



近红外光谱分析方法研究: 从传统数据到大数据

刘言, 蔡文生, 邵学广*

南开大学化学学院分析科学研究中心, 天津 300071

* 联系人, E-mail: xshao@nankai.edu.cn

2014-12-02 收稿, 2014-12-31 接受, 2015-02-03 网络版发表

摘要 近红外光谱分析技术作为一种绿色分析技术, 在许多领域中已得到广泛应用. 随着应用的深入和拓展, 近红外光谱的数据类型逐渐从传统数据变成近红外光谱大数据. 本文总结了近红外光谱的预处理、奇异样本筛选、多元校正和模型转移等技术及其在相关领域的应用. 对近红外光谱大数据分析技术的初步研究, 包括近红外光谱在工业品在线检测、不同批次产品鉴别中的应用以及近红外光谱物联网系统等也进行了综述. 此外, 对于近红外光谱大数据未来的发展及近红外光谱大数据云平台的基本功能、软硬件的设计与开发、建设过程中需要解决的问题等进行了详细阐述.

关键词

大数据
数据分析
近红外光谱大数据
近红外光谱分析
云平台

随着数字时代的到来, 人类对自然和社会认识的进一步加深, 人类的活动空间得到进一步扩展. 高度数字化的生活使得人类在科学研究、互联网应用、电子商务、移动运营等诸多领域均出现了大规模的数据增长, 大数据时代已经来临. 与传统的数据集合相比, 大数据可以通过挖掘和应用创造出巨大的价值, 因此迅速发展成为工业界、学术界乃至世界各国政府高度关注的热点. 大数据以其颠覆性的技术对国家治理模式、企业决策、组织和业务流程以及个人生活方式等均产生了巨大的影响.

1 大数据时代

作为一个新兴的概念, 大数据问题得到了学术界、工业界乃至政府机构的密切关注, 并对其产生浓厚的兴趣. *Nature* 于 2008 年针对大数据推出了专刊“Big Data”^[1], *Science* 于 2011 年推出专刊“Dealing with Data”, 围绕科学研究中大数据的问题展开讨论, 从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面讨论了大数据处理面临的各种问题, 说明了大数据对于科学研究的重要性^[2]. 2012

年, 美国奥巴马政府发布了“大数据研究和发展倡议”, 宣布投资 2 亿美元启动“大数据研发计划”. 这一计划使大数据上升到了国家战略层次, 使之成为各国关注的热点, 之后与大数据相关的研究成果呈现出爆炸性的增长并一直持续至今, 2012 年也因此成为大数据的元年.

2 大数据研究与应用

虽然大数据的重要性得到各行各业的一致认同, 但对大数据本身至今尚无确切统一的定义. 目前普遍认为, 大数据具有“4V”特点, 即数据体量(volumes)巨大、数据类别(variety)繁多、产生与处理速度(velocity)快、价值(value)密度低但可挖掘价值高. 目前, 大数据的研究内容主要集中在大数据采集、处理与集成、分析和解释的过程中产生的一系列热点和难点问题^[3]. 在数据采集技术迅速发展的同时, 数据的集成与整理技术也得到发展. 通过数据集成, 将结构复杂的数据转换为便于处理的数据结构, 通过对数据的整理保证数据的质量及可靠性. 数据分析是大数据处理流程中的核心部分, 通过对数据进行分

引用格式: 刘言, 蔡文生, 邵学广. 近红外光谱分析方法研究: 从传统数据到大数据. 科学通报, 2015, 60: 704–713

Liu Y, Cai W S, Shao X G. Analysis of near-infrared spectra: From tradition to big data (in Chinese). Chin Sci Bull, 2015, 60: 704–713, doi: 10.1360/N972014-01203

析,可以发现数据的价值.传统的数据处理分析方法,包括聚类分析、因子分析、相关分析、回归分析^[4]等仍然可以用于大数据分析.但由于大数据本身数据量大、实时性强的特点,使得传统方法在处理大数据时也存在众多局限性.因此,出现了许多专门针对大数据的分析方法,如散列法^[5]、布隆过滤器(Bloom Filter)^[6]、Trie树^[7,8]等.同时,针对不同类型的大数据,也存在不同的分析方法.如对文本进行分析的自然语言处理(NLP)技术^[9],对Web进行分析的Page Rank法^[10]和CLEVER法^[11],对多媒体进行分析的摘要系统以及对社交网络进行分析的概率法^[12]和线性代数法^[13]等.基于分布式文件系统^[14]、分布式数据库^[15]、批处理技术^[16-18]及开源平台的云技术^[19-21]为大数据分析奠定了基础.通过云技术实现了海量数据的高效存储、大数据的分布管理以及大数据分析的技术与平台.同时,可视化技术在大数据分析中也得到发展并发挥了重要作用^[22,23].

大数据目前已广泛应用于生命科学、医疗、商业、金融等行业中.在医疗领域中,大数据分析用于复杂疾病的早期诊断^[24]、心血管病的远程治疗^[25]、器官移植^[26]、HIV抗体的研究^[27]等,均取得了较好的效果.在生命科学领域,大数据技术用于基因组学^[28]、生物医学^[29]、生物信息学^[30]等方向.商业是大数据应用最广泛的领域,各种大数据的经典应用均来自此领域.如利用大数据关联分析,更准确地了解消费者的使用行为,挖掘新的商业模式^[31];利用大数据进行库存优化、物流优化、供应商协同等工作,缓和供需之间的矛盾、控制预算开支、提升服务等.此外,大数据技术还用于温室气体排放的检测^[32]、政府信息管理^[33]等公共领域.

3 近红外光谱技术

近红外光谱分析技术作为一种绿色分析技术,有分析速度快、操作简单、不需要样本预处理、可实现原位、无损、在线的定性定量分析等优点.实践证明,以近红外光谱分析为主的过程分析技术为工业信息化与自动化的深度融合起到了决定性的作用,它所提供的快速、实时测量信息可以使工农业生产过程保持最优化的控制,在显著提高产品质量的同时,降低生产成本和资源消耗.然而,由于近红外光谱存在吸收强度弱、光谱重叠严重等缺点,要想进行准确的定性定量分析,必须借助化学计量学方法所建立的高质量模型.因此,基于化学计量学的近红外光谱的数据处理技术研究得到了广泛的研究.目前,近红外光谱的数据分析技术研究主要集中在光谱预处理(包括变量筛选)技术、建模样本筛选技术、多元校正技术和模型转移技术等方面.

实验采集到的原始近红外光谱除了包含与样品相关的有用信息外,往往伴随随机噪音、背景干扰、杂散光等干扰信息,对校正模型的质量和未知样品预测的准确度将产生严重的影响.因此,在建立校正模型之前,通过光谱预处理技术消除各种类型的干扰显得十分关键和必要.常用的光谱预处理技术包括多元散射校正(MSC)^[34]、正交信号校正(OSC)^[35-39]、净信号分解^[40]、连续小波变换(CWT)^[41,42]等.如图1所示, MSC可以有效地消除样本颗粒分布不均匀及颗粒大小不同而产生的散射对其光谱的影响,而CWT可以有效地对样品光谱进行背景扣除.实践证明,光谱预处理技术可以有效地消除光谱中的变动背景及散射干扰,处理后的光谱可以有效提高校正模型的质量和未知样品预测的准确度.

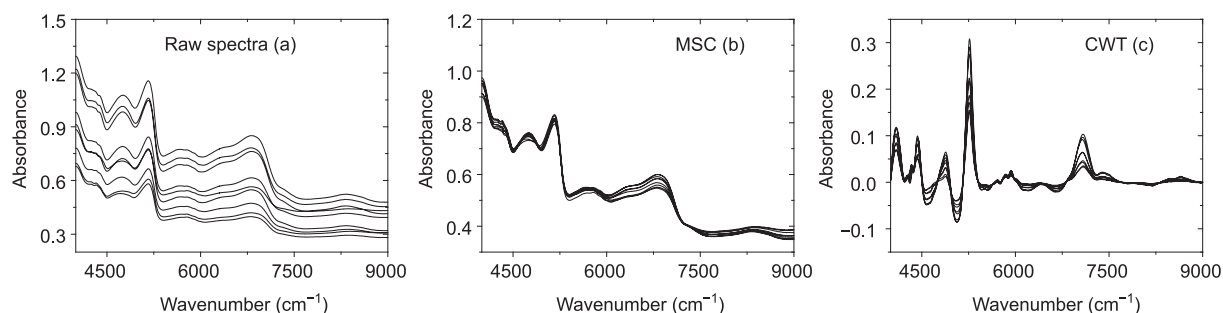


图1 不同预处理技术对光谱处理效果

Figure 1 Effect of different spectral processing methods

波长筛选技术对于得到代表样品信息的重要波长、删除冗余波长、提高模型预测精度和简化模型均具有重要意义,因此针对波长筛选开展了大量研究工作.常用的波长筛选技术包括模拟退火算法(SA)^[43]、遗传算法(GA)^[44~46]、粒子群算法(PSO)^[47,48]、连续投影算法(SPA)^[49~51]、无信息变量选择(UVE)^[52]、竞争性自适应重加权算法(CARS)^[53,54]等.我们在之前的工作中也提出了基于蒙特卡洛的无信息变量选择算法(MC-UVE)^[55]、随机检验算法(RT)^[56]、强影响变量算法(IVs)^[57]、基于特征投影图(LPG)的变量选择算法^[58]和基于局部线性嵌入的变量选择算法^[59]等.图2是LPG方法用于变量选择的示例,其中图2(a)为光谱数据集的特征投影图,其中实心三角形的点为特征投影图拐点处的变量,即根据共线性原则选择的变量;图2(b)中标出了选择的变量在实际光谱(经过连续小波变换处理)中的位置.可以看到,根据LPG法选出的变量均处于光谱的拐点位置,具有很强的共线性,对于建模的贡献较大,可以进行稳定的建模.

在实验和实际生产生活采集到的光谱之中,总

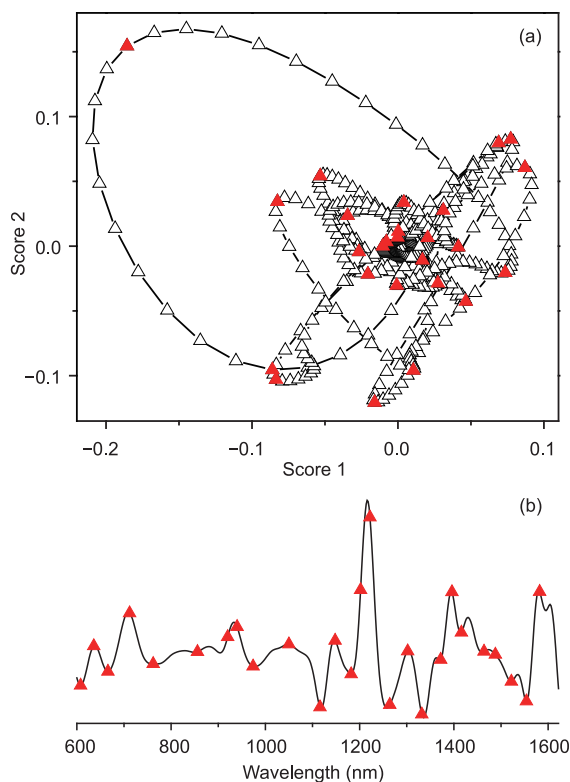


图2 (网络版彩色)(a) 特征投影图及(b) 变量选择结果
Figure 2 (Color online) (a) Latent projective graph and (b) selected variables

会存在一些落在总体样品分布之外的样本,该类样品被称作奇异样本(outlier).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.所以,奇异样本的有效识别对多元校正具有重要意义.奇异样本的识别方法大致可以分为3类:经典识别方法、稳健识别方法和基于统计学的识别方法.经典识别方法都是基于最小二乘估计,包括残差法^[60]、马氏距离^[61]、杠杆值^[62]和主成分得分图^[63]等.稳健识别方法则是通过寻找光谱矩阵稳健的均值和方差或者建立稳健的回归模型来寻找奇异样本.常用的方法包括椭圆多变量修剪法^[64,65]、最小体积椭圆估计^[66]、最小协方差行列式法^[67]、最小半球体积法和半数重采样法^[68]、M估计法^[69,70]等.基于统计学的识别方法则是通过重复地采样分析建模,然后通过统计参数来识别奇异样本,常用的方法为蒙特卡洛交叉验证法^[71].

多元校正技术一直是近红外数据分析的重点.在定量分析中,通过多元校正技术建立稳定的定量模型,才能对未知样品待测组分的浓度实现准确的定量.在所有的多元定量校正技术中,偏最小二乘(PLS)技术应用最为广泛,对于各类型样品的光谱,该技术均表现出良好的回归和预测性能.针对偏最小二乘技术,还提出了大量的改进算法,如iPLS^[72], boosting-PLS^[73], weighted-PLS^[74,75]和 multiblock-PLS^[76]等.其他一些非线性校正技术,如最小二乘支持向量回归机(LS-SVM)、高斯过程回归(GPR)和贝叶斯神经网络(BANN)、独立成分回归^[77]等,也引起了不少关注.类似地,通过多元校正技术建立稳定的定性分析模型,能够实现对未知样品的准确判别.在各类模式识别技术中,比较经典的算法为主成分分析(PCA)技术和偏最小二乘-判别分析(PLS-DA)技术.此外,还有支持向量机、人工神经网络,以及基于Fisher准则的典型变量分析技术(canonical variates analysis)等.对于药品、烟草和油品等大型近红外光谱数据库,基于光谱残差、标准偏差和相关系数等参数建立的库光谱搜索方法被认为是非常有效的识别技术.

为了建立准确和稳定的近红外光谱模型,通常需要准备大量的标准样品,测量它们的近红外光谱和分析目标的参考值.整个建模过程十分费时且花费巨大.而在实际应用中,由于仪器之间的硬件差异或测量条件的差异,一台仪器上建立的模型往往不

能直接用于另一台仪器。即使是同一台仪器,随着时间的推移或关键部件的更换,模型的预测性能也会发生变化。模型转移是解决上述问题的技术之一,通过模型转移将新仪器(源机)的光谱转移为原仪器(目标机)的光谱即可继续使用原来的模型。模型转移技术可以分为校正预测结果、校正光谱和校正模型3类。校正预测结果的技术是建立原仪器与新仪器模型预测结果之间的关系来进行转移,如SBC算法^[78]。校正光谱技术则是建立原仪器光谱与新仪器光谱之间的关系来进行转换,通过2台仪器光谱之间的转换,直接使用原来的模型或者建立新的模型进行预测。目前最常用的光谱校正技术包括DS算法^[79]、PDS算法^[80,81]、shenk算法^[82-84]、SWS算法^[85]和SST算法^[86,87]等。校正模型的方法与上述2类方法不同,核心是利用少数标准样品在目标机和源机上的光谱对源机上的模型进行修正,从而建立目标机上的模型,例如早期报道的B系数校正技术^[88]以及近来提出的基于正则化的校正技术^[89-92]。我们在之前的工作中也针对工业化生产中多台仪器需要同时校正的需求提出了一种多仪器光谱同时校正的模型转移方法^[93]。

4 近红外光谱大数据分析与应用

正是在近红外光谱数据分析技术的支持下,近红外光谱分析已广泛应用于食品、制药、石油化工、生命科学、农业等各个行业。在食品行业,近红外光谱技术用于各种类型食品品质参数的快速或在线分析,以及食品真伪和产地的快速鉴别,如肉品产地和品种的快速鉴定、鸡蛋种类的鉴别、食用油真伪和掺杂的鉴别、水果内、外部品质和产地的检测以及酒类和调味品中理化指标的检测等。在制药行业,近红外光谱能够简便、快速、准确地鉴别中药的种类、产地和真伪,并且能够快速测定中药材中有效成分的含量以及中药辅料的品质。在西药分析上,近红外光谱用于原辅料的真伪鉴别和成品药中的有效成分含量测定等方面。在石油化工业,近红外光谱广泛应用于生物燃料的快速检测分析、聚丙烯物性参数的快速分析和干粉灭火剂有效成分含量的测定等。在农业上,近红外光谱用于农作物种子品质的识别、土壤成分的分析、饲料化肥的原料及产品的成分分析等。此外,近红外光谱还在药检行业中用于对假药进行识别、在烟草工业中用于过程分析^[94]和质量控制^[95]等。

可以看到,随着近红外光谱技术在各个行业的

深入应用,近红外光谱数据逐渐呈现如下几个特征:首先,各个行业近红外光谱数据的数据量正在呈几何级数增长。以药检行业为例,为了市场上出售的药物进行监督检查,必须搜集全国所有不同类型、不同厂家、不同批次药物的近红外光谱来进行分析和建模,不同类型、厂家、批次的药物的近红外光谱汇集到一起,数据总量接近PB级别,超过以往任何类型的近红外光谱数据的数据量。同时,各个行业之中近红外光谱数据的类型也从单一类型向多类型数据进行转变。以工业生产为例,以往近红外光谱技术的应用仅仅是使用仪器采集一条生产线上产品的光谱数据,通过对采集到的光谱数据进行分析实现对该生产线产品的质量控制。在大规模的生产中,需要多台仪器对多条生产线甚至是不同厂家的数据进行采集。随着近红外光谱应用的深入,光谱的采集范围扩展到各种不同的原料以及不同生产节点的中间产品。不同生产线、不同仪器、不同类型原料和不同生产节点中间产品光谱的产生,大大丰富了近红外光谱的数据类型。此外,近红外光谱作为一种快速分析方法,采集光谱的速度非常快。正常情况下,单条光谱的采集时间仅为几秒钟。与之对应的,近红外光谱相关分析技术可以在极短的时间内实现对采集到的光谱进行处理并将结果输出。最后,在许多以在线分析为主流分析模式的行业中,较短的光谱采集时间使单位时间内产生的光谱数量异常巨大,这使得单条光谱的价值密度非常低,必须使用大数据挖掘技术对光谱数据集进行整体分析,才能挖掘到其中的有效信息。不难看出,传统的近红外光谱数据正逐渐转变为近红外光谱大数据。在此背景下,又出现了许多针对近红外光谱大数据的分析与应用。

在工业上,将近红外光谱与多元统计过程控制(multivariate statistical process control, MSPC)方法相结合,通过实时采集生产过程中物料的近红外光谱,并利用主成分分析建立模型计算统计量Hotelling T^2 对生产过程进行实时评价,建立了一种适用于生产过程分析与在线实时监测的新方法。图3为将该方法用于对生产线上3个月的样品进行监控的结果。可以明显看出第1个月的批次(No. 1)数据点大多在下方实线控制限以内,且批次内波动幅度较小,但以后2个月的2个批次(No. 2和No. 3)超过了下方实线控制限,且批次内波动幅度较大。由此说明原料或生产工艺随时间发生了变化,导致了物料的质量波动。实际应

用表明,利用该方法可以根据统计量实时监测生产过程中物料的变动情况,并根据不同批次统计量的进一步统计实现批次间相对稳定性的分析,成为生产过程质量控制的良好方法.从上面的应用中可以看出,该方法能够成功地对在线样品进行监测.但也应当看到,上述应用中该方法仅对一条生产线的样品进行检测,如果存在多条生产线的样品需要同时进行检测,就必须分别对每条生产线的样品进行建模,整个过程变得繁琐无比,建好的模型之间也没有通用性,此时方法的有效性将会大打折扣.因此,在面对数据类型不统一的大数据时,如何建立稳定且通用的检测模型并进行在线监控,将是今后工作的重点内容.

对于批次产品的生产过程而言,感兴趣的信息常常隐藏在未知的扰动中,而这些扰动在不同批次的生产中往往不同.通过分析和提取不同批次产品中的扰动信息,对于产品的在线质量分析与过程监控、提高产品的可靠性和改进生产有着重要的意义.交替三线性分解技术(alternating trilinear decomposition, ATLD)具有在复杂干扰下捕捉感兴趣信息的优势,将近红外光谱技术与ATLD方法相结合,可以建立一种过程分析的新方法,并将其用于不同批次产品的生产过程分析.实践证明,该方法可以在复杂干扰存在的情况下,捕捉批次共有的信息,并且利用该信息分析批次样品在生产过程中的变化.图4为使用传统的基于主成分分析(PCA)方法的过程分析与ATLD方法对生产线产品进行监控的结果.其中第11~14批次的样品(黑色叉形样品)在建模批次上不存在差别,而在样品批次上存在差别.基于主成分分析

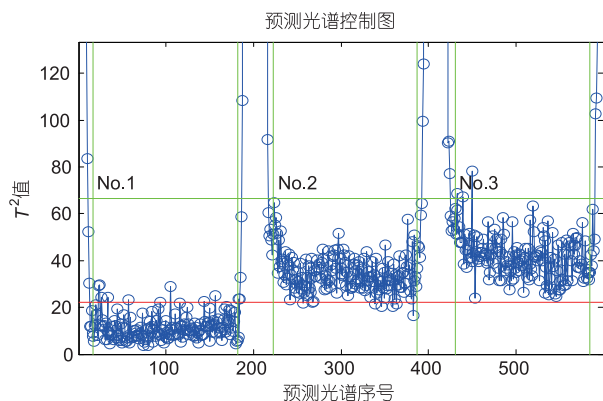


图3 (网络版彩色)生产过程质量控制图

Figure 3 (Color online) Quality control chart for a production process

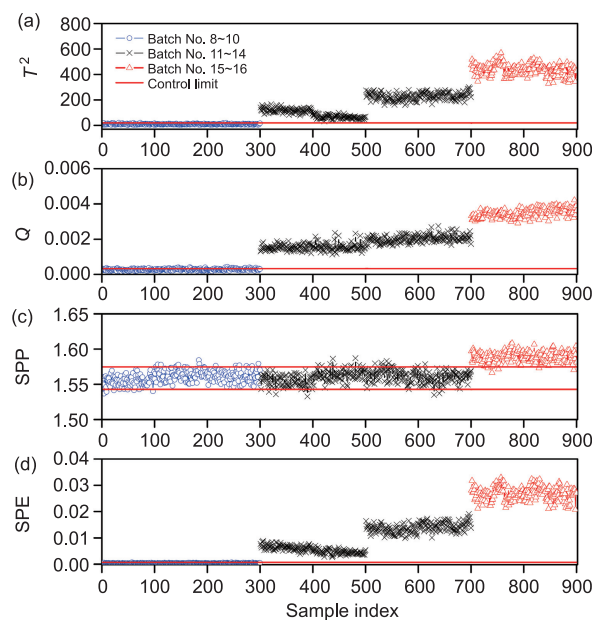


图4 (网络版彩色)建模样品(1~300)和验证样品(301~900)的过程监控图

Figure 4 (Color online) Control charts of the calibration (1~300) validation (301~900) samples

的算法对建模批次的判别出现了错判(图4(a)),但可以准确地将样品批次的差别进行判别(图4(b)).而ATLD方法对建模批次和样品均可以实现准确的判别(图4(c)和(d)).可以很清楚地看到,与传统基于主成分分析方法的过程分析相比较,ATLD方法在过程分析中表现良好^[94].此外,面对不同批次的产品,还可以将改进的系统聚类图与在线近红外光谱结合进行模式识别研究,实现对不同批次工业样品的辨别^[95].ATLD方法和改进的系统聚类图均可以有效地对不同批次的样品进行分析.然而,当分析数据的量很大时,整个分析过程将非常耗时.对于海量数据(TB甚至PB级别的数据),上述方法无法实现有效分析.因此,必须进一步开发和设计新的分析方法对近红外光谱大数据进行分析.

前面提到的近红外光谱技术在不同领域的各种应用,尽管分析对象已经从传统近红外光谱数据转变为近红外光谱大数据,分析方法仍然是传统的数据分析方法.这一点严重制约了近红外光谱技术的进一步发展,因为传统分析方法不能充分挖掘大数据中的有效信息.随着应用的深入,同样也出现了一些以大数据分析为基础的应用.如在某些领域中,一种基于近红外光谱技术的物联网系统正在逐

步形成. 该系统以近红外分析平台为基础, 通过网络系统搜集不同产地原料的光谱以及不同工厂的产品的的光谱. 然后使用数据挖掘等分析技术对搜集到的光谱进行分析, 得到不同产地原料的特点以及不同工厂对生产原料的不同要求. 以此分析结果为基础, 对原料的加工、存储点进行重新布置与调配, 在减少大众化原料运输时间的同时, 也保证了某些工厂对其所必需的特定原料的需求, 从而大大降低了运输和存储原料的成本. 该系统的建成对于工业化生产的统一管理与成本控制有十分重要的意义.

5 近红外光谱大数据的发展及应用前景

随着近红外光谱大数据的进一步发展, 近红外光谱技术的各个领域, 尤其是数据采集、存储与管理、数据分析以及数据可视化方面将会经历巨大的革新. 在数据采集方面, 以数据采集、处理和分析3项基本功能为基础设计的传统近红外光谱仪器将会逐步被小型化、智能化的新型近红外光谱仪器所取代. 新的近红外光谱仪器仅需包含数据采集与传输功能, 数据在经仪器采集之后可直接传输至数据终端进行处理与分析. 而在数据存储与管理方面, 传统的以实验室信息管理系统(laboratory information management system, LIMS)为基础的近红外光谱数据存储与管理系统将会升级和发展成为近红外光谱大数据管理系统. 该系统使用以分布式文件系统、分布式数据库为代表的大数据技术存储和管理不同类型和来源的近红外光谱大数据. 在数据分析方面, 面对各行各业数据体量巨大、种类繁多的近红外光谱大数据时, 传统的数据预处理、奇异样本筛选、多元校正和模型转移技术的分析能力将大打折扣. 因此, 必须发展以大数据算法为基础的新型近红外光谱分析技术. 新的算法能够在短时间内对TB甚至PB级别的数据进行分析后输出结果, 满足了大数据快速、实时分析的需求. 此外, 传统的近红外光谱分析由于数据类型简单、数据量小, 因此可以文本或图片的形式直接输出分析结果, 不存在可视化的问题. 而在输出近红外光谱大数据的复杂的分析结果时, 必须开发和设计新的大数据可视化方法, 对分析结果进行解释和输出. 上面提到的各种技术集成到一起, 组成了一个近红

外光谱大数据云计算平台. 该平台以网络化管理系统为基础, 通过新型仪器搜集和传输光谱数据, 并使用大数据管理系统存储和管理不同类型和来源的近红外光谱大数据. 同时, 云计算平台还可以使用大数据分析技术对光谱大数据进行分析, 挖掘数据中的有效信息, 并将结果以可视化的方式进行表达, 为生产生活提供相关的信息, 最终实现工业生产的信息化管理.

云计算平台在药检、食品、药品、农业等领域中有广阔的应用前景. 通过云计算平台可以管理和存储原料生产、在线产品和实验室研究等不同来源的近红外光谱数据. 同时, 使用大数据分析技术对搜集到的光谱大数据进行分析与挖掘, 然后将分析结果以可视化的方式进行输出, 可以有效地为产品的质量提供标准, 为原料的管理与存储、产品的销售以及上级有关部门的监控与执法提供可靠依据. 近红外光谱大数据云计算平台的建成对于药检、食品、药品、农业等领域实现的数字化、信息化的管理有重要的意义. 但也需要清醒地认识到, 该平台的实施与建设十分复杂而繁琐, 硬件和软件方面均存在大量的问题. 从硬件角度来说, 需要设计和选择不同样式的近红外光谱仪器以满足不同类型样品光谱采集. 此外, 为了对采集到的数据进行分析和管理, 需要组建大型服务器机群. 从软件角度来说, 必须开发出相关的大数据分析方法对采集到的近红外光谱大数据进行挖掘, 以获得有价值的信息; 还要建设相应的数据管理系统来存储和管理采集到的光谱与分析结果. 建设云计算平台需要做扎实的基础研究和系统的技术开发等工作, 属于典型的系统工程, 涉及诸多专业, 需要多方面的配合, 要实现满意的应用效果, 需要长期的努力.

随着生产力水平的不断提高, 近红外光谱技术将成为一种必须的分析手段. 而将近红外光谱技术与大数据进行融合, 建立一套基于大数据的近红外光谱分析管理技术, 对于近红外光谱在各行各业之中的深入发展, 有着十分重要的意义. 相信在不远的未来, 融合了大数据技术的近红外光谱分析技术一定能够在我国信息化与工业化深度融合的发展进程中发挥应有的作用.

致谢 衷心感谢航天材料及工艺研究所胡子君研究员等人给予的帮助.

参考文献

- 1 Graham-Rowe D, Goldston D, Doctorow C, et al. Big data: Science in the petabyte era. *Nature*, 2008, 455: 1–50
- 2 Reichman O J, Matthew B, Mark P H. Challenges and opportunities of open data in ecology. *Science*, 2011, 311: 703–705
- 3 Meng X F, Ci X. Big data management: Concepts, techniques and challenges (in Chinese). *J Comput Res Dev*, 2013, 50: 146–169 [孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. *计算机研究与发展*, 2013, 50: 146–169]
- 4 Wu X D, Zhu X Q, Wu G Q, et al. Data mining with big data. *IEEE T Knowl Data Eng*, 2014, 26: 97–107
- 5 Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun ACM*, 2008, 51: 117–122
- 6 Bloom B. Space/time tradeoffs in hash coding with allowable errors. *Commun ACM*, 1970, 13: 422–426
- 7 Buddhikot M M, Suri S, Waldvogel M. Space decomposition techniques for fast layer 4 switching. *Protoc High Speed Netw*, 1999, 25–41
- 8 Geraci F, Pellegrini M, Pisati P, et al. Packet classification via improved space decomposition technique. In: Makki K, Knightly E, eds. *Proceedings IEEE Infocom 2005*. Piscataway, NJ: IEEE, 2005, 1: 304–312
- 9 Genevieve B M, George H. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*, 2005, 12: 448–457
- 10 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. *Comput Network ISDN Syst*, 1998, 30: 101–117
- 11 Konopnicki D, Shmueli O. W3QS: A query system for the worldwide Web. In: Dayal U, Gray P M D, Nishio S, eds. *VLDB'95. Proceedings of the 21st International Conference on Very Large Data Bases*. San Francisco, CA: Morgan Kaufmann, 1995. 54–65
- 12 Ninagawa A, Eguchi K. Link prediction using probabilistic group models of network structure. In: Shin S Y, Ossowski S, Schumacher M, eds. *Proceedings of the 2010 ACM Symposium on Applied Computing*. New York: ACM, 2010. 1115–1116
- 13 Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorization. *ACM Trans Knowl Discov Data*, 2011, 10, doi: 10.1145/1921632.1921636
- 14 Ghemawat S, Gobioff H, Leung S T. The google file system. *Operat Syst Rev*, 2003, 37: 29–43
- 15 Chang F, Dean J, Ghemawat S, et al. Big table: A distributed storage system for structured data. *ACM Trans Comput Syst*, 2008, 26: 4–15
- 16 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. *Commun ACM*, 2008, 51: 107–113
- 17 Li C H, Zhang X F, Jin H, et al. MapReduce: A new programming model for distributed parallel computing (in Chinese). *Comput Eng Sci*, 2011, 31: 129–135 [李成华, 张新访, 金海, 等. MapReduce: 新型的分布式并行计算编程模型. *计算机工程与科学*, 2011, 31: 129–135]
- 18 Zhang X Y, Liu C, Nepal S, et al. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. *J Comput Syst Sci*, 2014, 80: 1008–1020
- 19 Shafer J, Rixner S, Cox A L. The Hadoop distributed filesystem: Balancing portability and performance. In: IEEE, ed. *2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS 2010)*. Piscataway, NJ: IEEE, 2010. 122–132
- 20 Yang C, Zhang X Y, Zhong C M, et al. A spatiotemporal compression based approach for efficient big data processing on Cloud. *J Comput Syst Sci*, 2014, 80: 1563–1583
- 21 Zhao J Q, Wang L Z, Tao J, et al. A security framework in G-Hadoop for big data computing across distributed Cloud data centres. *J Comput Syst Sci*, 2014, 80: 994–1007
- 22 Kaser O, Lemire D. Tag-cloud drawing: Algorithm for cloud visualization. *Comput Res Repository*, 2007, 70: 109–118
- 23 Rysavy S J, Bromley D, Daggett V. DIVE: A graph-based visual-analytics framework for big data. *IEEE Comput Graph Appl*, 2014, 32: 26–37
- 24 Liu R, Wang X D, Aihara K, et al. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network. *Med Res Rev*, 2014, 34: 455–478
- 25 Hsieh J C, Li A H, Yang C C. Mobile, cloud, and big data computing: Contributions, challenges, and new directions in telecardiology. *Int J Environ Res Public Health*, 2013, 10: 6131–6153
- 26 Massie A B, Kuricka L M, Segev D L. Big data in organ transplantation: Registries and administrative claims. *Am J Transplant*, 2014, 14: 1723–1730.
- 27 Gonzalez-Diaz H, Herrera-Ibata D M, Duarado-Sanchez A, et al. ANN multiscale model of anti-HIV Drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J Chem Inf Model*, 2014, 54: 744–755
- 28 O'Driscoll A, Daugeleite J, Sleator R D. “Big data”, hadoop and cloud computing in genomics. *J Biomed Inform*, 2013, 46: 774–781
- 29 Costa F F. Big data in biomedicine. *Drug Discov Today*, 2014, 19: 433–440
- 30 Dai L, Gao X, Guo Y, et al. Bioinformatics clouds for big data manipulation. *Biology Direct*, 2012, 7: 43–49
- 31 Spiess J, T'Joens Y, Dragnea R, et al. Using big data to improve customer experience and business performance. *Bell Labs Tech J*, 2014, 18: 3–17

- 32 Tang H, Yang X, Zhang Y J. Effort at constructing big data sensor networks for monitoring greenhouse gas emission. *Intl J Distrib Sens Netw*, 2014, 14: 1–7
- 33 Kim G H, Trimi S, Chung J H. Big-data applications in the government sector. *Commun ACM*, 2014, 57: 78–85
- 34 Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl Spectrosc*, 1985, 39: 491–500
- 35 Wold S, Antti H, Lindgren F, et al. Orthogonal signal correction of near-infrared spectra. *Chemom Intell Lab Syst*, 1998, 44: 175–185
- 36 Sjöblom J, Svensson O, Josefson M, et al. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom Intell Lab Syst*, 1998, 44: 229–244
- 37 Andersson C A. Direct orthogonalization. *Chemom Intell Lab Syst*, 1999, 47: 51–63
- 38 Fearn T. On orthogonal signal correction. *Chemom Intell Lab Syst*, 2000, 50: 47–52
- 39 Westerhuis J A, de Jong S, Smilde A K. Direct orthogonal signal correction. *Chemom Intell Lab Syst*, 2001, 56: 13–25
- 40 Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal Chem*, 1986, 58: 1167–1172
- 41 Ma C X, Shao X G. Continuous wavelet transform applied to removing the fluctuating background in near-infrared spectra. *J Chem Inf Comput Sci*, 2004, 44: 907–911
- 42 Chen D, Wang F, Shao X G, et al. Elimination of interference information by a new hybrid algorithm for quantitative calibration of near infrared spectra. *Analyst*, 2003, 128: 1200–1203
- 43 Kalivas J H, Roberts N, Sutter J M. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry. *Anal Chem*, 1989, 61: 2024–2030
- 44 Lucasius C B, Kateman G. Genetic algorithms for large-scale optimization problems in chemometrics—An application. *TrAC-Trends Anal Chem*, 1991, 10: 254–261
- 45 Arakawa M, Yamashita Y, Funatsu K. Genetic algorithm-based wavelength selection method for spectral calibration. *J Chemom*, 2011, 25: 10–19
- 46 Gu X Y, Xu K X, Wang Y. Application of wavelength selection algorithm to measure the effective component of Chinese medicine based on near-infrared spectroscopy (in Chinese). *Spectrosc Spect Anal*, 2006, 9: 1618–1620 [谷筱玉, 徐可欣, 汪曦. 波长选择算法在近红外光谱法中药有效成分测量中的应用. *光谱学与光谱分析*, 2006, 9: 1618–1620]
- 47 Xu L, Jiang J H, Wu H L, et al. Variable-weighted PLS. *Chemom Intell Lab Syst*, 2007, 85: 140–143
- 48 Zou H Y, Wu H L, Fu H Y, et al. Variable-weighted least-squares support vector machine for multivariate spectral analysis. *Talanta*, 2010, 80: 1698–1701
- 49 Araujo M C U, Saldanha T C B, Galvao R K H, et al. The successive projections algorithm for variable selection in spectroscopic multi-component analysis. *Chemom Intell Lab Syst*, 2001, 57: 65–73
- 50 Galvao R K H, Araujo M C U, Fragoso W D, et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemom Intell Lab Syst*, 2008, 92: 83–91
- 51 Acebal C C, Grunhut M, Lista A G, et al. Successive projections algorithm applied to spectral data for the simultaneous determination of flavour enhancers. *Talanta*, 2010, 82: 222–226
- 52 Centner V, Massart D L, de Noord O E, et al. Elimination of uninformative variables for multivariate calibration. *Anal Chem*, 1996, 68: 3851–3858
- 53 Li H D, Liang Y Z, Xu Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal Chim Acta*, 2009, 648: 77–84
- 54 Fan W, Li H D, Shan Y, et al. Classification of vinegar samples based on near infrared spectroscopy combined with wavelength selection. *Anal Methods*, 2011, 3: 1872–1876
- 55 Cai W S, Li Y K, Shao X G. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom Intell Lab Syst*, 2008, 90: 188–194
- 56 Xu H, Liu Z C, Cai W S, et al. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom Intell Lab Syst*, 2009, 97: 189–193
- 57 Shao X G, Zhang M, Cai W S. Multivariate calibration of near-infrared spectra by using influential variables. *Anal Method*, 2012, 4: 467–473
- 58 Shao X G, Du G R, Jing M, et al. Application of latent projective graph in variable selection for near infrared spectral analysis. *Chemom Intell Lab Syst*, 2012, 114: 44–49
- 59 Shan R F, Cai W S, Shao X G. Variable selection based on locally linear embedding mapping for near-infrared spectral analysis. *Chemom Intell Lab Syst*, 2014, 131: 31–36
- 60 Walczak B. Outlier detection in multivariate calibration. *Chemom Intell Lab Syst*, 1995, 28: 259–272

- 61 Mark H. Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis. *Anal Chem*, 1987, 59: 790–795
- 62 Paul S R, Fung K Y. A generalized extreme studentized residual multiple outlier detection procedure in linear regression. *Technometrics*, 1991, 33: 339–348
- 63 de Groot P J, Postma G J, Melssen W J, et al. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. *Anal Chim Acta*, 2001, 446: 71–83
- 64 Gnanadesikan R, Kettenring J R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 1972, 28: 81–124
- 65 Walczak B, Massart D L. Robust principal components regression as a detection for outliers. *Chemom Intell Lab Syst*, 1995, 27: 41–54
- 66 Howington E B, Gray J. Matrix-based genetic algorithm for computing the minimum volume ellipsoid. *IEEE Trans Evol Comput*, 2009, 13: 1247–1260
- 67 Rousseeuw P J, Katrien V D. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 1999, 41: 212–223
- 68 Egan W J, Morgan S L. Outlier detection in multivariate analytical chemical data. *Anal Chem*, 1998, 70: 2372–2379
- 69 Yohai V J. High breakdown point and high efficiency robust estimates for regression. *Ann Stat*, 1987, 15: 642–656
- 70 Serneels S, Croux C, Filzmoser P, et al. Partial robust M-regression. *Chemom Intell Lab Syst*, 2005, 79: 55–64
- 71 Liu Z C, Cai W S, Shao X G. Outlier detection in near-infrared spectroscopic analysis by using Monte Carlo cross-validation. *Sci China Ser B-Chem*, 2008, 51: 751–759
- 72 Norgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc*, 2000, 54: 413–419
- 73 Shao X G, Bian X H, Cai W S, et al. An improved boosting partial least squares method for near-infrared spectroscopic quantitative analysis. *Anal Chim Acta*, 2010, 666: 32–37
- 74 Jing M, Cai W S, Shao X G. Multiblock partial least squares regression based on wavelet transform for quantitative analysis of near infrared spectra. *Chemom Intell Lab Syst*, 2010, 100: 22–27
- 75 Liu Z C, Cai W S, Shao X G. A weighted multiscale regression for multivariate calibration of near infrared spectra. *Analyst*, 2009, 134: 261–266
- 76 Xu H, Cai W S, Shao X G. Weighted partial least squares regression by variable grouping strategy for multivariate calibration of near infrared spectra. *Anal Methods*, 2010, 2: 289–294
- 77 Shao X G, Liu Z C, Cai W S. Extraction of chemical information from complex analytical signals by a non-negative independent component analysis. *Analyst*, 2009, 134: 2095–2099
- 78 Bouveresse E, Hartmann C, Massart D L. Standardization of near-infrared spectrometric instruments. *Anal Chem*, 1996, 68: 982–990
- 79 Wang Y D, Veltkamp D J, Kowalski B R. Multivariate instrument standardization. *Anal Chem*, 1991, 63: 2750–2756
- 80 Wang Y D, Lysaght M J, Kowalski B R. Improvement of multivariate calibration through instrument standardization. *Anal Chem*, 1992, 64: 562–564
- 81 Wang Z Y, Dean T, Kowalski B R. Additive background correction in multivariate instrument standardization. *Anal Chem*, 1995, 67: 2379–2385
- 82 Shenk J S, Westerhaus M O, Templeton W C. Calibration transfer between near infrared reflectance spectrophotometers. *Crop Sci*, 1985, 25: 159–161
- 83 Shenk J S, Westerhaus M O. New standardization and calibration procedures for NIRS analytical systems. *Crop Sci*, 1991, 31: 1694–1696
- 84 Shenk J S, Westerhaus M O. Near infrared reflectance analysis with single and multiproduct calibrations. *NIR News*, 1993, 4: 13–15
- 85 Norgaard L. Direct standardisation in multi wavelength fluorescence spectroscopy. *Chemom Intell Lab Syst*, 1995, 29: 283–293
- 86 Chen Z P, Li L M, Yu R Q, et al. Systematic prediction error correction: A novel strategy for maintaining the predictive abilities of multivariate calibration models. *Analyst*, 2011, 136: 98–106
- 87 Du W, Chen Z P, Zhong L J, et al. Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Anal Chim Acta*, 2011, 690: 64–70
- 88 Forina M, Drava G, Armanino C, et al. Transfer of calibration function in near-infrared spectroscopy. *Chemom Intell Lab Syst*, 1995, 27: 189–203
- 89 Kunz M R, Kalivas J H, Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of Tikhonov regularization. *Anal Chem*, 2010, 82: 3642–3649
- 90 Kunz M R, Ottaway J, Kalivas J H, et al. Impact of standardization sample design on Tikhonov regularization variants for spectroscopic calibration maintenance and transfer. *J Chemom*, 2010, 24: 218–229
- 91 Shahbazikhah P, Kalivas J H. A consensus modeling approach to update a spectroscopic calibration. *Chemom Intell Lab Syst*, 2013, 120: 142–153

- 92 Farrell J A, Higgins K, Kalivas J H. Updating a near-infrared multivariate calibration model formed with lab-prepared pharmaceutical tablet types to new tablet types in full production. *J Pharm Biomed Anal*, 2012, 61: 114–121
- 93 Liu Y, Cai W S, Shao X G. Standardization of near infrared spectra measured on multiinstrument. *Anal Chim Acta*, 2014, 836: 18–23
- 94 Liu J J, Xu H, Cai W S, et al. Discrimination of industrial products industrial products by on-line near infrared spectroscopy with an improved dendrogram. *Chin Chem Lett*, 2011, 22: 1241–1244
- 95 Liu J J, Ma X, Wen Y D, et al. On-line near infrared spectrscopy combined with alternating trilinear decomposition for process analysis of industrial production and quality assurance. *Ind Eng Chem Res*, 2011, 50: 7677–7681

Analysis of near-infrared spectra: From tradition to big data

LIU Yan, CAI WenSheng & SHAO XueGuang

Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, China

Near-infrared spectroscopy has been widely used in analytical sciences because of its speed, accuracy, environmental friendliness and simplicity. The data type associated with near-infrared spectra is becoming big data because of the development of the process in many fields. The methods used for processing of near-infrared spectra, including spectral preprocessing, outlier detection, multivariate calibration, modeling transfer, and their applications are summarized in this paper. The research approaches for big data analysis of near-infrared spectra, including online detection of industrial products, the identification of products from different batches, and the web of things constructed using near-infrared spectra, are also introduced. Additionally, the prospects for further development of the big data of near-infrared spectra and the cloud platform for near-infrared spectra are discussed.

big data, data analysis, big data of near-infrared spectra, near-infrared spectral analysis, cloud platform

doi: 10.1360/N972014-01203