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。

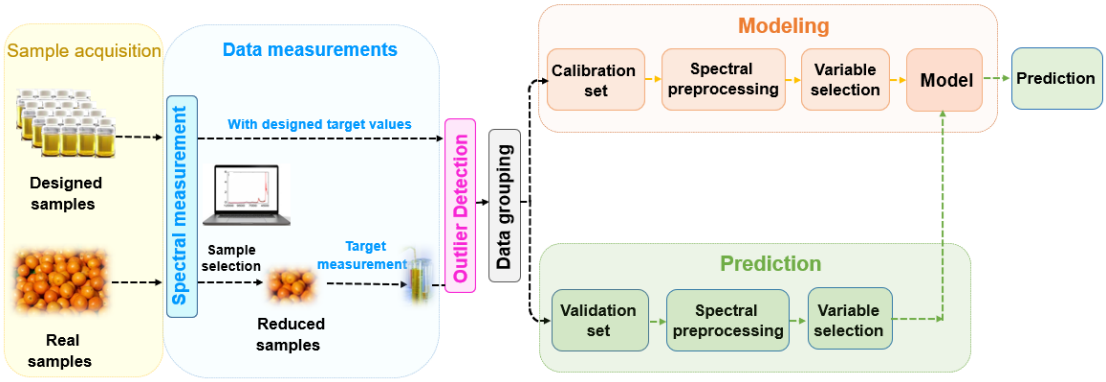

图 1　光谱分析结合化学计量学结合对复杂样品定性定量分析的过程

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。

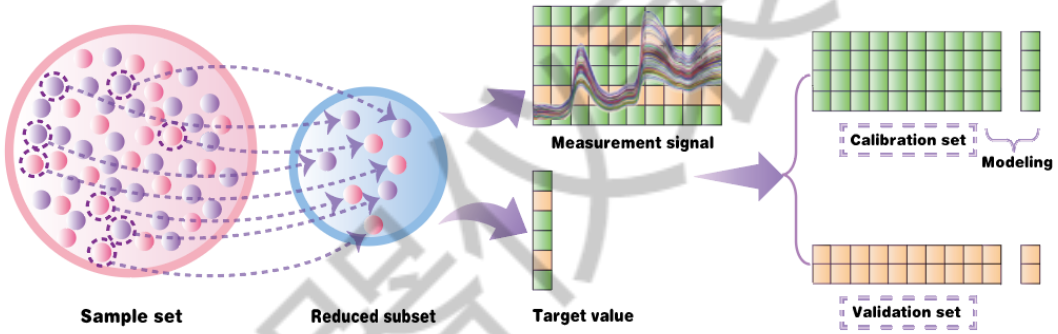

图 2　代表性样本选择过程

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。