阵。通过 ELM 来预测结果。

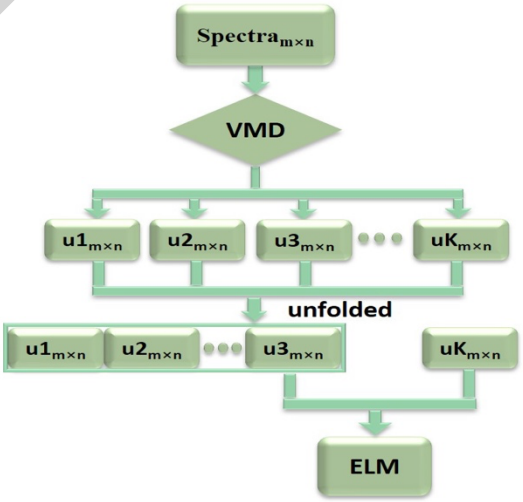

图 3　VMD-UELM 的示意图

为了验证 VMD-UELM 的有效性，将样品数据集进行划分。在血液数据集中，选择 173 个样品作为训练集，58 个样品作为预测集。在燃料油数据集中，选择 76 个样品作为训练集，38 个样品作为预测集。二元中药数据集的训练集和预测集分别为 50 和 25 个样品。以 RMSEP 和 R 作为预测集的评价标准来评价预测性能，并与 ELM、PLS 进行了比较。对于 ELM 的两个关键参数激励函数和隐藏层节点数，根据 R 的均值与标准偏差的比值（MSR）随不同激励函数和隐藏层节点数的变化情况来确定。其中激励函数包含了 sin、sig、ribas、hardlim、radbas，隐藏层节点数范围为 1-100。对于血液数据集，使用 sig 激励函数且隐藏层节点数 59 时，MSR 达到最大值。在燃料油数据集上，MSR 的最大值出现在使用 sin 激励函数且隐藏层节点数为 30 的情况下。二元中药数据集 MSR 最大值与 sin 激励函数和隐藏层节点数 100 相对应。对于 PLS 的参数—因子数（latent variable，LV），采用蒙特卡洛交叉验证（MCCV）与 F 检验相结合来确定。血液、燃料油、二元中药数据集的 LV 分别为 11、4、7。

3 实验数据

本研究采用了 3 个光谱数据集来验证 VMD-UELM 方法的性能。其中，血液数据集包含了 231 个血液样品的近红外漫反射和透射光谱，以及血红蛋白、葡萄糖和胆固醇的含量。该数据集由 Norris[15] 提供，并可从以下网址下载：http://www.idrcchambersburg.org/shootout2010.html。光谱是使用近红外 6500 光谱仪（NIR systems, Inc., Silver Springs, USA）进行测定的，波长范围为 1100-2498 nm，间隔为 2 nm，共包括 700 个波长点，其光谱图如图 4 所示。在本研究中，使用反射光谱血红蛋白的含量作为目标值进行分析。

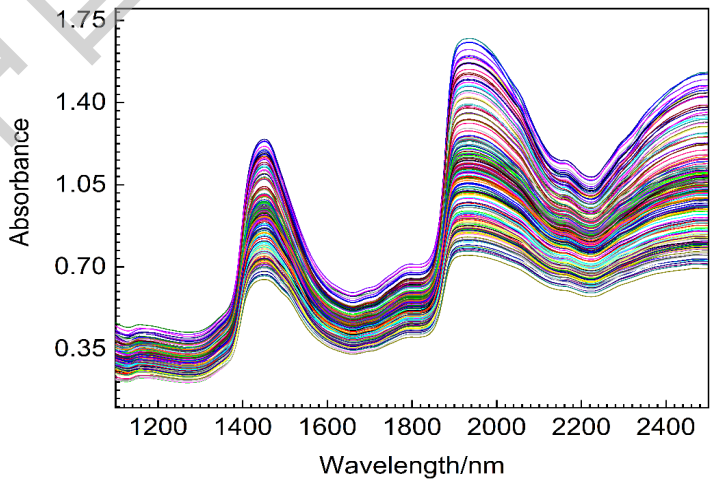

图 4　血液数据集的光谱

燃料油数据集来源于 Wentzell[16] 等人的研究，可从 https://myweb.dal.ca/pdwentze/downloads.html 下载。该数据集包含 115 个轻汽油和柴油燃料的紫外光谱，以及相应的多环芳烃、双环芳烃、单环芳烃和饱和烃的含量。光谱数据是使用 Cary 3 紫外-可见分光光度计（Varian Instruments, San Fernando, Calif.）进行测量的。每个样品有 572 个波长点，波长范围为 200-400 nm，间隔为 0.35 nm。根据网站说明，由于第 115 号样品是一个奇异样品，因此在分析中没有使用。图 5 展示了剩余 114 个样品的光谱。在本研究中，以双环芳烃的含量作为分析的目标。

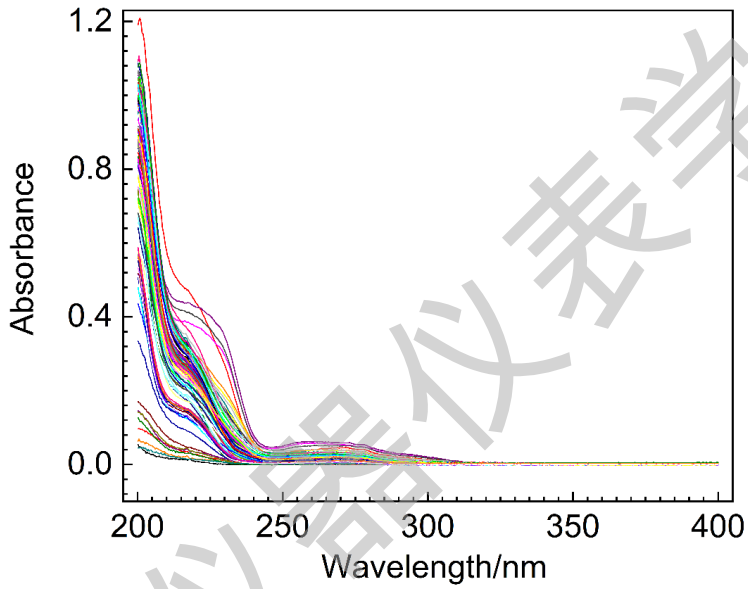

图 5　燃料油数据集的光谱图

二元中药数据集包含三七和莪术，是由本组实验室收集的。从天津多家药店购买三七和莪术。由于草药含有一定量的水分，因此在 60℃下干燥至恒定重量。这些草药被研磨成粉末，过 120 目不锈钢筛，并储存在 60 mm×100 mm 的密封塑料袋中。将处理过的草药粉末以不同的质量百分比混合，并确保每个样品中四种草药的总质量分数为 100%，共有 75 个样品。在 Vertex 70 近红外光谱仪（Bruker Optics Inc., Ettlingen, Germany）上以 2 cm^{-1} 的间隔对二元中药从 12000 到 4000 cm^{-1} 进行测量。本章对三七的含量进行分析预测。图 6 显示了样品的近红外光谱。

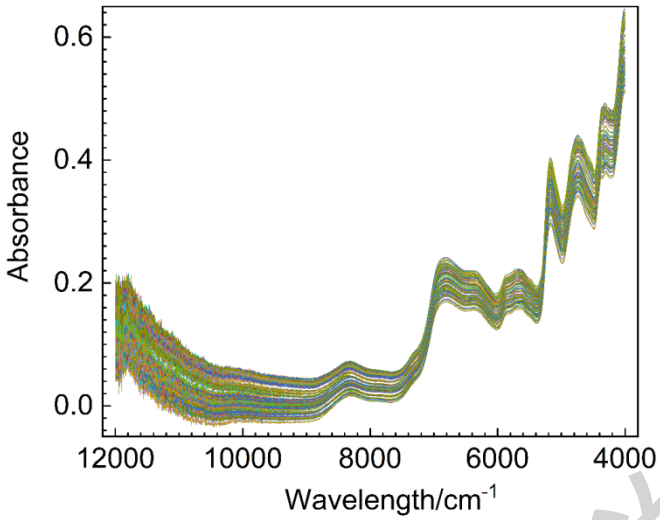

图 6　二元中药数据集的光谱图

4 结果讨论

4.1 模态分量数

模态分量的数量 *K* 决定了将每个光谱分解成多少个模态分量。选择合适的 *K* 值非常重要，过多或过少都会导致问题。如果分解过多，会出现虚假模态和过度分解问题，降低精度并增加计算负载；而分解过少则会导致分解不充分，信息不全。图 7 显示了血液数据集中 VMD-UELM 建模的 RMSEP 随模态分量数量变化的趋势，可以看出随着模态分量数量的增加，初始阶段 RMSEP 缓慢增加，之后有下降趋势，最低值出现在模态分量为 5 时。因此，血液数据集的最佳 *K* 值为 5。对于燃料油数据集，在图 8 中可以看出随着 *K* 的增加，RMSEP 先上升后下降，当 *K* 为 5 时 RMSEP 最低，而 *K* 为 6 时急剧上升。因此，燃料油数据集的最优模态分量数为 5。在二元中药数据集中，根据图 9 显示的 RMSEP 随模态分量数量变化的波动，当模态分量为 5 时 RMSEP 最低。因此，二元中药数据集的最佳 *K* 值也为 5。

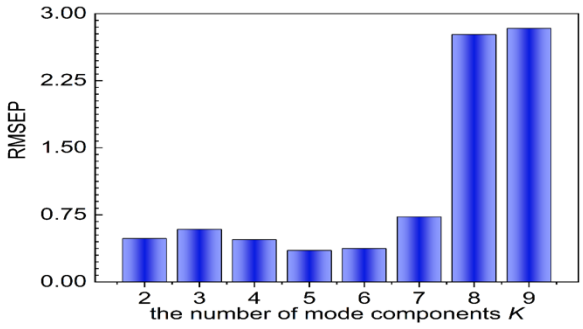

图 7　血液数据集 VMD-UELM 建模的 RMSEP 随模态分量数 *K* 的变化

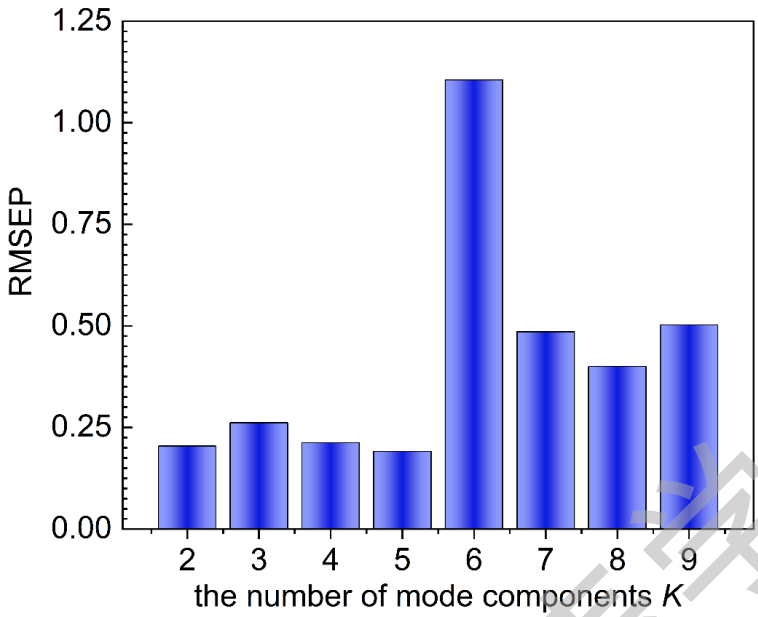

图 8　燃料油数据集 VMD-UELM 建模的 RMSEP 随模态分量数 *K* 的变化

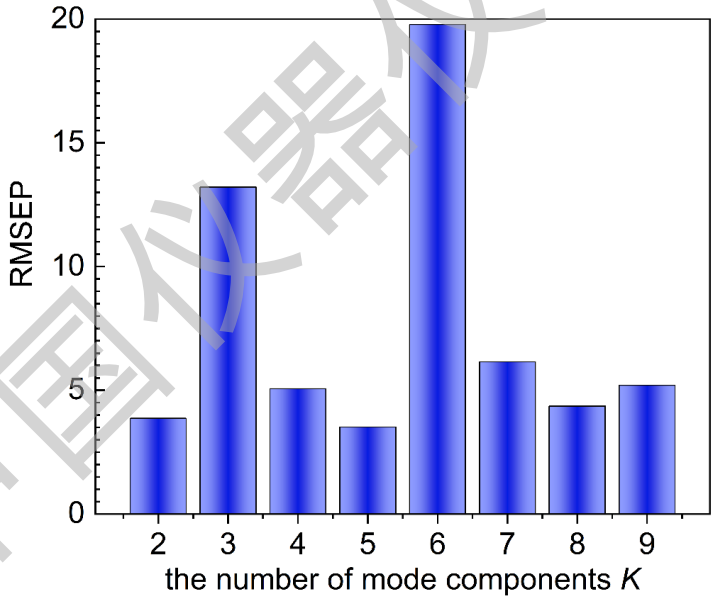

图 9　燃料油数据集 VMD-UELM 建模的 RMSEP 随模态分量数 *K* 的变化

4.2 光谱的 VMD 分解

VMD 将每个光谱信号分解为多个模态分量，从而提取出不同的光谱信息。通过分析第 3 号样品的血液数据集分解结果，可以更好地理解这些模态分量的含义。在图 10 中，原始近红外反射光谱被分解成了 5 个 u 分量，按照提取的顺序进行绘制。每个频率块可能携带着不同的信息，并对模型的贡献也各不相同。通过观察图 9，可以看到 u1 波动很小，整体趋

于平缓；u2 的波动相对较大，有明显的起伏，出现两个明显峰值；从 u3 到 u5，频率波动逐渐加快，出现多个峰值，且没有噪音干扰，这表明这些模态分量中包含了许多有用的信息。

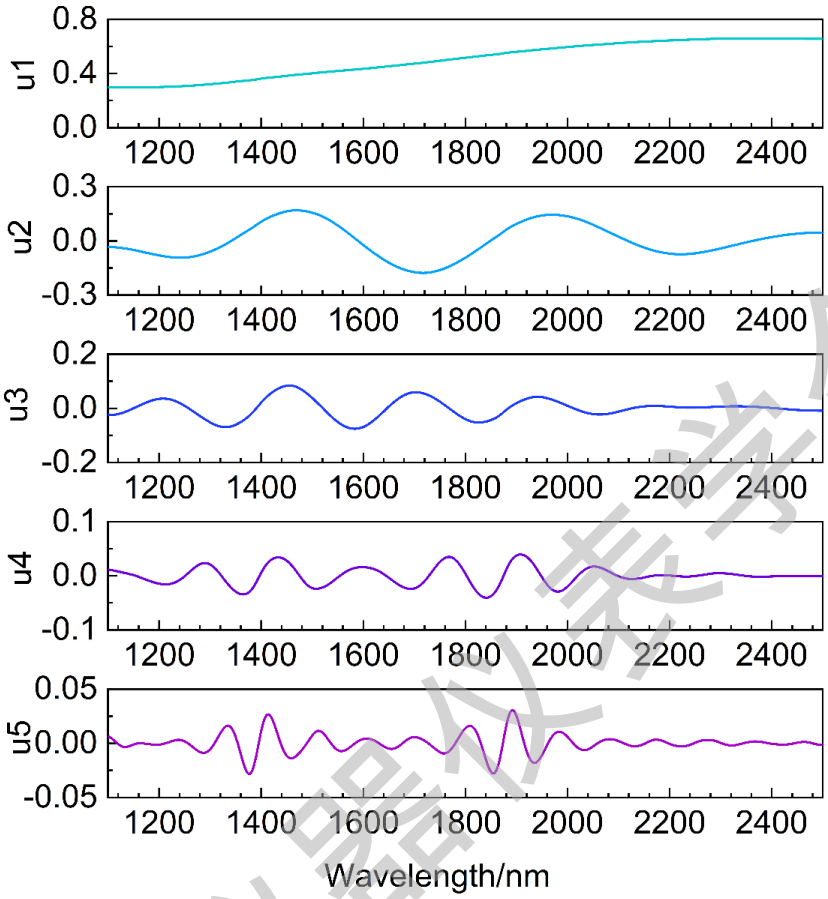

图 10　血液数据集中 3 号样品的光谱变分模态分解图

对于燃料油，以 20 号样品为例，图 11 显示光谱被分解为五个 u 分量。前三个 u 分量在波长范围为 200-280 nm 之间表现出较大的波动，而在 280-400 nm 范围内变化比较平稳，这表明这些分量可能包含了大量有用信息。相较于前三个分量，u4 的波动幅度更大，看起来更像是噪音成分。u5 的频率变化更快，所包含的有用信息更少。

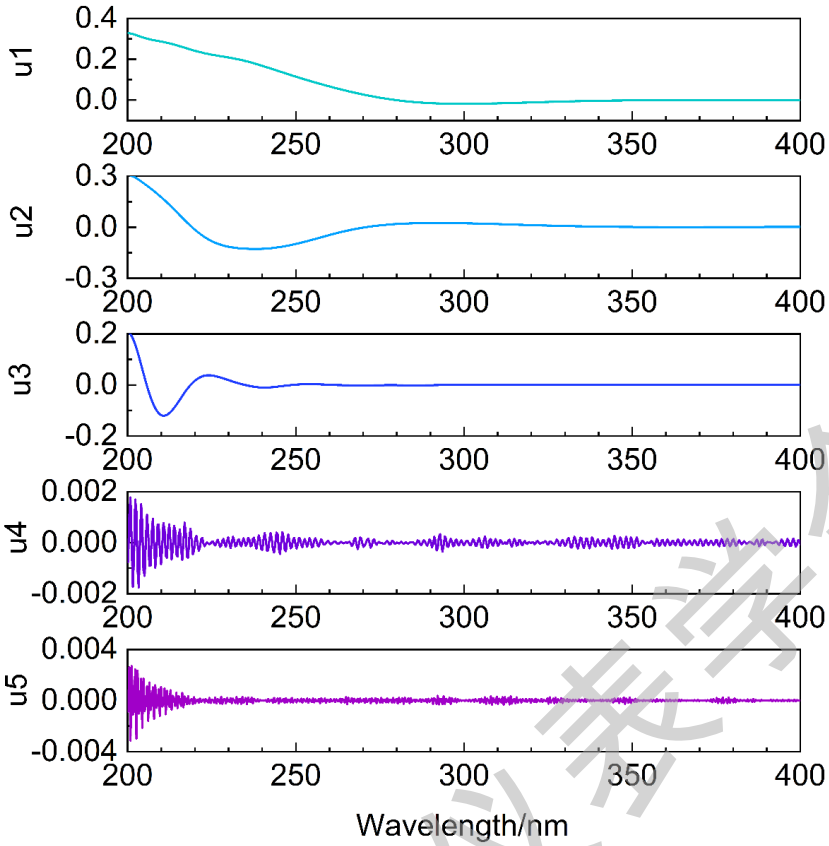

图 11　燃料油数据集中 20 号样品的光谱变分模态分解图

二元中药数据集以 31 号样品为例。图 12 显示，u1 和 u2 分量在 12000-8000 cm^{-1} 处波动平缓，在 8000-4000 cm^{-1} 处波动较大，可能包含较多有用信息。后 3 个 u 分量的曲线频率变化很快，峰值数量明显增加，显示出潜在的噪声干扰，并且在 8000-4000 cm^{-1} 波长范围内呈现出白噪音，对模型的贡献程度不大。

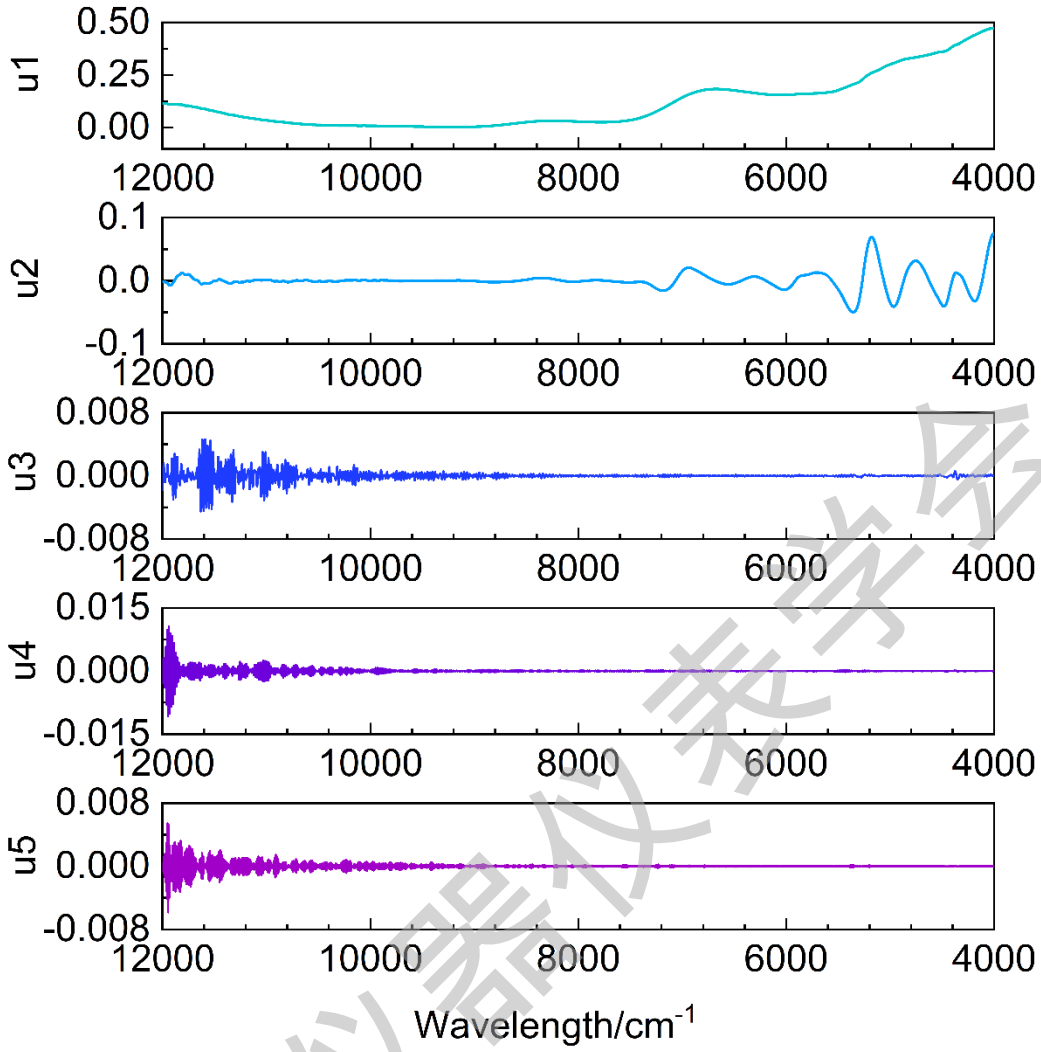

图 12　二元中药数据集中 31 号样品的光谱变分模态分解图

4.3 不同方法的预测结果

为了验证所提方法的预测性能，使用 ELM、PLS 进行比较。预测集的 RMSEP 和 R 用来作为检验模型性能的指标。表 1 总结了不同方法得出的预测结果。从表中可以看出，对于血液数据集，ELM 比 PLS 有较低的 RMSEP 和较高的 R，说明 ELM 优于 PLS。此外，VMD-UELM 在三种方法中 R 最高，RMSEP 最低。这说明对血液样品的近红外反射光谱进行 VMD 分解并展开建模可以进一步提高 ELM 的预测能力。

对于燃料油数据集，VMD-UELM、ELM 和 PLS 的 R 都在 0.98 以上，其改善空间较小。ELM 的 RMSEP 小于 PLS，VMD-UELM 的 RMSEP 是最小的。在二元中药数据集中，VMD-UELM 比 ELM、PLS 的 RMSEP 更小，相关系数 R 更大。结果表明，与其 ELM 和 PLS 相比，所提出的方法能更有效地提高模型的准确度。

表 1　通过 VMD-UELM、ELM、PLS 方法得出的预测结果的比较

Dataset	Method	RMSEP	R
blood	PLS	0.4310	0.9613
	ELM	0.3373	0.9766
	VMD-UELM	0.3217	0.9808
gasoil	PLS	0.1935	0.9896
	ELM	0.1907	0.9885
	VMD-UELM	0.1718	0.9901
Chinese traditional medicine	PLS	4.4467	0.9929
	ELM	2.3846	0.9979
	VMD-UELM	2.1734	0.9985

5 结论

基于变分模态分解和极限学习机的优势，提出了一种变分模态分解展开极限学习机方法，称为 VMD-UELM。验证了其在预测血液、燃料油和二元中药样品的可行性，并分别对血红蛋白、双环芳烃和三七的含量进行定量分析。首先利用 VMD 将每个光谱分解为多个模态分量。然后，将这些模态分量沿变量方向展开成一个扩展矩阵。最后，利用 ELM 在扩展矩阵和目标值之间建立定量模型。VMD-UELM 充分利用了光谱局部特征信息。与 ELM 和 PLS 相比，该方法有更低的 RMSEP 和更高的 R。结果表明，VMD-UELM 具有更好的定量预测性能。

参考文献：

[1] Wang T Y, Xie C J, You Q, et al. Qualitative and quantitative analysis of four benzimidazole residues in food by surface-enhanced Raman spectroscopy combined with chemometrics[J]. Food Chemistry, 2023, 424: 136479.

[2] Eissa M S, Darweish E. Insights on ecological spectroscopic techniques recently adopted for

pharmaceutical analysis: A comprehensive review from the perspective of greenness assessment metrics systems application[J]. Trac Trends in Analytical Chemistry, 2024, 170: 117435.

[3] Oliveira L G, Araújo KC, Barreto MC, et al. Applications of chemometrics in oil spill studies[J]. Microchemical Journal, 2021, 166: 106216.

[4] Wang H P, Chen P, Dai J W, et al. Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues[J]. Trac-Trends in Analytical Chemistry, 2022, 153: 116648.

[5] Parastar H, Tauler R. Big (Bio)chemical data mining using chemometric methods: A need for chemists[J]. Angewandte Chemie-International Edition, 2022, 61: e201801134.

[6] Aroca-Santos R, Cancilla JC, Matute G, et al. Identifying and quantifying adulterants in extra virgin olive oil of the picual varietal by absorption spectroscopy and nonlinear modeling[J]. Journal of Agricultural and Food Chemistry, 2015, 63(23): 5646-5652.

[7] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: a new teaming scheme of feedforward neural networks[C]. IEEE International Joint Conference on Neural Networks, 2004.

[8] Dragomiretskiy K, Zosso D. Variational mode decomposition[J]. IEEE Transactions on Signal Processing. 2014, 62, 3, 531-544.

[9] Chen Q M, Chen J H, Lang X, et al. Self-tuning variational mode decomposition[J]. Journal of the Franklin Institute-Engineering and Applied Mathematics, 2021, 358: 7825-7862.

[10] Lahmiri S. Intraday stock price forecasting based on variational mode decomposition[J]. Journal of Computational Science, 2016, 12: 23-27.

[11] Hu H L, Wang L, Tao R. Wind speed forecasting based on variational mode decomposition and improved echo state network[J]. Renewable Energy, 2021, 164: 729-751.

[12] Wang Z J, Wang J J, Du W H. Research on fault diagnosis of gearbox with improved variational mode decomposition[J]. Sensors, 2018, 18(10): 3510.

[13] Bian X H, Wu D Y, Zhang K, et al. Variational mode decomposition weighted multiscale support vector regression for spectral determination of rapeseed oil and rhizoma alpiniae offcinarum adulterants[J]. Biosensors, 2022, 12: 586.

[14] Bian X H, Li S J, Lin L G, et al. High and low frequency unfolded partial least squares

regression based on empirical mode decomposition for quantitative analysis of fuel oil samples. Analytica Chimica Acta. 2016, 925: 16-22.

[15] Kuenstner J T, Norris K H. Spectrophotometry of human hemoglobin in the near infrared region from 1000 to 2500 nm[J]. Journal of Near Infrared Spectroscopy, 1994, 2(2): 59-65.

[16] Wentzell P D, Andrews D T, Walsh J M, et al. Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration[J]. Canadian Journal of Chemistry, 1999, 77: 391-400.

[17] Tsekouras G J, Dialynas E N, Hatziargyriou N D, et al. A non-linear multivariable regression model for midterm energy forecasting of power systems[J]. Electric Power Systems Research, 2007, 77(12): 1560-1568.

[18] Wang L, Liu H Z, Liu L, et al. Prediction of peanut protein solubility based on the evaluation model established by supervised principal component regression[J]. Food Chemistry, 2017, 218: 553-560.

[19] Xu Q S, Liang Y Z. Monte carlo cross validation[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 56: 1-11.

[20] Helland I S, Sæbo S, Almoy T, et al. Model and estimators for partial least squares regression[J]. Journal of Chemometrics, 2018, 32(9): e3044.

[21] 韩立群. 人工神经网络[M]. 北京：北京邮电大学出版社，2006.

[22] Huang G B, Chen L, Siew C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Transactions on Neural Networks, 2006, 17(4): 879-892.

[23] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20: 273- 293.

通讯作者简介：

卞希慧，女，1983 年生，教授，主要进行化学计量学算法研究及其在中药、食品、环境等方面的应用研究。

E-mail：bianxihui@163.com

第一作者简介：

沈亮亮，男，2000 年，硕士研究生，研究方向为化学计量学建模方法研究。

E-mail: 3138417205@qq.com