基于霸王龙优化算法的复杂样品近红外光谱变量选择方法

Tyrannosaurus Optimization Algorithm for Near-infrared Spectral Variable Selection for Complex Samples

卞希慧[1*]，杨文博[1]，赵春焱[2]，刘鹏[1]，谭小耀[1]

（1.天津工业大学 化学工程与技术学院 化学工程与工艺系，天津 300387；2.天津工业大学 经济与管理学院 信息管理与信息系统系，天津 300387）

摘要：近红外光谱分析技术以其操作简便、快速无损、环境友好等优势，与多元校正方法相结合在复杂样本的定量分析中得到了广泛应用。然而，复杂样品的光谱数据常常面临谱峰重叠、变量冗余等问题，这些因素可能影响分析结果的准确性。为了提高模型的预测能力，需要在建模前对光谱数据进行变量选择。霸王龙优化算法（Tyrannosaurus Optimization Algorithm, TROA）概念简单，易于实施，收敛速度快且能避免局部最优。将霸王龙优化算法离散化与偏最小二乘结合，用于四组数据集的近红外光谱变量选择。为验证该方法的有效性，将该方法与全光谱PLS以及六种变量选择方法相比较。结果表明，霸王龙优化算法在这四组数据的近红外光谱变量中选择的变量最少且所选变量更具代表性。

关键词：偏最小二乘建模；变量选择；光谱分析；霸王龙优化算法；离散化

1 引言

光谱分析技术由于其简单、快速、准确、无损的优点在分析方法中备受瞩目，且被应用于化工[1]、石油[2]、制药[3]、材料[4]等各个领域。然而，由于样品的复杂性，光谱在分析过程中难免会由于光谱的吸收强度产生光谱重叠、谱带复杂、信息冗余等大量负面影响，不仅会增加所建模型的复杂度，也会对分析结果造成很大的影响。为了克服此问题，在建模前需要对与目标组分相关的变量进行筛选[5]。

随着化学计量学的不断发展[6]，多种变量选择方法已被提出，如变量投影重要性（Variable Importance In Projection, VIP）是一种以PLS为基础的变量选择技术[7]。在偏最小二乘（Partial Least Squares, PLS）建模中，各个变量会沿着方差最大的方向进行投影，以评估它们与模型输出之间的相关性[8]。竞争性自适应重加权算法（Competitive Adaptive Reweighted Sampling, CARS）是一种高效的波长选择方法，它利用自适应重加权采样

（Adaptive Reweighted Sampling, ARS）技术从PLS模型中挑选出具有较大回归系数绝对值的波长点[9]。无信息变量消除算法（Uninformative Variable Elimination, UVE）[10]由Centner等人首次提出并应用于NIR光谱数据。该算法的主要目的是减少最终偏最小二乘法（PLS）模型中包含的变量数，从而降低模型的复杂性并改善PLS模型的性能。蒙特卡罗-无信息变量消除方法（Monte Carlo-Uninformative Variable Elimination, MC-UVE）**错误!未找到引用源。**是在无信息变量消除方法（UVE）的基础上，融合了蒙特卡洛采样原理。这种方法利用蒙特卡罗技术代替了UVE方法中的留一交叉验证（LOOCV），以计算变量的稳定性值。这种改进使得MC-UVE能够更有效地从数据的不同角度提取并表达样本光谱与待测组分性质之间的复杂关系，通过这种方式，MC-UVE能够可靠地估计每个变量的稳定性，从而有望解决过拟合问题[12]。随机检验-偏最小二乘法（Randomization Test-Partial Least Squares, RT-PLS）是由邵学广课题组在2009年提出的一种特征选择方法[13]。随机检验（Random Test, RT）是利用整个样本的分布规律对于某类假设而进行检测的统计学方法.随机检验-偏最小二乘法能保证光谱数据不会在校正集中变化，并能检测保留的波长数与所建模型的关系，留下适当有信息的波长，建立最优的模型。

相较于基于单一指标和统计学的变量选择方法，基于群体智能优化算法的变量选择方法借助于其强大的算力，在处理大型复杂问题时具有较大优势。而目前已经发展了许多群体智能优化算法，但只有少部分运用于光谱变量选择中。霸王龙优化算法（Tyrannosaurus Optimization Algorithm, TROA）是印度Sahu等人[14]人受霸王龙狩猎行为的启发于2023年提出的一种新的群体智能优化算法。该算法基于霸王龙的狩猎行为，霸王龙以一定的成功率追逐猎物，而猎物则以一定的速度试图逃跑，霸王龙通过狩猎空间寻找最优解。相较于其它群体智能优化算法，霸王龙优化算法概念简单，易于实施，收敛速度快且能避免局部最优，但是尚未应用于近光谱的变量选择中。

本研究首次将霸王龙优化算法离散化，并探讨其在近红外光谱变量选择中的可行性。以药片、六元调和油、橘汁和土壤样品的近红外光谱作为研究对象，通过研究偏最小二乘因子数、离散化函数以及最大迭代次数对模型性能的影响。采用最佳离散化函数下的TROA算法选择与待分析组分相关的近红外光谱变量，并构建偏最小二乘（Partial least squares, PLS）模型。为验证该方法的有效性，将该方法与全光谱PLS，以及六种变量选择方法相比较，包括无信息变量消除法（UVE）、蒙特卡罗-无信息变量消除法（MC-UVE）、随机检验（RT）、竞争性自适应加权采样（CARS）、灰狼优化算法（Grey Wolf Optimizer, GWO）[15]和鲸鱼优化算法（Whale Optimization Algorithm, WOA）[16]。结果表明，霸王龙优化算法在这四

组数据的近红外光谱变量中选择的变量最少且所选变量更具代表性。因此，霸王龙优化算法是一种有效和高效的复杂样品近红外光谱变量选择方法，为分析化学领域的数据处理提供了一种新方法。

2 实验

2.1 霸王龙优化算法

霸王龙优化算法是一种模拟霸王龙狩猎、迁徙和竞争行为的群体智能优化算法。该算法将优化问题转化为一个生态系统模型，其中每个霸王龙代表一个解决方案，其适应度表示该解决方案的优劣程度。算法通过模拟霸王龙的狩猎行为来更新解决方案，寻找更好的解决方案。霸王龙优化算法模拟了霸王龙狩猎行为，主要分为：初始化、狩猎和选择。通过迭代最终得到最优猎物的位置即全局最优解。

（1）初始化

TROA 是一种基于种群的算法，它在搜索空间中随机生成猎物的数量。假设 x 是猎物的位置或地点，在搜索空间的上限和下限内随机生成，如公式(1)所示。

$$X_i = \mathrm{rand}(np, \dim) * (ub - lb) + lb \tag{1}$$

其中，$X_i = [x_1, x_2, \cdots x_n]$是猎物的位置，$i = 1, 2 \cdots n$，n 是维数，np 是种群数量，dim 是搜索空间维数，ub 是上限，lb 是下限。

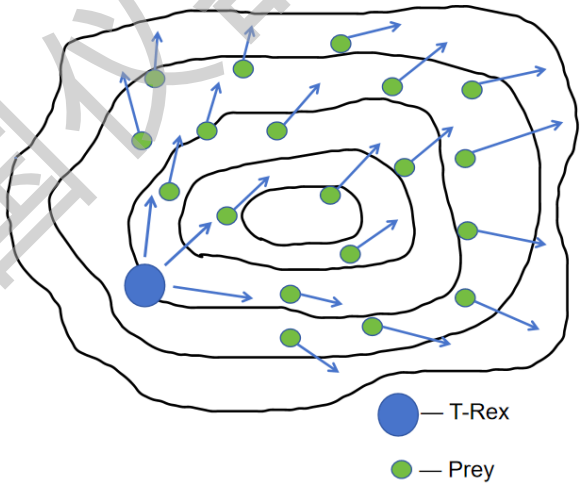

图 1　猎物和霸王龙的狩猎区

图 1 中，绿色圆圈代表猎物位置，黑色圆圈代表霸王龙位置。

（2）狩猎

霸王龙就像狮子、狼等顶级掠食者一样，当霸王龙看到离它最近的猎物时，它就会试图捕猎。有时猎物会保护自己免受狩猎或逃跑。霸王龙狩猎时，幼体会追逐和捕捉猎物，因此

霸王龙狩猎时，他会随机狩猎。

$$X_{new} = \begin{cases} X_{new} & if\ Rand() < Er \\ Random & else \end{cases} \tag{2}$$

$$E_r = Rand() * (1 - (\mathrm{t}/\mathrm{Max\_iteration})) \tag{3}$$

其中，Er 是到达分散猎物的估计值，即当霸王龙开始捕猎时，猎物开始破碎，并通过更新其位置来狩猎猎物，如公式(2)所示。Max_iteration 为最大迭代次数。

$$X_{new} = x + rand() * sr * (tpos * tr - target * pr) \tag{4}$$

其中，sr 是狩猎成功率，介于[0,1]之间。如果成功率为 0，则表示猎物已经逃跑，狩猎失败，必须相应地更新猎物位置。目标是猎物到霸王龙位置的最小位置。霸王龙的奔跑速度为 tr。但在这里，我们只考虑猎物被猎杀，且只有一只霸王龙。假设霸王龙的平均奔跑速度为 30 英里/小时，行走速度为 6.7 英里/小时[31]。因此，霸王龙的奔跑速度在[0.067,0.3]之间。pr 是猎物的奔跑速度，在[0,1]之间，但在这里，必须考虑猎物的奔跑速度应小于霸王龙的速度。$tpos$为当前霸王龙的位置，$target$为当前目标猎物的位置。霸王龙优化算法流程图，如图 2 所示。

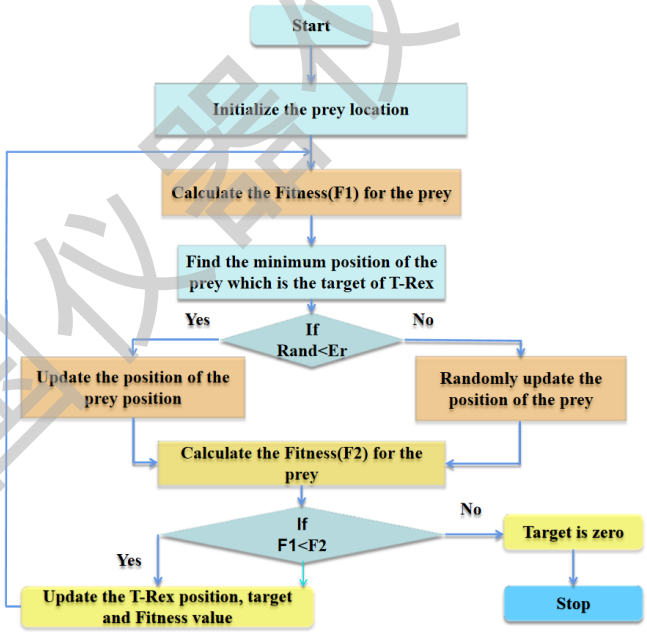

图 2　TROA 的流程图

（3）选择

选择过程取决于猎物的位置，即目标猎物的当前位置和先前的位置。如果霸王龙狩猎失败，如果猎物逃跑或保护自己不被捕猎，则猎物的位置会变为零。这是通过比较适应度函数来实现的。

$$X_i^{k+1} = \begin{cases} update\ the\ target\ position\ if & f(X) < f(X_{new}) \\ target\ \ is\ \ zero & otherwise \end{cases} \tag{5}$$

其中，f(X)是初始随机猎物位置的适应度函数，f(X_{new})是更新猎物位置的适应度函数。

霸王龙优化算法是一种高效且功能强大的工具，它被广泛应用于解决各类优化问题。在处理变量选择这一关键任务时，提出了一种离散化的 TROA-PLS 方法，以应对挑战。

在实际应用中，变量集中往往包含了许多冗余或干扰变量，这些变量不仅无助于模型的建立，反而可能对其准确性造成负面影响。因此，有效的变量选择显得尤为重要。通过精心挑选出对模型有实质贡献的变量，能够从复杂的光谱数据中提取出关键的特征信息，这不仅有助于降低数据的维度，还能简化模型的结构，并显著提高预测的效率。

在本研究中，为了将霸王龙优化算法中的位置信息转化为二进制形式，采用了三种不同的传递函数进行离散化处理，分别是 atan、V 形和 sigmoid 传递函数。这些传递函数在处理不同数据集时，其预测性能的影响是有所不同的。因此，对这三种传递函数进行了对比研究，以了解它们在不同数据集上的预测性能。总的来说，通过使用适当的传递函数，可以有效地将霸王龙的位置信息进行二进制转换，从而在处理优化问题时，能够更精确地进行变量选择，提高模型的预测效率。

atan 传递函数公式为：

$$x_i^k(t+1) = |\arctan(x_i^k(t))| \tag{6}$$

其中，$x_i^k(t+1)$是第 i 只霸王龙在第 k 维的第 t+1 迭代时的值。

V 形传递函数公式为：

$$x_i^k(t+1) = |erf(\frac{\sqrt{\pi}}{2} x_i^k(t))| \tag{7}$$

sigmoid 传递函数公式为：

$$x_i^k(t+1) = \frac{1}{1+e^{-x}} \tag{8}$$

在完成了两种值的转换之后，接下来需要利用公式（9）对这些转换后的值进行进一步的转换处理。

$$x_i^k(t+1) = \begin{cases} 0 & \text{if} \quad x_i^k(t+1) < rand \\ 1 & \text{if} \quad x_i^k(t+1) \geq rand \end{cases} \tag{9}$$

在这个情况下，引入了一个介于 0 和 1 之间的随机数 rand。通过使用一个传递函数，将连续的解映射为离散的变量选择值，即 0 或 1。在这里，1 代表该变量被选中，而 0 则表示

该变量未被选中。

在离散化 TROA-PLS 算法中，核心的参数包括迭代次数 t 和猎物数量 z。其中，迭代次数 t 是影响 TROA-PLS 过程的关键因素。如果迭代次数过少，可能就不能获得最优的变量组合，这可能会降低模型的预测精度。然而，过多的迭代次数可能会增加模型的复杂性和计算量，从而延长计算时间。因此，在建模阶段，确定一个合适的 t 值至关重要。在本研究中，设定迭代次数 t 的范围为 1 至 500 次。TROA-PLS 变量选择的具体过程描述如下：

（1）初始化猎物的位置 x。每个猎物的位置利用公式（1）在搜索空间的上限和下限内随机生成。

（2）更新猎物的位置：当$Rand(\,) \geq Er$时，猎物位置进行随机更新，当$Rand(\,) < Er$时，猎物利用公式(4)实时更新自己的位置；接下来用不同离散函数离散化处理猎物的位置X_{new}。

（3）重复上述步骤直至符合终止条件（最大迭代次数）。

（4）输出最佳猎物向量即最优解X_i^{k+1}。在目标值对应的光谱变量和训练集中的最优猎物向量和之间建立 PLS 模型。

（5）对于预测阶段，选取与偏最小二乘模型训练时相同的光谱变量作为输入参数。这些变量将用于对预测集进行预测，以得到预测集中各样本的预测值。

为了全面评估霸王龙优化算法在变量选择方面的性能，需要关注几个关键的性能评价指标。这些指标包括：

1.变量选择的数量：这是指通过算法构建模型时，最终被选入模型的变量数量。一个高效的变量选择算法应该能够有效地剔除不相关的变量，从而使得被选入模型的变量数量尽可能少。这样，就能确保所选的变量几乎都是关键变量，这对于模型的性能和解释性都至关重要。

2.预测集均方根误差（RMSEP）：这是衡量模型预测能力的一个重要指标。它计算的是模型在预测集上的预测值与实际值之间的差异。理想情况下，希望这个差异尽可能小，这意味着模型的预测准确性高。其公式如下：

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{m_{pred}} \left(\hat{c}_i - c_i\right)^2}{m_{pred}}} \tag{10}$$

其中c_i表示实际值，$\bar{c}_i$表示所有样品实际值的均值，$\hat{c}_i$表示模型预测值，m_{pred}为预测集样品的个数。

3.交叉验证均方根误差（RMSECV）：这是一个评估模型稳定性和泛化能力的重要指标。

它基于交叉验证过程，计算模型在不同子集上的表现。一个低的 RMSECV 值通常意味着模型具有较好的泛化能力，其公式：

$$\mathrm{RMSECV}=\sqrt{\frac{\sum_{i=1}^{\mathrm{m}_{train}}\left(\hat{c}_i-c_i\right)^2}{\mathrm{m}_{\mathrm{train}}}} \tag{11}$$

其中 c_i 表示实际值，$\bar{c}_i$ 表示所有样品实际值的均值，$\hat{c}_i$ 表示模型预测值，$\mathrm{m}_{\mathrm{train}}$ 为训练集样品的个数。

4.相关系数 R：这是一个衡量模型预测值与实际值之间线性关系强度的指标。它的取值范围是-1 到 1，接近 1 或-1 表示强相关，接近 0 表示弱相关或无关，其公式：

$$\mathrm{R}=\sqrt{1-\frac{\sum_{i=1}^{m_{\mathrm{pred}}}\left(\hat{c}_i-c_i\right)^2}{\sum_{i=1}^{m_{pred}}\left(c_i-\bar{c}_i\right)^2}} \tag{12}$$

2.2 实验数据

为了评估霸王龙优化算法的性能，本研究选取了四组多样化的近红外光谱数据集进行测试。这些数据集中包括：药片样本中的活性成分、六元混合油样品中葵花籽油的成分、橘汁样本中的蔗糖成分，以及土壤样本中的有机质成分。通过这四种不同的数据类型，可以全面地验证霸王龙优化算法在处理近红外光谱数据时的准确性和效率。

在数据处理方面，对于药品数据集采用网站上的分组方法，将 310 个药片样品有效地分为 232 个样本的训练集和 78 个样本的预测集。将 51 个六元调和油样品采用 Kennard-Stone（KS）算法精确地划分为含有 34 个样品的训练集和 17 个样品的预测集。将 218 个橘汁样品采用网站上的分组方法，有效地划分为含有 150 个样品的训练集和 68 个样品的预测集。将 102 个土壤样品采用网站上的分组方法，划分为含有 82 个样品的训练集和 20 个样品的预测集。

药片数据集包含在实验室、中试工厂和生产规模生产的 310 个近红外光谱数据，可分为（A、B、C、D）4 种类型，包括药片的 NIR 数据和相对活性物质（API）含量[17]。其相对活性物质含量在（4.6 %- 9.8%）之间。光谱有 404 个变量，波数范围在 10500-7400cm^{-1} 之间，分辨率为 8cm^{-1}。该数据集由 M. Dyrby 测得，可以在 https://www.models.life.ku.dk/Tablets 上下载。图 3a 和图 3b 分别为药片数据集的近红外光谱图和样品中活性组分浓度分布图。

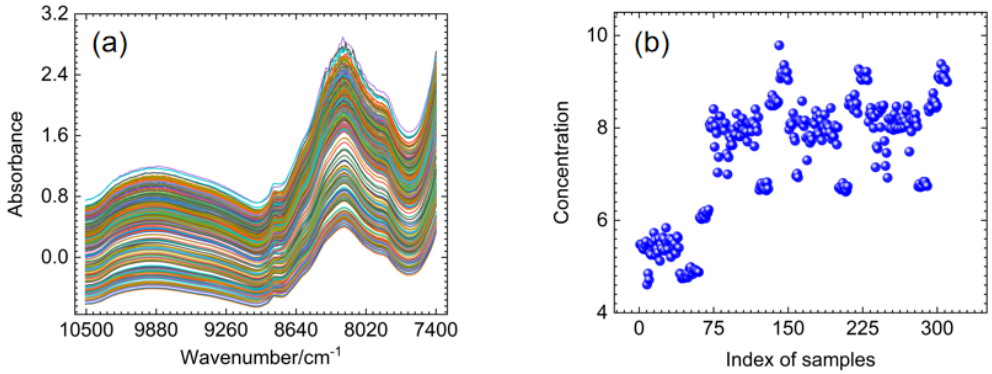

图 3　药片数据集的近红外光谱图（a）活性组分浓度分布图（b）

六元调和油数据集包含了 6 种不同类型的食用油：葵花籽油、芝麻油、大豆油、米油、玉米油和花生油[18]。这些油被混合制成了 51 份六元调和油样品，每份样品的质量均匀分布，且不重复，质量间隔相等。每个调和油样品的质量被设定为 5 克，因此 51 个样品的总质量为 255 克。由于六种油被混合，每种油的理论总质量为 42.5 克。在 51 个样品中，每种单一油的质量从 0 开始，按照等差数列的求和公式计算，质量区间为 0.0333 克。理论上，每种油的最大质量为 1.665 克。为了降低不同油之间的线性相关性，玉米油的质量按照递增顺序排列，而其他油的质量则通过 Matlab 软件进行随机打乱。根据设计的理论质量，使用精度为 0.0001 克的分析天平称量每种油的实际质量。在称重过程中，使用塑料滴管将每种油加入玻璃瓶中。然后，根据实际质量计算出每种单油的实际质量百分比。样品的光谱分析是在布鲁克光学公司生产的 Vertex 70 近红外光谱仪上进行的，该设备产自德国埃特林根。在此次测量中，记录了从 12000cm^{-1}-4000cm^{-1} 的波数范围，共采集了 4148 个数据变量。为了提高信噪比，扫描次数被设置为 64 次。仪器的分辨率为 4cm^{-1}，扫描间隔约为 1.93cm^{-1}。选择宽度为 2mm 的石英样品池用于测量样品。每个样品测量三次，平均光谱用于最终分析。每次测量后，用醇洗涤并干燥样品池。本文以葵花籽油的含量作为目标组分，图 4a 和 4b 分别为六元调和油数据集的近红外光谱图和葵花籽油组分浓度分布图。

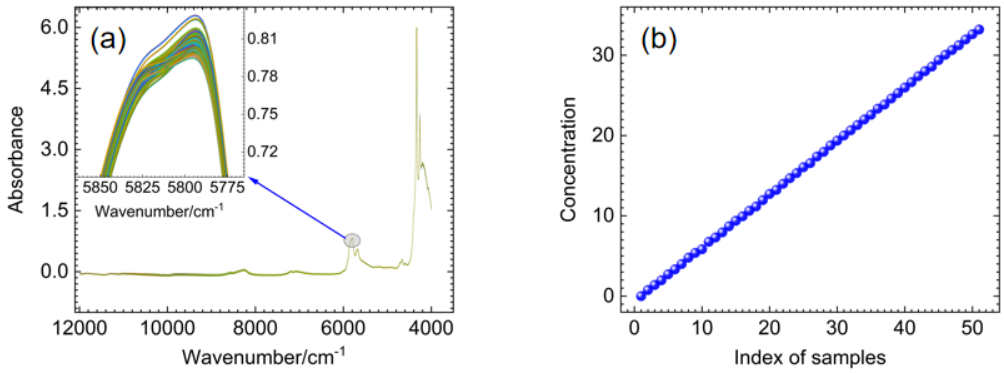

图 4　六元调和油数据集的近红外光谱图（a）葵花籽油组分浓度分布图（b）

橘汁数据集包括 218 个橘汁样本的近红外光谱及橘汁样品中的蔗糖浓度[19]。近红外光谱的波长范围从 1100nm-2498nm，包含 700 个不同的波长变量。此数据集由 Marc Meurens 教授提供，可以在 http://www.ucl.ac.be/mlg 上下载。图 5a 和 5b 分别为橘汁数据集的近红外光谱图和橘汁数据集中蔗糖组分浓度分布图。

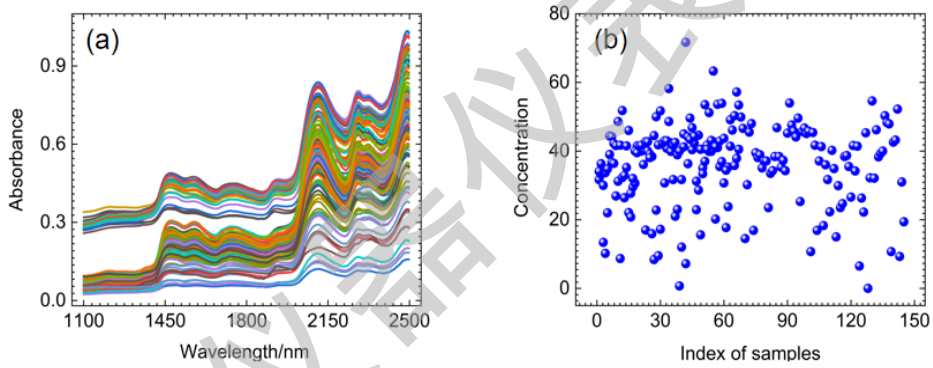

图 5　橘汁数据集的近红外光谱图（a）蔗糖组分浓度分布图（b）

土壤数据集包括 108 个土壤样品的近红外光谱数据，通过使用 NIR Systems 6500 NIR 分光光度计（Foss Analytical, Sweden）在 400 至 2500nm 的波长范围内记录光谱[20]。光谱范围为 400nm 至 2500nm，波长间隔为 2nm，共有 1050 个波长变量。这些样本来源于瑞典北部阿比斯科（681210 N, 181490 E）的长期野外实验样本来自 36 个地块，每个地块有三个子样本（在两个不同的层位，一个来自下部，两个来自上部），总共得到 108 个样本。此数据集可在网站 http://www.models.kvl.dk/nirsoil.上下载。在这项研究中，土壤有机质（SOM）含量在 550℃时测量为灼烧损失被视为有利害关系的财产。在删除了通过采样误差剖面分析异常值检测方法[31]检测到的 6 个异常值剩余的 102 个样本。图 6a 和图 6b 分别为土壤数据集的近红外光谱和样品中有机质的浓度分布图。

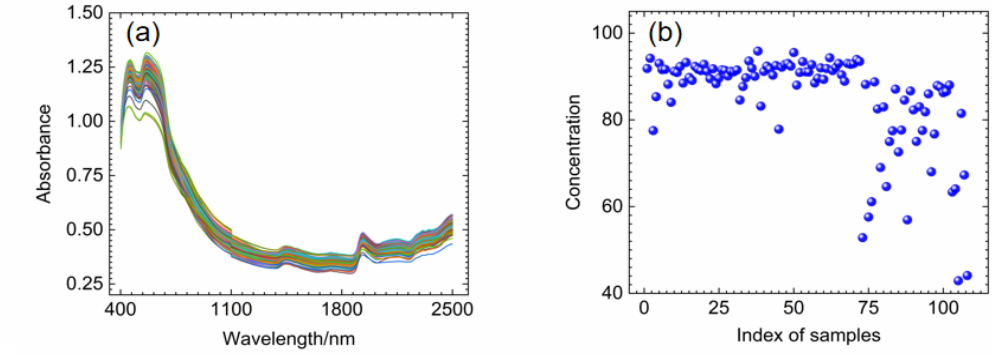

图 6　土壤数据集的近红外光谱图（a）有机质浓度分布图（b）

3 结果与讨论

3.1 偏最小二乘因子数

为了提高所建的 TROA-PLS 模型的准确度，将本实验使用的四种数据进行了处理，首先确定了最佳的偏最小二乘建模的因子数。图 7 为药片数据集中活性组分的交叉验证均方根误差（RMSECV）随着偏最小二乘因子数（LV）的变化图。从图中可以看出，LV 取 1 时，RMSECV 取到最大值，说明因子数太小会欠拟合，进而导致建模效果差。随着 LV 从 1 增大到 8，RMSECV 逐渐减小且当 LV 等于 8 时，RMSECV 取到最小值，光谱中的有用信息被完全提取出来。当 LV 从 8 增大到 25，RMSECV 逐渐增大。由于因子数越大，拟合效果越好，因子数越小，则欠拟合，所以本文中的实验取 7 作为药片数据集进行偏最小二乘回归建模的最佳因子数。此外，对六元调和油样品中的葵花籽油成分、橘汁样品中的蔗糖成分和土壤样品样品中的有机质成分进行了类似分析，这三组数据集的偏最小二乘因子数分别取 20、7 和 12。

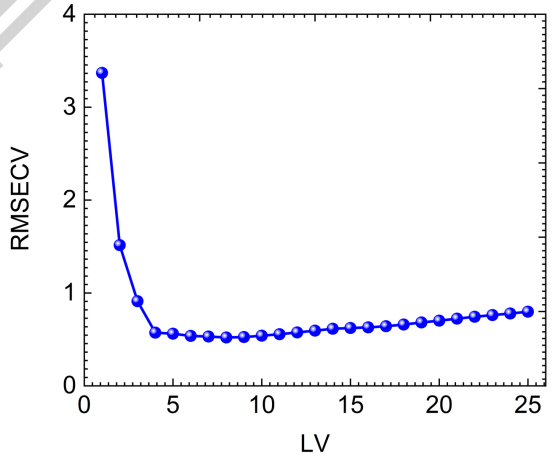

图 7　PLS 对药片数据集中活性组分建模时的 RMSECV 随因子数变化图

3.2 离散化函数

为了确保霸王龙优化算法的离散化过程能够有效地进行，需要考虑不同的离散化函数对优化结果的影响。离散化函数的选择对于算法的性能至关重要，因为不同的函数会导致不同的寻优效果。在本项研究中，选择了三种常用的离散化函数：atan 函数、V 形函数和 sigmoid 函数，以探讨它们在霸王龙优化算法中的表现。通过将这些函数应用于算法，并观察随着迭代次数增加，预测误差的均方根（RMSEP）如何变化，可以评估每种离散化函数的效能。在对药片数据进行分析后，观察到全局最优解的均方根误差（RMSEP）在不同离散化函数下，随迭代次数的变化情况如图 8 所示。图 8 描绘了在使用 atan、V、sigmoid 这三种不同数学模型对药片样品中的活性成分含量进行预测时，随着迭代次数增加，最优预测结果的均方根误差（RMSEP）的变化趋势。通过图像分析，可以明显看到，在不断增加的迭代次数下，三种不同的离散函数所处理的药片样本数据的均方根预测误差（RMSEP）呈现了一致的降低趋势，并逐渐达到一个平稳状态。详细地，注意到，当迭代次数处于 1 到 100 的区间时，sigmoid 离散函数对应的 RMSEP 迅速下降。而在 100 至 460 次迭代过程中，RMSEP 减少幅度相对较小；在 460 至 500 次迭代时，RMSEP 值虽然收敛，但 RMSEP 值仍然较高。这表明 sigmoid 函数在寻优过程中的表现相对较差。对于 V 形离散函数，其在迭代次数 1 至 200 的区间内 RMSEP 迅速下降，而在 200 至 500 次迭代中，RMSEP 值维持在较低水平并稳定，这表明算法已经收敛并接近最优值。在对比不同离散函数对于药片数据集的优化效果时，观察到 atan 离散函数在初始的 1 至 100 次迭代过程中表现出了最为显著的性能提升，其性能下降幅度最为突出。当迭代次数进一步增加，从 100 至 500 次时，atan 函数继续稳步改进并收敛，最终以三个函数中最低的 RMSEP 值收效。这一发现表明，在寻求药片数据集最佳解决方案的过程中，atan 离散函数的效能明显超过了 sigmoid 和 V 形离散函数。此外，对六元调和油样品中的葵花籽油成分、橘汁样品中的蔗糖成分和土壤样品样品中的有机质成分进行了类似分析，也同样确认了 atan 离散函数为最佳离散函数。

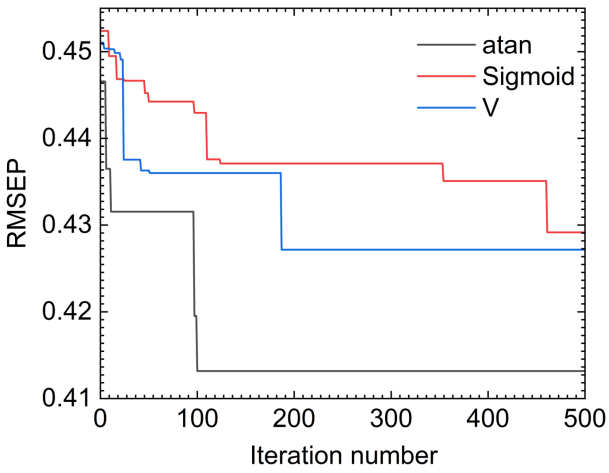

图 8　TROA 采用不同离散函数对药片数据集中活性组分含量预测时 RMSEP 随迭代次数的变化图

3.3 霸王龙优化算法的迭代次数

迭代次数是霸王龙优化算法在初始化阶段设置的一个主要参数。若迭代次数过少，模型可能无法充分寻找最优变量组合，从而导致预测精度较差。而若迭代次数过多，则会增加模型复杂度，使计算过程变得更加耗时。因此，需要确定一个合适的迭代次数，以提高模型的预测性能，降低模型复杂度。以土壤样品中有机质组分为研究对象，对迭代次数进行优化。其中狩猎成功率 sr 取一般值 0.8、霸王龙奔跑速度 tr 取一般值 0.3、猎物奔跑速度取一般值 0.25，猎物数量取 50，偏最小二乘因子数为 7，迭代次数的范围为 1~500，考察土壤样品中有机质组分的 RMSEP 随迭代次数变化情况，采用 atan 函数对土壤数据集进行离散化，结果如图 9 所示。由图 9 可知，在迭代次数从 0 增加到 100 的过程中，RMSEP 显著下降，这表明算法在这一阶段内进行了有效的优化搜索。当迭代次数继续增加，从 100 次迭代到 400 次迭代时，RMSEP 的下降速度开始减缓，并逐渐趋于平稳。最终，在迭代次数达到 400 至 500 次时，RMSEP 已经达到最小值。因此，可以推断出，在迭代次数大约为 400 至 500 次时，算法已经达到或接近了最优值。这一结果表明，该算法在经过大约 400 至 500 次迭代后，其性能提升的空间已经非常有限，可以认为此时算法的性能已经达到了一个较为稳定和优化的状态。同时，在其他三个数据集中发现虽然霸王龙优化算法收敛速度快，但为了与其他变量选择方法进行比较，将迭代次数设定 500，这样各种变量选择方法均能够达到最优值，并且实现收敛。因此，可以认为，在迭代次数为 500 时，算法性能最佳且稳定性良好。

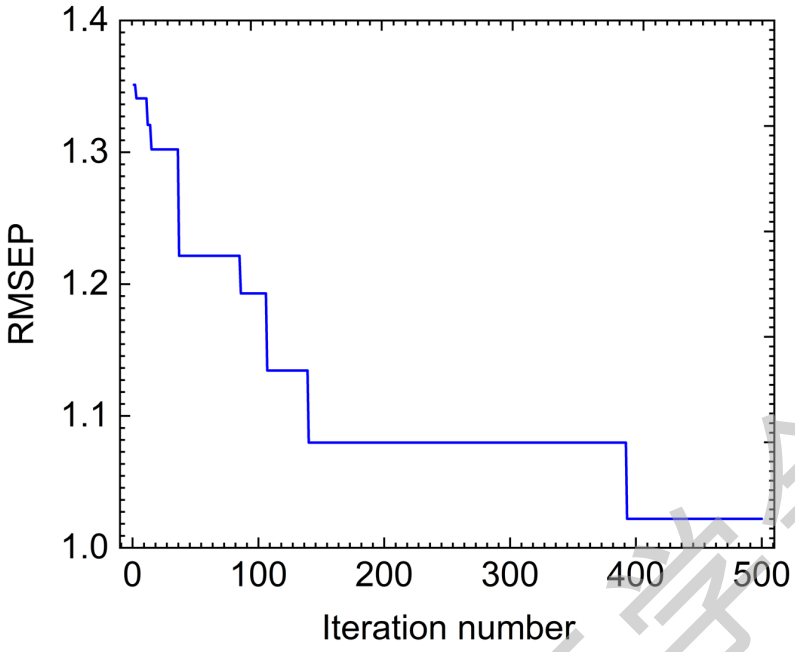

图 9　土壤样品中有机质组分 RMSEP 随迭代次数变化图

3.4 不同变量选择方法选择变量的分布情况

在最佳参数设置下，采用了四个不同的数据集，并运用了 TROA 方法进行变量选择。接着，基于所选变量，构建了偏最小二乘（PLS）模型，并与传统的 PLS 模型性能进行了对比。为了体现 TROA 在复杂样品近红外光谱变量选择中的优越性，还加入了其他几种变量选择技术，包括基于单一指标的变量选择方法 CARS、基于统计学的变量选择方法 UVE、MCUVE、RT 和基于群体智能优化算法的变量选择方法 GWO 和 WOA。这些方法筛选出的变量同样用于建立 PLS 模型，以便进行综合分析。

以药片数据集研究对象，采用不同变量选择方法，选择的变量分布图如图 11 所示。根据图 11 观察到，在处理药片数据集时，TROA 算法所选择的变量数量明显少于 UVE、MC-UVE、RT、GWO 和 WOA 等变量选择方法，同时也比 CARS 这种变量选择方法选取的变量要少。此外，TROA 算法选出的变量在各波段都有，说明选择的变量具有代表性。

对于基于统计学的变量选择方法，它们在选取变量的总体趋势上呈现出较高的一致性。对于群体智能优化的变量选择，GWO 和 WOA 在选取变量上存在较高的一致性。

然而，TROA 和 CARS 在选择变量上表现出了差异性。综合分析可知，在药片数据的变量选择过程中，TROA 算法不仅在选择变量的数量上占有优势，其性能也明显超过了其他几种变量选择方法。

因此，可以得出结论，在处理药片数据时，TROA 算法在变量选择方面展现出了其优越

性，既能够有效减少所需变量的数量，又能保持选变量的多样性和均衡性，从而可能提供更高效的数据分析结果。此外，对其他三种数据集也进行类似的分析，得到了相同的结论。

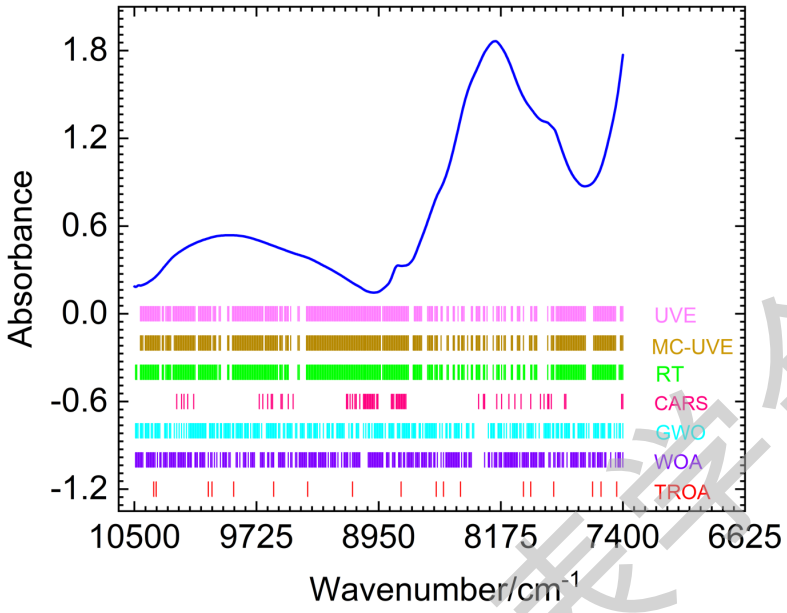

图 11　药片数据集训练集的平均光谱及采用不同变量选择方法选择变量的分布图

3.5 预测结果

在最优参数下，将霸王龙优化算法分别应用于四种数据集的近红外光谱变量选择。利用筛选出的光谱变量构建 PLS 校正模型，同时将其与全光谱建立的 PLS 校正模型以及六种变量选择方法相比较，包括无信息变量消除法（UVE）、蒙特卡罗-无信息变量消除法（MC-UVE）、随机检验（RT）、竞争性自适应加权采样（CARS）、灰狼优化算法（GWO）和鲸鱼优化算法（WOA）。利用 RMSEP 和相关系数（R）来评价模型的性能，其中 RMSEP 表示预测值和真实值的偏差，R 表示预测值和真实值的相关性，RMSEP 值越小、R 值越大均表明模型具有较好的预测性能。

表 1　四组数据采用不同变量选择方法后 PLS 建模结果

数据集	方法	变量数	RMSEP	R
药片	PLS	404	0.4752	0.9336
	UVE-PLS	295	0.4630	0.9369
	MCUVE-PLS	295	0.4591	0.9380
	RT-PLS	305	0.4559	0.9389
	CARS-PLS	63	0.5296	0.9216

数据集	方法	变量数	RMSEP	R
	GWO-PLS	266	0.4305	0.9457
	WOA-PLS	277	0.4237	0.9474
	TROA-PLS	18	0.4200	0.9510
六元调和油	PLS	4148	8.5219	0.7023
	UVE-PLS	225	5.8557	0.8868
	MCUVE-PLS	150	5.1698	0.8943
	RT-PLS	145	5.4055	0.8839
	CARS-PLS	33	7.4184	0.8265
	GWO-PLS	2706	6.6854	0.8157
	WOA-PLS	2122	6.4439	0.8211
	TROA-PLS	24	5.1399	0.9281
橘汁	PLS	700	4.3936	0.8132
	UVE-PLS	455	4.1961	0.8289
	MCUVE-PLS	400	4.1337	0.8354
	RT-PLS	215	4.2903	0.8116
	CARS-PLS	17	3.9988	0.8435
	GWO-PLS	433	4.0683	0.8379
	WOA-PLS	323	3.9293	0.8528
	TROA-PLS	15	3.7120	0.8771
土壤	PLS	1050	1.4403	0.9736
	UVE-PLS	515	1.6458	0.9624
	MCUVE-PLS	420	1.8559	0.9495
	RT-PLS	485	1.7778	0.9539
	CARS-PLS	63	1.9781	0.9490

数据集	方法	变量数	RMSEP	R
	GWO-PLS	683	1.2579	0.9782
	WOA-PLS	681	1.2119	0.9809
	TROA-PLS	55	1.0763	0.9854

根据表 1 的数据分析，可以观察到 TROA-PLS 在处理不同数据集时，包括药片数据、六元混合油数据、橘汁数据和土壤数据，所选择的变量数量明显少于其他几种变量方法。更为重要的是，TROA-PLS 在这些数据集中不仅选择了较少的变量，而且对应的 RMSEP 值更低，R 值更高，表明 TROA-PLS 模型的预测性能更为出色，且模型具有更高的精度。这些数据的结果表明了 TROA 在变量选择和模型建立方面的性能和效率，展现了其优越性。

4 结论

本研究首次将霸王龙优化算法离散化并与偏最小二乘结合，应用于药片、六元调和油、橘汁和土壤四组近红外光谱数据集的变量选择以及组分含量预测。采用预测均方根误差作为算法适应度指标。针对算法中偏最小二乘因子数、离散化函数和迭代次数 3 个参数进行优化，将优化后的算法应用于药片、六元调和油样品、橘汁样品和土壤样品的光谱数据，以进行变量选择。通过筛选出的变量，构建了偏最小二乘（PLS）预测模型，并分别预测了样品中的活性成分、葵花籽油成分、蔗糖成分以及有机质成分的含量。经过参数优化后，TROA 在离散化过程中更加高效。通过对 PLS 因子数和迭代次数的细致调整，成功提升了算法的性能。为了进行变量选择，利用优化后的算法对光谱数据进行了深入分析。通过筛选出最具代表性的变量，构建了一个偏最小二乘（PLS）预测模型。该模型能够准确地预测样品中的活性成分、葵花籽油成分、蔗糖成分和有机质成分的含量。将 TROA-PLS 应用于药片、六元调和油、橘汁和土壤样品的近红外光谱定量分析，与全光谱 PLS 以及六种变量选择方法相比较，结果表明，相较于全光谱 PLS 以及六种变量选择方法相比较，TROA-PLS 所选取的变量数最少且所选变量更具代表性，且 RMSEP 值更低、R 值更高，因此，霸王龙优化算法是一种有效和高效的复杂样品近红外光谱变量选择方法，为分析化学领域的数据处理和模型构建提供了一种高效、可靠的优化工具。

参考文献：

[1] P. Ravichandiran, V.K. Kaliannagounder, A.P. Bella, et al. Simple Colorimetric and

Fluorescence Chemosensing Probe for Selective Detection of Sn2+ Ions in an Aqueous Solution: Evaluation of the Novel Sensing Mechanism and Its Bioimaging Applications. Analytical Chemistry, 2021, 93(2): 801-811.

[2] R. Gavard, D. Lozano, A. Guzman, et al. Rhapso: Automatic Stitching of Mass Segments from Fourier Transform Ion Cyclotron Resonance Mass Spectra. Analytical Chemistry, 2019, 91(23): 15130-15137.

[3] L. Coic, P.Y. Sacre, A. Dispas, et al. Selection of Essential Spectra to Improve the Multivariate Curve Resolution of Minor Compounds in Complex Pharmaceutical Formulations. Analytica Chimica Acta, 2022, 1198: 339532.

[4] B. Lima, P.L. Hayes, P.M. Wood-Adams. Lamellar Orientation at the Surface of Isotactic Polystyrene Thin Films Analyzed by Sum Frequency Generation Spectroscopy. Analytica Chimica Acta, 2023, 1248: 340904.

[5] D. Wu, P. Nie, Y. He, Z. Wang, H. Wu. Spectral Multivariable Selection and Calibration in Visible-Shortwave Near-Infrared Spectroscopy for Non-Destructive Protein Assessment of Spirulina Microalga Powder. Int. J. Food Prop, 2023, 16: 1002-1015.

[6] W. Cai, Y. Li, X. Shao. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems, 2008, 90: 188-194.

[7] I.G. Chong, C.H. Jun. Performance of Some Variable Selection Methods When Multicollinearity Is Present. Chemometrics and Intelligent Laboratory Systems, 2005, 78: 103-112.

[8] M. Sjostrom, S. Wold, W. Lindberg, et al. A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables. Analytica Chimica Acta, 1983, 150: 61-70.

[9] H.D. Li, Y.Z. Liang, Q.S. Xu, et al. Key Wavelengths Screening Using Competitive Adaptive Reweighted Sampling Method for Multivariate Calibration. Analytical Chimica Acta, 2009, 648: 77-84.

[10] V. Centner, D.L. Massart, B.M. Vandeginste, et al. Elimination of Uninformative Variables for Multivariate Calibration. Analytical Chemistry, 1996, 68: 3851-3858.

[11] A.D. Belegundu, J.S. Arora. A study of mathematical programming methods for structural

optimization. Part I: theory. Int J Numer Methods Eng, 1985, 21(9): 1583–1599.

[12] W.S. Cai, Y.K. Li, X.G. Shao. A Variable Selection Method Based on Uninformative Variable Elimination for Multivariate Calibration of Near-Infrared Spectra. Chemometrics and Intelligent Laboratory Systems, 2008, 90: 188-194.

[13] H. Xu, Z.C. Liu, W.S. Cai, et al. A Wavelength Selection Method Based on Randomization Test for Near-infrared Spectral Analysis. Chemometrics and Intelligent Laboratory Systems, 2009, 97(2): 189-193.

[14] D.S.V.S. Manohar, S. Padarbinda, C.P. Kumar. Tyrannosaurus Optimization Algorithm: A New Nature-Inspired Meta-Heuristic Algorithm for Solving Optimal Control Problems. e-Prime - Advances in Electrical Engineering. Electronics and Energy, 2023, 5: 100243.

[15] S.M. Mirjalili, A.L. Mirjalili. Grey Wolf Optimizer. Advances in Engineering Software, 2014, 69: 46-61.

[16] X. Cui, J. Zhu, L. Jia, et al. A novel heat load prediction model of district heating system based on hybrid whale optimization algorithm (WOA) and CNN-LSTM with attention mechanism. Energy, 2024, 312133536-133536.

[17] E. Gamillo. New Study Finds T. Rex Walked at a Slow Pace Of Three Miles Per Hour. https://www.smithsonianmag.com/smart-news/new-study-finds-that-t-rex-walked-at-slow-pace-of-3-miles-per-hour-180977572/.

[18] M. Dyrby, S.B. Engelsen, L.M. Nørgaard, et al. Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. Applied Spectroscopy, 2002, 56: 579-585.

[19] 刘丽美. 基于近红外光谱的多元调和油定量分析方法研究. 天津工业大学, 2016.

[20] N. Benoudjit, E. Cools, M. Meurens, et al. Chemometric Calibration of Infrared Spectrometers: Selection and Validation of Variables by Non-Linear Models. Chemometrics and Intelligent Laboratory Systems, 2003, 70: 47-53.

[21] R. Rinnan, A. Rinnan. Application of Near Infrared Reflectance (NIR) and Fluorescence Spectroscopy to Analysis of Microbiological and Chemical Properties of Arctic Soil. Soil Biology & Biochemistry, 2007, 39: 1664–1673.

通讯作者简介：

卞希慧，女，1983 年生，天津工业大学化学工程与技术学院教授，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

杨文博，男，2001 年生，硕士研究生，研究方向为基于群体智能优化的化学计量学算法研究。

E-mail：18822085929@163.com