光谱多元建模中代表性样本选择方法研究综述

张可欣，张强，刘鹏，卞希慧*

（天津工业大学 化学工程与技术学院，天津 300387）

摘要：在多元建模中，模型性能很大程度上受到建模所用样本的影响。随着分析仪器的发展，样本光谱信息的获取越来越容易。样本量不很大时建模样本的增多可以提高模型的预测性能。然而，过多的样本可能导致冗余信息，而且样本目标值的测量通常费钱且耗时，提高模型性能的代价高昂。因此，需要从大量样本中选择出代表性样本。本综述总结了化学计量学领域提出的 19 种代表性样本选择方法，并首次将这些方法分为基于抽样的方法、基于距离的方法、基于聚类的方法、基于变量选择的方法、基于实验设计的方法、基于奇异样本检测的方法和基于预处理的方法等七类。并对每种方法的原理、优缺点以及适用范围进行总结，为选择代表性样本方法提供参考。

关键词：化学计量学；光谱分析；多元建模；样本集选择

1 引言

光谱分析结合化学计量学方法已成为复杂样本快速定性定量分析的重要手段。在采用化学计量学进行复杂样本的光谱分析时，首先需要获得大量的实际样本，并采集样本的光谱数据。然后通过光谱信息从最初采集到的大量样本中选择出简化的样本集。再通过传统的方法测得这些简化样本集中样本目标值的含量。有时在可行性研究中，样品是通过一定的实验设计方式按照给定的比例配置而成，这时的目标值通常是已知的且没有重复，不需要从大量样本中选择简化的样本集的步骤。不论是实际样本选择出的简化样本集还是实验设计的样本集，奇异样本的存在都会破坏模型的预测效果。因此在样本测量完成后，需要通过奇异样本识别算法找到并剔除奇异样本。再将样本集进一步划分为校正集（calibration set）和验证集（validation set），其中校正集用来建立模型，验证集用来验证模型的效果。依次对校正集样本光谱进行预处理和变量选择，最后建立多元校正或化学模式识别模型。将验证集样本通过同样的光谱预处理、变量选择，将选择后的变量代入到模型中得到预测值，整个过程如图 1 所示。

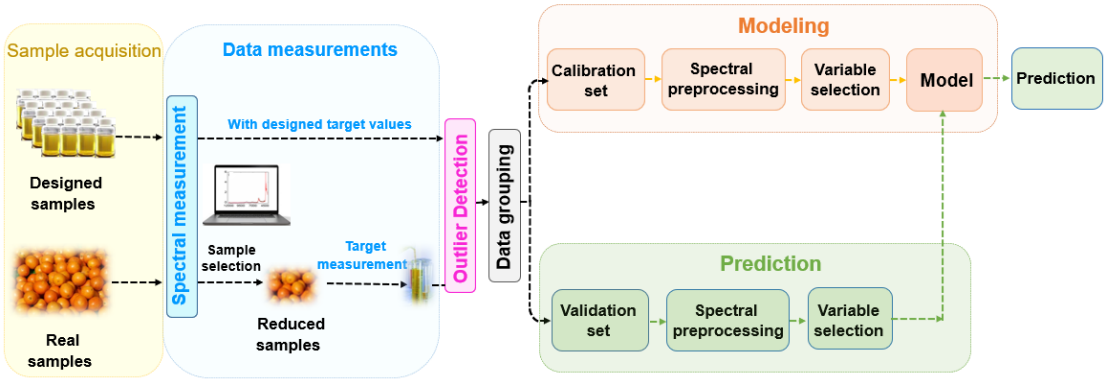

图 1　光谱分析结合化学计量学结合对复杂样品定性定量分析的过程

在上述化学计量学建模过程中，有两个过程都涉及到代表性样本的选择。第一个过程是从大量样本集中选择出简化的样本子集，第二个过程是从简化的样本集中选择出代表性的样本作为校正集，如图 2 所示。因此，代表性样本的选择是化学计量学的一个重要研究内容。从 1969 年 KS 算法被提出后，又有大量的代表性样本选择方法被提出，但是没有对这些方法的系统的综述、分类以及比较。

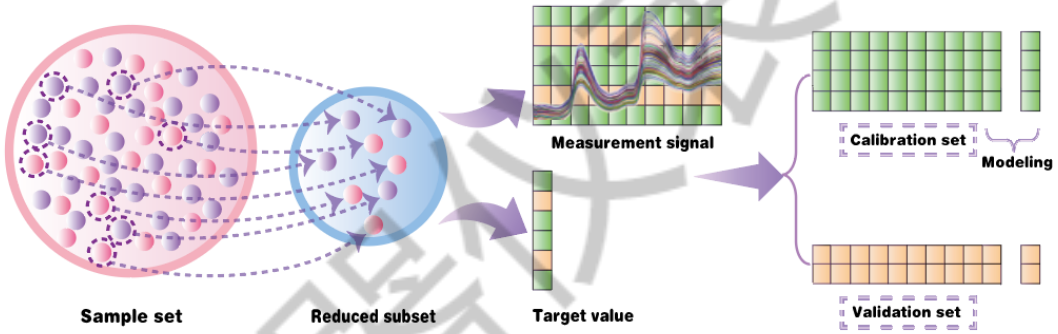

图 2　代表性样本选择过程

本文首次将代表性样本选择方法根据其原理分为七大类，即基于抽样的方法、基于距离的方法、基于聚类的方法、基于实验设计的方法、基于变量选择的方法、基于奇异样本检测的方法和基于预处理的方法，如图 3 所示。系统综述了每种代表性样本选择方法的原理，并对它们的优缺点进行了比较。

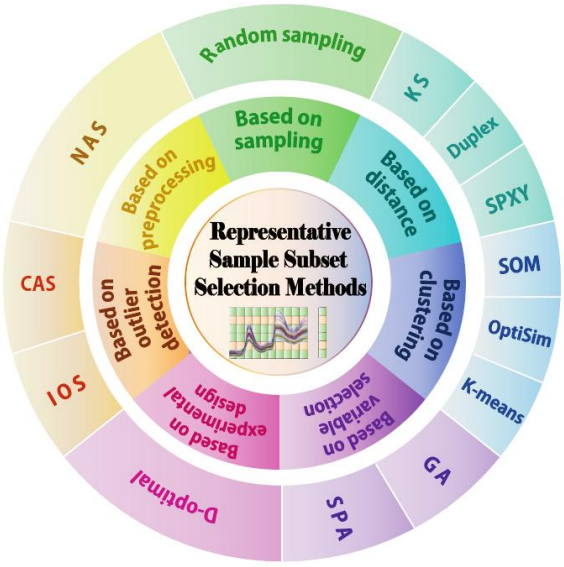

图 3　代表性样本选择方法

2　基于抽样的样本选择方法

随机抽样（Random Sampling，RS）是第一个对样本集进行划分而不需要特殊处理的方法，其基本思想是从样本集中随机抽取一部分样本组成校正集，其余样本作为验证集。该选择方法简单，无需对数据进行排序、筛选或计算，但每次随机挑选校正集样本可能存在很大差异，不能保证所选样本具有代表性。

3 基于距离的样本选择方法

3.1 Kennard-Stone 算法

Kennard-Stone 算法（KS）是 Kennard 和 Stone[1]在 1969 年提出的一种代表性样本选择方法。该方法以计算样品测量的光谱信号间的欧式距离为基础，旨在选择出覆盖范围广、且分布均匀的代表性样本。首先，使用欧式距离计算得到距离最大的两个样本，并将它们放入校正集。之后，分别计算剩余样本与已选样本的距离，从剩余样本中选择样本间距离最小的样本，并从这些距离最短的样本中选择距离最大的样本，将其放入校正集。最后，重复以上步骤，直至校正集样本达到预定的数量。

KS 算法的优点是样本在校正集中按空间距离均匀分布，建立的模型具有良好的预测性能。然而，KS 方法在高浓度范围内效果最好，在低浓度范围内，光谱的差异难以区分，会导致性质空间分布不均匀。

3.2 Duplex 算法

Duplex 算法是 Snee 等人[2]在 1977 年提出的一种基于 KS 的改进方法。Duplex 算法通过

比较样本光谱之间的欧氏距离，对校正集和验证集的样本同时选择，以保证校正集和验证集都具有代表性。首先，使用欧氏距离在所有样本中找出最远的两个样本并放入校正集中。然后，在剩余样本中找出与校正集样品距离最远的样本放入验证集中。其次，选择离校正集中最初选的两个最远的样本放入到校正集中，选取离验证集中最初选取的最远的样本放入到验证集中。最后，直到验证集满足数目要求，将剩余样本加入到校正集。

3.3 Puchwein 算法

Puchwein 算法是奥地利化学计量学专家 Puchwein[3]在 1988 年提出的一种基于马氏距离的代表性样本选择方法。首先，根据所有样本点至样本集中心点的马氏距离，选择距离最远的点加入到校正集中。其次，设定一个阈值，规定比阈值小的样本均被排除在外。然后，选择剩余样本中，离中心点距离最远的点（而且要比阈值大）加入校正集中，重复此过程。直至没有样本符合条件。该方法校正集样本的数目由阈值的大小来决定，如果阈值太小，则样本数过多，否则，样本数则过少。因此必须设定不同的阈值，重复几次上述实验过程，样本数目才能达到期望值。

3.3 SPXY 算法

KS 算法仅考虑了光谱 X 的信息，没有考虑参考值 Y 的影响。为了解决 KS 算法的缺点，巴西的 Galvão 等人[4]在 2005 年提出的一种基于 X-Y 联合距离（Sample set Partitioning based on joint X–Y distances，SPXY）的校正集和验证集分组的方法。该算法是在 KS 算法的基础上发展起来的，其中光谱空间之间的距离和所测组分浓度的空间距离被用作选择最佳样本集作为校正集的参考。在计算样本 u 和样本 v 之间的距离时，采用了同时考虑光谱数据变量 x 和目标参考值变量 y 的新的距离定义 $d_{xy}(u,v)$。

$$d_{xy}(u,v)=\frac{d_x(u,v)}{max_{u,v\in[1,n]}d_x(u,v)}+\frac{d_y(u,v)}{max_{u,v\in[1,n]}d_y(u,v)};\mathrm{u,v}\in[1,\mathrm{N}] \quad (1)$$

式中 $d_x(u,v)$为以光谱 x 为特征参数计算的样本 u 和 v 之间的欧氏距离；$d_y(u,v)$为以目标参考值 y 为特征参数计算样本 u 和 v 之间的距离；n 为样本的总数目。由于 SPXY 必须使用光谱和目标值，而目标值的测量通常费时费钱，因此，该方法主要用于校正集和验证集分的分组，很少用于从大量实际样本选择简化的样本集。

3.4 基于谱相似性的样本选择算法

基于谱相似性的样本选择算法（Sample Selection based on Spectral Similarity，FS）是孙等人[5]在 2021 年提出的代表性样本选择方法。首先，计算未知样本 $X_t(i)$与样本池 X 中第 j 个的欧氏距离 $d_x(i,j)$，进行子集划分。其次，为验证集选择最相似的样本，去除验证集中的

重复样本。最后，计算剩余样本 X_r 中 $X_t(i)$和第 j 个样本之间的欧氏距离。候选样本数和验证集样本数分别计算如下：

$$N_{can} = n \times N_t - N_{rcan} \quad (2)$$

$$N_v = g \times N_t - N_{rv} \quad (3)$$

N_{can}、N_v 和 N_t 分别表示候选、验证和独立测试集样本的数量。N_{rcan} 和 N_{rv} 分别是候选和验证集样本的重复冗余样本数。

在复杂样本的光谱定性定量分析过程中，一开始提出的代表性样本选择方法是基于距离的方法。随着其它化学计量学方法的发展，一些开始应用于聚类、预处理、变量选择以及奇异样本检测的方法也逐渐引入到代表性样本选择中，由此发展了跟多的代表性样本选择方法。

4 基于聚类的样本选择方法

聚类分析（Clustering Analysis）又称为无监督的化学模式识别，是把相似的样本通过静态分类的方法分成多个类别的分析过程。主成分分析（Principal Component Analysis，PCA）、系统聚类分析（Hierarchical Cluster Analysis，HCA）、K 均值（K-Means）聚类算法、自组织映射（Self Organizing Map，SOM）等都是常用的聚类算法。目前 Naes、自组织映射、最优 K 相异性算法和 K 均值聚类都被引入到代表性样本选择中。

4.1 Næs 算法

Næs 算法是由 Isaksson 和 Næs[6]于 1990 年提出的一种基于聚类分析原理的代表性样本选择方法。其思想是，聚类一起覆盖整个感兴趣的空间，并且当样本在同一聚类中时，它们包含相似的信息。因此，最好从每个样本集群中只选择一个样本，而不是从少数样本集群中使用几个样本。Næs 算法先把光谱进行聚类分析，设定类的数目与所期望的校正集样本数目一致。然后从每一类内，选择离中心最远的样本点加入至校正集中。

4.2 自组织映射算法

自组织映射（Self Organizing Map，SOM）是一种由 Kohonen 等人于 1981 年提出的一种竞争学习网络。该方法于 1996 年被比利时 Massart 课题组的吴等人[7]应用于代表性样本选择。SOM 用于代表性样本选择的原理是同一神经元中样本的性质类似，从每个神经元中选择一些样本作为校正集样本，将该神经元中的剩余样本作为验证集样本。所有簇内的校正集样本相加即为整个数据集的校正集。SOM 对校正集和验证集进行划分的主要目的是将样本从 m 维空间映射到二维空间。当样本在原始空间中具有相似属性时，它们会被映射到同一个节点上，并从映射到同一个节点上的样本中随机选择校正集。

4.3 最优 K 相异性算法

最优 K 相异性算法（Optimizable K-Dissimilarity，OptiSim）是 Clark[8]于 1997 年提出的一种代表性样本选择算法。最优 K 相异性算法涉及 3 个参数：K 定义每一次迭代中子样本集的大小；R 定义一个有效的候选样本与任何一个已经选定的样本之间所允许的最小相似性；M 为所选的代表性子集的总数目。首先，从数据集中随机选择一个样本，在剩下的样本集中创建一个候选样本缓冲池、一个回收站和子样本集。其次，从候选缓冲池中随机取出一个样本，如果它到任何一个已选定样本的相似性小于 R，引入子样本集，否则放入回收站。最后，重复上一个步骤，直到子样本集包含 M 个样本或者候选缓冲池耗尽。

OptiSim 的优点是能够快速选择既有代表性又有多样性的样本，并能灵活控制代表性和多样性之间的平衡。改变 K 值是调整代表性和多样性之间的平衡的一种方法。K 值越小，样本的代表性越强；K 值越大，样本的多样性越强。

4.4 K 均值聚类算法

K 均值聚类算法由 Daszykowski 等人[9]于 2002 年引入到代表性样本选择中。首先，将样本集 X 分成 k 个簇，选择要找到的簇数 k。其次，将样本随机分配到 k 个簇中，计算每个簇的均值 X_j。然后，将每个样本重新分配给它最近的 X_j，然后根据公式（4）计算每个簇的中心 X_j 与实验样本之间的欧式距离平方和 E。重复上一个步骤，直到 E 收敛，即最相似。

$$E = \sum_{j=1}^{k} \sum_{i=1}^{m} (x_i - \bar{x}_j)^2 \qquad (4)$$

5 基于实验设计的样本选择方法

实验设计是在多因素多水平下，通过数学设计寻找最佳实验条件的方法。很多实验设计方法已被提出，包括全因子设计、中心复合设计、D-最优设计和混合设计。利用这些实验设计方法，不仅可以寻找到全局最优条件，而且可以大大减少实验次数。将实验设计的思路进入到代表性样本选择，可以用最少的样本数量，增加确定实验的最佳子集（即校正集）的可能性。

5.1 D-最优设计

西班牙的 Ferre 和 Rius[7]于 1996 年将 D-最优设计（D-Optimal）引入到代表性样本选择中。该方法的原理是选取线性回归模型的信息矩阵 $|X^tX|$ 的行列式最大的样本。矩阵是包含 n 个样本和 m 个变量的方差协方差矩阵，其中 n 是要选择的样本数。当所选样本跨越整个数据空间时，该矩阵的行列式最大。首先，因为 m 远大于 n，所以对样本进行居中处理后，进行 PCA 预处理，使变量数量减少到 n-1 个，现在 X 变成了 n 个样本和 n-1 个变量的分数

矩阵。然后，应用 D 最优设计对$y = \sum \beta_i x_i + e$的线性模型进行样本选择，其中 x_i 是潜在变量 i。这种线性模型仅用于样本的选择。对每个类分别执行，从每个类中选择 3/4 的样本，放在一起作为校正集。如果 3/4 的样本不是整数，则使用与随机选择相同的方法将其舍入。当变量数量大于样本数量时，由于信息矩阵的奇异性，不能直接应用 D-最优设计。

6 基于变量选择的样本选择方法

由于光谱数据含有上千个波长点，并不是所有的波长变量都与目标组分相关。因此需要从采集的波长变量中选择代表性样本信息的重要波长，删除冗余波长。合适的光谱变量选择可以增加模型的解释性，简化模型并提高模型的预测精度。大量的变量选择方法被提出，如模拟退火、遗传算法、连续投影算法、无信息变量消除、随机检验、竞争性自适应重加权算法和灰狼算法等。现如今，遗传算法和连续投影算法已应用于代表性样本选择中。

6.1 遗传算法

遗传算法（Genetic Algorithms，GA）于 1998 年被意大利的 Forina 等人[10]引入到代表性样本的选择当中。GA 旨在优化两个目标：最大化特征子集的分类准确性和最小化所选特征的数量。为了获得具有代表性的子集，使用两个拟合度函数，如公式(5)所示，第一个拟合度函数表示所选子集内样本的不相似度，使用欧氏距离作为不相似度的度量。

$$Dis = \{2/N(N-1)P\}\left\{\sum_{i<j}^{N}\sum_{l}^{P}\left(X_{il} - X_{jl}\right)^2\right\}^{1/2} \quad (5)$$

N 是子集中样本的个数，P 是描述符的个数。第二个适应度函数是积矩相关系数 PMCC 值的平均值，用描述符 MP 的 PMCC 的平均值作为拟合函数。

$$MP = (1/P)\sum_{i}^{all\ variables} r_i \quad (6)$$

GA 选择良好目标函数的方法优于传统方法。但其收敛速度较慢，容易陷入局部最优；算法的随机性比较大，多次运行的结果不均匀；容易陷入过早成熟或整个种群进化停滞。

6.2 连续投影算法

连续投影算法（Successive Projections Algorithm，SPA）是 Filho[11]在 2004 年提出的一种代表性样本选择方法。该算法在选择代表性样本时会同时考虑光谱数据 X 和目标参考值变量 y，根据分析中涉及的化学物种的光谱剖面选择样本。

SPA 是一种迭代正演选择方法，在仪器响应矩阵 X_{cal} 上运行，其行和列分别对应于校正样本和光谱变量。SPA 从一列 x_0 开始，确定其余各列中哪一列在与 x_0 正交的子空间 S_0 上的投影最大。这一列（用 x_1 表示）可视为包含 x_0 中未包含的最大信息量的一列。在下一次迭代中，SPA 将分析局限于子空间 S_0，将 x_1 作为新的参考列，并继续上述步骤。SPA 选择

的样本子集冗余度小，而且考虑了 X-Y 相关性，因此所选样本在建模过程中具有代表性。对于减少多元建模中的实验和计算工作量，以及不同仪器之间的校正转移都很有价值。

6.3 最优预测校正子集方法

荷兰的 Andries 和比利时的化学计量学专家 Heyden[12]在 2023 年提出一种名为最优预测校正子集（Optimally Predictive Calibration Subset，OPCS）的代表性样本选择方法，该方法通过选择具有最优预测能力的最佳拟合样本来减少校正集。首先，利用一种名为使用变量的 PLS 回归系数 bj 的显著性的最终复杂度自适应模型（Final Complexity Adapted Models using significance of the PLS regression coefficients b*j* of the variable *j*, FCAM-SIG）的变量选择方法后得到原始大的全局校正集的全局 PLS 模型。然后，根据残差递增对全局校正样本进行排序后，选择排序后的校正集的不同放大分数。对于每个分数，通过交叉模型验证（Cross Model Validation，CMV）确定最优预测能力和相应的最优 PLS 复杂度。在对所有分数执行 CMV 后，确定具有最佳拟合样本和最佳预测能力的分数。

OPCS 方法的优点是子集中不包含奇异样本，所选择的最佳拟合样本不需要代表全局校正集，而只需支持基于 OPCS 的模型，因此，OPCS 模型中的样本数量通常比传统的代表性样本选择方法所选的样本数量少。

7 基于奇异样本检测的样本选择方法

奇异样本（outlier）有时也称为异常值、不规则点、离群点或界外点，至今没有严格的定义，一般是指那些落在总体之外的样本向量。奇异样本的存在会在一定程度上影响甚至改变整体数据的分布趋势，从而影响校正模型的准确性。因此，奇异样本检测也是化学计量学的一个重要研究内容。奇异样本的识别方法有残差法、最小体积椭球估计法、隔离森林等。其中简单区间计算方法、隔离林奇异值检测方法和组合分析信号方法可以用来选择代表性样本。

7.1 简单区间计算方法

简单区间计算（Simple Interval Calculation，SIC）方法是由俄罗斯化学计量学专家 Rodionova 和 Pomerantsev[13]在 2008 年提出一种用于代表性样本选择的方法。该方法明确地选择校正集中最重要的样本，并使用该子集进行模型开发，而不会显著降低预测能力。首先，随机选择一个初始样本。然后，计算每个样本点与起点之间的差距。再根据计算得到的差距，将样本点按照差距的大小排序，通常从最小到最大。最后，选择前 n 个样本点作为代表性样本（n 是根据需要选择的样本数量）。

SIC 方法不仅考虑 X 值，还考虑 y 值。它适用于大型样本集，可以帮助减少样本集的大小，从而减少计算的开销。

7.2 隔离森林奇异样本检测和子集选择方法

隔离森林奇异样本检测和子集选择（Isolation forest Outlier detection and Subset selection，IOS）算法是中南大学的张志敏课题组[14]在 2016 年提出的代表性样本选择方法。隔离森林（Isolation Forest，IForest）最初是由南京大学的周志华等人提出的一种奇异样本检测算法。张等人提出的 IOS 算法可以在隔离森林的基础上同时检测奇异样本和选择有代表性的样本子集。首先，构建一个包含所有样本的 IForest，并设置树的数量为 1000，然后将所有样本放入 IForest 中，重复此过程数次，取所得分数的平均值。然后，根据样本的奇异得分对样本进行排序，从代表性样本中排除奇异样本。最后，选择所有奇异得分大于 0.5 的样本作为奇异样本，从剩余样本中均匀选择所需数量的样本作为代表性的正态普通样本。

IOS 可以剔除奇异样本，选择没有 y 值的代表性样本。只测量其 y 值，可以缩短时间，节省资源。此外，它还可以用于检测预测样本中的异常值，减少模型的冗余度和易于更新模型。

7.3 组合分析信号算法

组合分析信号（Combined Analytical Signal，CAS）是 Pomerantsev 和 Rodionova[15]在 2023 年提出一种用于选择代表性样本的方法。CAS 是一个已知分布的变量，可以作为多变量数据的分析信号，可以进一步扩展到多块数据。代表性样本的选择包含两个问题，第一个问题是简化集的选择（Reduced Set Selection，RSS）。首先，从顶部选择几个样本作为校正集。然后，建立模型进行预测，并获得质量特征（通常是 RMSEP）。如果质量不令人满意，在校正集中增加额外的样本，重复直到收敛。第二个问题是测试集的选择（Test Set Selection，TSS）。首先，将所有样本分成两个子集：前 10%构成“极端”子集，其余构成“常规”子集。再使用随机或其他选择方法，将“常规”样本的所需分数放入测试集，将相同比例的“极端”样本添加到测试集中，从底部选择它们。然后，使用由未包含在测试集中的样本组成的校正集开发模型，预测测试集并检查可能在那里找到的奇异值，将测试集的奇异值替换为来自校正集的“常规”样本。最后，通过增加模型的复杂性，重复上一步骤，以确保测试集不包含奇异。该方法的主要优点是不需要大量的计算和时间成本。

8 基于预处理的样本选择方法

实验采集到的原始光谱除了包含与样本相关的有用信息外，往往也掺杂着干扰信息，包括随机噪音、背景干扰、杂散光以及测样器件引起的光谱差异，这对校正模型的质量和未知样本预测的准确度将产生严重的影响。因此，在建立化学模式识别和多元校正模型前，通常需要对光谱进行预处理以消除各种干扰。光谱预处理方法包括多元散射校正、标准正态变量、正交信号校正、净分析物信号、小波变换等。其中基于净分析物信号的方法已经用于代表性样本选择。

8.1 净分析物信号方法

净分析物信号（Net Analyte Signal，NAS）是由东北大学的贺忠海课题组[16]于 2018 年引入到代表性样本选择的方法。NAS 方法能够将光谱信息转换为与浓度相关的度量。首先，使用一组样本的浓度和光谱来计算投影矩阵、NAS 向量和标量值。然后，通过投影矩阵与样本谱的乘积计算候选样本的 NAS 向量。通过范数计算得到 NAS 的标量值。计算候选集与选择的样本集之间的距离，依次将距离最大的样本加入选择的样本集。最后，测量分析物的浓度，使样品可以用作校正样品。

采用 NAS 计算的样本选择方法有效地增加了不需要测量 y 值的样本的代表性。以测量较少原始样本的浓度值为代价，可以获得标量 NAS 值的分布。随后的候选样本通过其 NAS 标量规范值进行评估。从现有样本中选择具有最大 NAS 距离值的样本进入校正集，以确保其代表性。

8.2 半监督选择方法

半监督选择（Semi-Supervised Selection，SS）的代表性样本选择方法也是由东北大学的贺等人[17]在 2018 年提出。首先，使用 KS 方法选择部分样本集，并测量其浓度。其次，根据净分析物信号的标量值分布选择另一部分样本集。如果某个样品的净分析物信号与现有净分析物信号值相比具有明显差异，则将该样品添加到校正集中。然后测量样品中的相关分析物，以便将样品用作校正样品。

SS 方法只需测量具有代表性的样本的浓度，在参考值测量上节省了大量的时间和金钱。当采集到一些有代表性的样本时，可以更新校正模型。样本选择对在线光谱测量有很大的好处，可以方便地获得许多样本的光谱。

本文综述了现有的 19 种的代表性样本选择算法的原理，算法的英文全称、缩写、提出时间、第一作者以及文章的被引次数总结在表 1 中。可以看出，在这七类代表性样本选择方法中，应用最广泛的一类方法是基于距离的方法。KS 算法作为最早被提出的一种代表性样本选择方法，其使用频次远远高于其它算法。Duplex 和 SPXY 也具有较高的使用频率。

表 1　代表性样本选择方法总结

Algorithm name	Abbreviation	Categories	Year	First Author	Citations	Proposed literature
Random Sampling	RS	Sampling-based	unknown	unknown	unknown	unknown
Kennard-Stone	KS	Distance-based	1969	R.W. Kennard, L.A. Stone	3504	[1]
Duplex	Duplex	Distance-based	1977	R.D. Snee	1432	[2]
Puchwein	Puchwein	Distance-based	1988	G. Puchwein	67	[3]
Sample set partitioning based on joint x-y distance	SPXY	Distance-based	2005	R.K.H. Galvão	820	[4]
Sample selection based on spectral similarity	FS	Distance-based	2021	Y. Sun	14	[5]
Næs	Næs	Clustering-based	1990	T. Næs	79	[6]
Self-organizing maps	SOM	Clustering-based	1996	W. Wu	264	[7]
Optimizable K-Dissimilarity Selection	OptiSim	Clustering-based	1997	R.D. Clark	200	[8]
K-means clustering algorithm	K-means	Clustering-based	2022	M. Daszykowski	325	[9]
D-optimal designs	D-optimal	Experimental design-based	1996	J. Ferre	59	[7]
Genetic	GA	Variable	1998	C.P. Millan	14	[10]

algorithm		selection-based				
Successive projections algorithm	SPA	Variable selection-based	2004	H.A.D. Filho	64	[11]
Optimally predictive calibration subset	OPCS	Variable selection-based	2023	J.P.M. Andries	0	[12]
Isolation forest Outlier detection and Subset selection	IOS	Outlier detection-based	2016	W.R. Chen	25	[13]
Simple interval calculation	SIC	Outlier detection-based	2008	O.Y. Rodionova	36	[14]
Combined analytical signal	CAS	Outlier detection-based	2023	A.L. Pomerantsev	1	[15]
Net analyte signal	NAS	Preprocessing-based	2018	Z.H. He	4	[16]
Semi-supervised selection	SS	Preprocessing-based	2018	Z.H. He	11	[17]

9 结论

本文首次系统总结了 19 种的代表性样本选择算法原理、优缺点，并将它们分为基于抽样、基于距离、基于聚类、基于实验设计、基于变量选择、基于奇异样本检测、基于预处理等七大类，为研究者选择合适的化学计量学方法提供依据，也为新的代表性样本选择方法的提出提供思路。

参考文献:

[1] R.W. Kennard and L.A. Stone. Computer aided design of experiments. Technometrics, 1969, 11(1): 137-148.

[2] R.D. Snee. Validation of regression models: methods and examples. Technometrics, 1977, 19(4): 415-428.

[3] G. Puchwein. Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra. Analytical Chemistry, 1988, 60(6): 569-573.

[4] R.K.H. Galvão, M.C.U. Araujo, G.E. José, et al. A method for calibration and validation subset partitioning. Talanta, 2005, 67(4): 736-740.

[5] Y. Sun, M. Yuan, X.Y. Liu, et al. A sample selection method specific to unknown test samples for calibration and validation sets based on spectra similarity. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 258, 119870.

[6] T. Isaksson and T. Næs. Selection of samples for calibration in near-infrared spectroscopy. Part II: Selection based on spectral measurements. Applied Spectroscopy, 1990, 44(7): 1152-1158.

[7] W. Wu, B. Walcmk, D.L. Massart, et al. Artificial neural networks in classification of NIR spectral data: design of the training set. Chemometrics and Intelligent Laboratory Systems, 1996, 33(1): 35-46

[8] R.D. Clark. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. Journal of Chemical Information and Computer Sciences, 1997, 37(6): 1181-1188.

[9] M. Daszykowski, B. Walczak, D.L. Massart. Representative subset selection. Analytica Chimica Acta, 2002, 468(1): 91-103.

[10] C.P. Millan, M. Forina, C. Casolino, R. Leardi. Extraction of representative subsets by potential functions method and genetic algorithms. Chemometrics and Intelligent Laboratory Systems, 1998, 40(1): 33-52.

[11] H.A. Dantas, R.K.H. Galvao, M.C.U. Araújo, et al. A strategy for selecting calibration samples for multivariate modelling. Chemometrics and Intelligent Laboratory Systems, 2004, 72(1): 83-91.

[12] J.P.M. Andries, Y.V. Heyden. Calibration set reduction by the selection of a subset containing the best fitting samples showing optimally predictive ability. Talanta, 2024, 124943.

[13] W.R. Chen, Y.H. Yun, M. Wen, et al. Representative subset selection and outlier detection via isolation forest. Analytical Methods, 2016, 8(39): 7225-7231.

[14] O.Y. Rodionova, A.L. Pomerantsev. Subset selection strategy. Journal of Chemometrics, 2008,

22(11-12): 674-685.

[15] A.L. Pomerantsev, O.Y. Rodionova. Subset selection using combined analytical signal. Microchemical Journal, 2023, 190, 108654.

[16] Z.H. He, Z.H. Ma, J.M. Luan, et al. An active learning representative subset selection method using net analyte signal. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 196, 311-316.

[17] Z.H. He, Z.H. Ma, M.C. Li, et al. Selection of a calibration sample subset by a semi-supervised method. Journal of Near Infrared Spectroscopy, 2018, 26(2): 87-94.