第39卷 第5期 质 谱 学 报 Vol. 39 No. 5
2018年9月 Journal of Chinese Mass Spectrometry Society Sep. 2018

数据预处理技术和机器学习方法在质子转移反应质谱中的应用

孙 运[1],陈一冰[2],褚美娟[1],蒋学慧[1],汪 瞲[1],郭冰清[1]

(1.天津大学精密仪器与光电子工程学院,天津 300072;
2.中国人民解放军总医院呼吸内科,北京 100853)

摘要:质子转移反应质谱(PTR-MS)法是一种用于检测挥发性有机物(VOCs)的分析技术。它具有检测限低、响应速度快、无需样品前处理、实时分析等特点,在大气化学、环境化学、食品、生物医学等领域得到广泛应用。随着PTR-MS应用的扩展和样品种类的增加,如何从复杂的质谱数据中提取特征,并寻找内在规律,对分析算法的处理能力提出了更高的要求。本工作从数据预处理技术和机器学习方法两方面展开论述,归纳了具有PTR-MS特点的数据预处理技术,总结了不同机器学习算法在PTR-MS数据分析中的应用,并讨论了它们的优点和不足。

关键词:质子转移反应质谱(PTR-MS);挥发性有机物(VOCs);数据预处理;机器学习
中图分类号:O657.63 **文献标志码**:A **文章编号**:1004-2997(2018)05-0513-11
doi:10.7538/zpxb.2017.0181

Review of Data Pre-processing Techniques and Machine Learning in PTR-MS

SUN Yun[1], CHEN Yi-bing[2], CHU Mei-juan[1], JIANG Xue-hui[1], WANG Yan[1], GUO Bing-qing[1]

(1. *School of Precision Instrument and Optoelectronics Engineering, Tianjin University, Tianjin* 300072, *China*;
2. *Respiratory Medicine, Chinese PLA General Hospital, Beijing* 100853, *China*)

Abstract: Proton transfer reaction mass spectrometry (PTR-MS) is an analytical technique developed for the detection of volatile organic compounds (VOCs). It offers many advantages for VOCs analysis, such us ultra-low detection limits, very short response, no sample preparation, real-time analysis, etc. It has been applied in atmospheric chemistry environmental chemistry, food and biomedical. With the expansion of applications of PTR-MS and the increase of sample types, how to analyze the features from complex

收稿日期:2017-11-23;**修回日期**:2018-02-27
基金项目:国家重大科学仪器设备开发专项:质子转移反应质谱仪器研制及应用示范(2013YQ090875);天津市应用基础与前沿技术研究计划:用于环境监测的离子漏斗-质子转移反应离子源质谱研究(15JCYBJC23300)资助
作者简介:孙 运(1988—),男(汉族),河北唐山人,博士研究生,从事质谱仪器研制及应用。E-mail: yunsun@tju.edu.cn
通信作者:蒋学慧(1982—),女(汉族),河北正定人,工程师,从事质谱数据处理与信息挖掘。E-mail: jiangxuehui82@163.com
网络出版时间:2018-05-14;**网络出版地址**:http://kns.cnki.net/kcms/detail/11.2979.TH.20180511.1647.010.html

data and find out the inherent rules have put forward higher requirements on the processing ability of the algorithm. Therefore, this paper discussed the data preprocessing techniques and machine learning methods. Firstly, we summarized the data preprocessing methods with PTR-MS features. The data generated by the instrument cannot be directly used for statistical analysis, otherwise it will bring great error. Therefore, data pre-processing is an essential step. It includes several steps, such as denoising, normalization, and concentration calculation. The purpose of preprocessing is to get data matrix for subsequent analysis. Next, we focused on the use of machine learning methods for data analysis in PTR-MS, and the advantages of this techniques would be demonstrated as well as the drawbacks. The machine learning method can be divided into two parts. Usually unsupervised methods are common choices for initial data analysis. For further analysis and a priori knowledge, a supervised analysis would be a better way. These methods use this knowledge to learn rules and patterns related to classes in the data, and then use these rules and patterns to predict classes in newly acquired data sets. The main goal of all surveillance techniques is to find the relationship between the predictor (VOC) matrix and the response vector. In general, the combination of the unsupervised and supervised methods is a good idea. PTR-MS is a soft ionization technique, however, the presence of a few fragments will still cause great difficulties in spectral analysis, especially for unknown mixtures, which is the main reason why spectral analysis of PTR-MS differs from other mass spectrometry methods. Perhaps, the data fusion of different platform instruments and different samples will be a good way to solve this problem.

Key words: proton transfer reaction mass spectrometry (PTR-MS); volatile organic compounds (VOCs); data pre-processing; machine learning

质子转移反应质谱(PTR-MS)由奥地利因斯布鲁克大学的Lindinger等[1]研发于20世纪90年代。经过20多年的发展,PTR-MS的应用从早期的环境污染物分析,扩展到食品科学、生物制药、医学诊断、防化安保等领域,主要用于挥发性有机物(VOCs)检测[2]。VOCs的种类多达上万种,比如,在食品领域,水果的香气物质有2 000多种;在医学领域,人体中呼气、血液、尿液等样品释放的VOCs均在几百种以上[3]。如何更好地在复杂未知混合物中进行VOCs成分的检测,并通过数据分析深度挖掘VOCs的特征是非常重要的。

不同于电子轰击源,PTR-MS采用的是一种软电离方式,通过质子转移反应将VOCs离子化。这是一个需要精密控制的过程,温度、湿度、压强、电场的变动都会对仪器的信号输出产生影响。如果直接用仪器产生的数据来进行统计分析,会带来极大的误差,所以,数据的预处理必不可少[4-5]。随着化学计量学和软件技术的发展,以及蛋白组学、代谢组学等领域分析需求的促进,机器学习算法得到了极大的应用和推广,能够很好的帮助人们进行特征提取,解读数据中有价值的信息[6-8]。目前,已经有很多机器学习方法应用到PTR-MS的数据分析中[9]。

1 质子转移反应质谱仪

PTR-MS主要包括进样系统、空心阴极放电源、漂移管、质量分析器、检测系统5大部分[10-11]。其中,空心阴极放电源主要用于产生水合氢离子(H_3O^+)。随后,H_3O^+进入漂移管,与VOCs在电场和气流作用下不断碰撞。在碰撞过程中,质子亲和势比水高的VOCs与H_3O^+发生质子转移反应,这些携带质子的VOCs以离子形式进入后续的质谱系统。在PTR-MS中,四极杆、离子阱、飞行时间等质量

分析器各具特点[12-14]。其中,四极杆和飞行时间质量分析器的应用最多。四极杆质量分析器具有良好的定量检测能力,最早被用作 PTR-MS 的质量分析器[12]。飞行时间质量分析器根据不同质量离子到达检测器时间的差异实现对离子的鉴别,具有质量范围宽、分辨率高等优点,并且在高分子质量和混合物分析方面显示了很大优势[14]。

2　质子转移反应质谱数据预处理

从 PTR-MS 仪器中得到的检测信号是原始数据,需要进行科学合理的预处理才可用于统计分析。数据预处理通常包括降噪、基线校准、峰形校准等[15-17]。PTR-MS 数据预处理流程示于图 1。

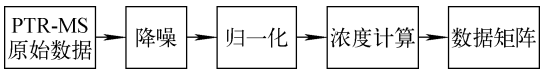


图 1　PTR-MS 数据预处理流程

Fig. 1　Process of the data preprocessing in PTR-MS

2.1　降噪

PTR-MS 仪器噪声主要来源于电子器件,此外,也与传输线及屏蔽状况、仪器环境温度等因素有关。仪器噪声往往表现为随机噪声,对于这种随机误差的处理,采用多次测量求取平均值是比较常规的做法,但是该法并不适用所有场合。

Cappellin 等[4]强调了质量轴的稳定性是均值法的前提。相比于四极杆质量分析器,飞行时间质谱的谱图质量轴不稳定性更加突出。飞行时间质谱检测中,一次脉冲产生的离子信号非常小,需要叠加多次脉冲产生的离子信号。每次分析都有一定的扫描速度,可能有 400 张谱图(即 400 次扫描),甚至是更多的谱图被叠加。对于一个质量数,需要每一张谱图的轴点位置都得到校正,均值法才会可靠。虽然在分析样品之前都会利用标准品对质量轴进行校准,但受限于标准品的个数,只有少数质量数能够得到校准。因此,该课题组采取了较为简单的操作,不改变信号值,仅根据校准的质量数偏移谱图。采用该方法,所得的质量误差小于 1×10^{-6}。

Hewitt 等[18]提到均值法在 VOCs 浓度较高并且波动较大时不适用。他们对噪声进行了定量分析,从而优化驻留时间及补偿仪器输出信号。利用 m/z 63、69、70 这 3 个质量数的信号对仪器噪声进行定量计算,调节驻留时间范围为 0.1～20 s,每个质量数至少采样 170 个数据点,并定义了噪声统计值(noise statistic, *NS*):

$$NS = \frac{\text{mean signal (cps)}}{\sqrt{\text{mean (cps)} \times \text{dwell (s)}}} \quad (1)$$

其中,cps(counts per second)表示离子信号计数值。结果发现,仪器噪声近似为高斯分布。对于设定的阈值(平均值$\pm2\times NS$),检测值中有 2%在上偏差范围内,2%在下偏差范围内,并且这一规律独立于质量数、驻留时间、计数率。基于此,他们提出了"2%"法则,认为在阈值(平均值$\pm2\times NS$)以外的数值变化很有可能不是噪声,而是由真实浓度变化引起的。

此外,还可以应用很多标准的降噪方法[19-20],特别是小波变换[21]。小波变换具有多分辨率性、尺度内相关性和时域局部化等特点,适用于不稳定信号的平滑降噪,同时不引起信号失真,使信号的原始特征得到最大程度的保留。

2.2　水团簇因素的补偿与归一化

在 PTR-MS 漂移管内,会有水团簇离子生成,这些团簇离子使得谱图复杂化。虽然可以通过增强漂移管内的 E/N 值(E 为电场强度,N 为气体数密度)抑制团簇的生成,但不能完全消除。考虑到团簇效应,Gouw 等[22]引入了一个新的参数 X_R 进行归一化。通过实验测定,绝大多数 VOCs 的 X_R 值为 0.5,但苯和甲苯的 X_R 值接近 0,因为它们与 $H_3O^+(H_2O)$的反应较慢。如果 X_R 为负值,说明有一部分 $H_3O^+(H_2O)$的碎片峰在漂移管和质量分析器的接口处生成。据此对计数值进行归一化,记为 *NCPS*(normalized counts/s):

$$NCPS = 10^6 \frac{i[RH^+]}{i[H_3O^+] + X_R \times i[H_3O^+(H_2O)]} \quad (2)$$

Tani 等[23]定义了 *SCPS*,对检测的离子进行归一化。*SCPS* 是所检测的物质信号总离子数,按照 10^6 计数的母离子,对 2 kPa 漂移管压强进行归一,示于式(3):

$$SCPS = \frac{\sum i[RH^+] \times 10^6}{\sum_{n=1}^{3} i[H_3O^+(H_2O)_{n-1}]} \times \frac{2}{P_{drift}} \quad (3)$$

Sinha 等[24]采用 PTR-MS 检测吡咯，讨论了湿度的影响。以 cps 计量的检测信号按照漂移管压力 200 Pa，温度为 298.15 K 以及 1.0×10^6 母离子的条件进行归一化，并分为 4 种情况：1) 只有 m/z 19；2) 只有 m/z 19、37；3) 只有 m/z 19、37、55；4) 只有 m/z 19、37、55、73。

$$NCPS = 10^6 \frac{i[RH^+]}{\sum_{n=1}^{x} i[H_3O^+(H_2O)_{n-1}]} \times \frac{2}{P_{drift}} \times \frac{T_{drift}}{298.15} \quad (4)$$

这里 $x=1,2,3,4$，分别对应上述 4 种情况。

2.3 计数值转化浓度的计算方法

通常，VOCs 的检测都需要有对应的标准物质，利用标准曲线进行定量计算。尽管这种方法的精度较高，但不是所有的物质都有标准品，并且在检测未知混合物时很难确定标准品的种类。PTR-MS 可以通过化学反应计算 VOCs 浓度。当 VOCs 分子的质子亲和势大于水时，即可发生质子转移反应。用 R 表示 VOCs 分子，上述反应可以用式(5)表示：

$$H_3O^+ + R \xrightarrow{k} RH^+ + H_2O \quad (5)$$

式中，k 表示质子转移反应过程的反应速率常数。参照文献[1,25]报道，在反应区末端的产物离子浓度$[RH^+]$可以表示为：

$$[RH^+] = [H_3O^+]_0[1 - e^{-k[R]t}] \approx [H_3O^+]_0[R]kt \quad (6)$$

式(6)中，$[H_3O^+]_0$ 为反应试剂 H_3O^+ 的初始浓度，$[R]$为待测物 R 的浓度，t 为离子通过漂移管的平均时间。因为待测物 R 的浓度远小于 H_3O^+ 的浓度，其只与少量的 H_3O^+ 发生质子转移反应，H_3O^+ 信号强度在反应前后可以近似不变。由此，可得待测物 R 的浓度计算公式为：

$$[R] = \frac{[RH^+]}{[H_3O^+]_0} \times \frac{1}{kt} = \frac{i[RH^+]T_{H_3O^+}}{i[H_3O^+]T_{RH^+}} \times \frac{1}{kt} \quad (7)$$

式(7)中，计数率 $i[RH^+]$和 $i[H_3O^+]$为待测物和 H_3O^+ 的离子计数值，可由检测系统测得。它们与 RH^+ 和 H_3O^+ 的浓度成比例，比例系数分别为 T_{RH^+} 和 $T_{H_3O^+}$，反应速率常数 k 可通过查阅文献得到，平均反应时间 t 可以直接测得，也可以通过计算公式求出[26]，这里不再叙述。

从式(7)可以看出，反应速率 k 对待测组分浓度的影响很大。Zhao 等[27]利用量子化学方法得出反应物结构，然后通过平均极化理论(average-dipole-orientation, ADO)计算了 78 种烃类物质和 58 种非烃类 VOCs 与 H_3O^+ 的质子转移反应速率。新计算的反应速率常数可以为 PTR-MS 定量检测 VOCs 提供可靠的参考值。

Keck 等[28]对浓度计算方程做了进一步修正。他们提出，RH^+ 的浓度在漂移管入口到出口间不断增加，并且增速大于 H_3O^+ 浓度减少的速度，这种流动效应会造成计算结果的偏差。为此，考虑 RH^+ 浓度随着时间变化的函数，经过一系列推导，对浓度计算式(7)进行了修正。

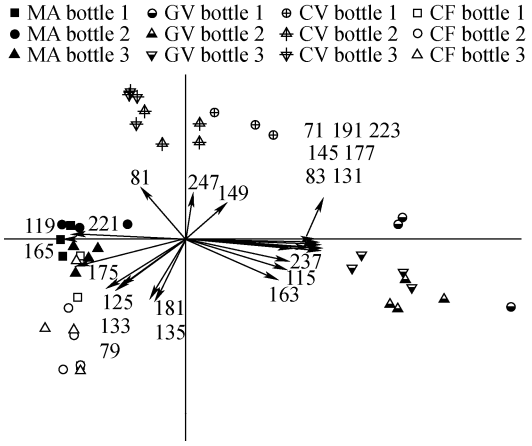
3 多变量统计分析

PTR-MS 能够获得多达 500 个不同 m/z 的谱图，但相比于色谱和光谱等仪器，谱图信息仍不够丰富。此外，PTR-MS 仪器的分析检测不含分离过程，谱图中的单个谱峰可能来自不同物质。所以，针对少量的具有代表性的物质成分进行快速检测，弥补因缺少分离过程造成的信息不足是 PTR-MS 谱图分析的一个策略。机器学习方法提供了多种渠道分析和理解复杂的数据，帮助获取有用信息[29-31]。通常最初的算法选择无监督模式，用于数据探索和挖掘，各个数据的权重是相同的。该算法能够在缺乏经验的前提下，研究 VOCs 与样品间的复杂关系。当有了先验知识，有监督模式算法将是更好的选择。

3.1 无监督学习方法

3.1.1 主成分分析 主成分分析(principal component analysis, PCA)是机器学习中应用非常广泛的无监督学习方法。PCA 方法是通过构造原随机变量的一系列线性不相关的组合反映原变量的信息[32]。其目标是用低维子空间表示高维数据，使得在误差平方和意义下能更好地描述原始数据。该法常被用于食品领域中产地、种类的区分[33-41]。

Boscaini 等[33]用 PCA 法处理 PTR-MS 检测数据的结果示于图 2。图中的每一个点对应一种样品,每条线对应一个质量数。线与线之间的角度代表二者的相关性,角度小说明正相关,反之说明相关性较差。从图中可以看出,PCA 方法很好地区分了 4 种不同品种的葡萄酒。



注:缩写字母 MA、GV、CV、CF 代表 4 种不同品种的葡萄酒

图 2　PCA 分析举例

Fig. 2　An example of PCA

Farneti 等[34]用 PCA 法找出了 21 种质量数用于描述番茄挥发物的模型,并区分了番茄的不同成熟阶段。张丹丹等[35]将 PTR-TOF MS 采集到的 3 个不同产地的闽北水仙茶的挥发性指纹图谱进行数学统计分析,利用 PCA 法提取了 3 个主成分,累计贡献率达到 84.66%,表明 PTR-TOF MS 结合分类算法可以有效区分不同产地的闽北水仙茶。除此之外,PCA 法常作为其他多变量分析法的第一步处理[42-44]。

从传统主成分分析方法的计算过程可以看出,进行 PCA 计算的关键是算出变量的协方差矩阵或者相关矩阵,相关矩阵可以从协方差矩阵得到,所以可以把问题都归结为协方差矩阵的计算。这个过程对离群值非常敏感,所以导致接下来所计算的相关矩阵、特征值和特征向量也受其影响,产生不合理的结果。协方差矩阵对离群值敏感,主要因为其计算过程中要使用均值向量,而均值向量只是对多维数据的简单求平均值,这种计算方法很容易受到离群值的影响,使得协方差矩阵不是稳健的估计量。稳健主成分分析方法则可以有效解决这个问题,通过构造一个稳健的协方差矩阵,降低离群值对协方差矩阵的影响[45]。

3.1.2　层次聚类分析法　另一种常用的无监督统计方法是聚类分析。聚类分析是根据各个样品或指标的数量对事物进行分类,在分类过程中不必给出分类的标准,是一种探索性的分析。聚类分析中的层次聚类分析法(hierarchical cluster analysis, HCA)是最常用的,其基本思想是将 N 个样品看成 N 类,然后将性质相近的 2 类合并为 1 个新类,再从剩下的 $N-1$ 类找出最接近的 2 类合并成 $N-2$ 类,以此类推,直到所有样品合成一类。这个过程可以用一系列的嵌套聚类树完成[46-47]。

Sánchezlópez 等[48]利用 PTR-TOF MS 结合 HCA 研究了生产浓缩咖啡的热水萃取工艺,并利用 Ward 最小方差法和半平方欧氏距离法对 46 个初步确定的化合物规范化时间强度特征进行了层次聚类分析。Ciesa 等[49]用 PTR-MS 法检测 7 种现代和 35 种老式苹果品种,分析单个水果在收获和存储期间释放出的 VOCs 信息,用 HCA 法评估了各品种释放的 VOCs 差异。Pozo-Bayón 等[50]用 HCA 法对 PTR-MS 和 GC/MS 检测的奶酪饼干数据进行分析,研究其存储时间。

但在处理过程中,每次合并分类将会直接影响接下来对新类的处理,造成每一个步骤的效果变差,降低聚类结果的质量。针对此问题,人们发展了双聚类算法,通过分别对矩阵的行和列聚类,然后合并聚类结果[51-52]。

3.2　有监督学习方法

有监督学习方法利用已知某种特征的样本进行训练,建立数学模型,再利用这一模型将所有新数据样本映射为响应的输出结果,从而实现预测的目的。因此,有监督学习方法的主要目标是发现样本与变量响应之间的关系。为了检测判别模型的识别能力,通常采用另一组已知类别的样本组成测试集,将训练中得到的正确判断率作为识别率,用测试样本集所得到的准确识别率称为预测率,一般情况下,识别率均优于预测率[53]。有监督学习方法包含很多,这里只介绍 PTR-MS 中常用的分析方法。

3.2.1　偏最小二乘判别　偏最小二乘判别分析(partial least squares discrimination analy-

sis，PLS-DA)是偏最小二乘回归分析的变形，是在很大程度上可以取代主成分分析、多元线性回归的判别分析统计方法。不同于主成分分析，PLS是同时对自变量矩阵(样本数据矩阵)X和相应变量Y进行分解，并力图建立它们之间的回归关系。它适用于解释多变量，并且存在多重共线性、观测样本少以及干扰较大的情况，尤其对于二元分类问题，可以获得很好的分类效果[54-55]。

Ruth等[56]通过PTR-MS与PLS-DA的结合，成功区分了牛奶脂肪。首先，利用PTR-MS分析、感官分析和经典化学分析3种方式评估食品的处理过程。随后，利用PLS-DA处理PTR-MS数据，预测基质(黄油/奶油)以及样品的感官等级。采用十倍交叉验证机制模型，正确区分了89%的样品。结果表明，PTR-MS和PLS-DA的结合是质量控制和制度控制的潜在应用方法。

Nooshin等[57]用PTR-MS法分析了来自5个欧洲国家的192个橄榄油样品的顶空挥发性化合物，提出了3个不同偏最小二乘法PLS-DA模型，分别用于区分样品的原产国、意大利国内的原产地和更小范围的产地，并用交叉验证方法评估模型的正确率。第一个模型对于区分橄榄油的原产国有86%的正确率，其中只有法国的正确率较低，为40%；第二个模型，只适用于区分意大利国内不同原产地的橄榄油，其正确率达到了74%；第三个模型，则用于意大利国内更小地域的橄榄油产地区分，正确率只有52%，这可能是因为在该尺度内橄榄油VOCs成分的组成及比例较为近似。

随着对PLS-DA法的应用与研究[58-60]，对该方法进行了很多改进。正交偏最小二乘判别分析(orthogonal PLS-DA，OPLS-DA)利用正交信号校正思想，滤除了自变量矩阵和相应变量矩阵的无关信息，所以OPLS-DA能够更好地区分组间差异，提高模型的有效性和解析能力，更加适用于多类区分问题[61]。

3.2.2 随机森林 决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树[62]。Brieman等[63]提出的随机森林(random forest，RF)算法是一种包含多个互不关联决策树的分类器，其构建主要考虑数据的随机性选取和待选特征的随机选取，对异常值和噪声具有很好的容忍度，且不容易出现过拟合。

Kistler等[64]利用PTR-TOF首次研究了饮食特点对营养性肥胖症老鼠呼气的影响。在呼气成分数据分析中，用RF进行特征识别，但是存在信息丢失、分类效果差等问题[65]。为了解决这些问题，他们提出了RF++算法，在传统RF算法中执行基于主题的引导过程，通过包外误差对模型进行评估。实验结果表明，从以植物为原料的饮食到几种半纯化加工食物的变化，会影响实验鼠呼气中VOCs的特征。

Granitto等[66-67]利用随机森林-递归特征消除(RF-RFE)算法对PTR-MS分析工农业产品产生的谱图进行相关特征鉴别，并与支持向量机-递归特征消除(SVM-RFE)方法做比对，利用多次重复的实验估计无偏的泛化误差。结果表明，在小种群的特征提取方面，RF-RFE法比SVM-RFE法更可靠，RF-RFE比SVM-RFE更适用于指纹识别工农业产品的PTR-MS谱图。

3.2.3 其他有监督学习方法 判别分析是在分类已经确定的条件下，根据某一研究对象的特征判别其类型归属的一种多元统计方法。按照数学模型可分为线性判别和非线性判别。其判别准则有多种，例如费舍尔准则、最小平方准则、最小距离准则、最大概率准则等[68]。Thekedar等[69]利用线性判别分析方法(linear discriminant analysis，LDA)将病人和对照组的呼气成分与室内空气成功进行了区分，进一步减小了外源性环境因素产生的VOCs背景影响。

人工神经网络(artificial neural network，ANN)是由大量处理单元互联组成的非线性、自适应信息处理系统。它是在现代神经科学研究成果的基础上提出的，试图通过模拟大脑神经网络处理、记忆信息的方式进行信息处理。Cancilla等[70]利用ANN模型对18位肺癌患者和22位健康人员的呼气成分的PTR-MS谱图进行了分析，讨论了是否考虑葡萄糖摄取因

素下的统计结果。当不考虑葡萄糖摄取因素时,只用8种质量数作为独立变量就可以建立一个MLP模型,精准度达到93%,所挑选出的质量数有助于检测和诊断肺癌疾病。

3.3 机器学习方法在PTR-MS谱图分析中面临的困难

不同的机器学习方法有各自的优缺点,这决定了它们的适用性。关于算法本身的特点,这里不再论述。本文主要结合PTR-MS谱图的特点,介绍这些算法在应用中面临的困难。相比于GC/MS,PTR-MS虽然具有快速检测的特点,但是定性能力不足,缺少色谱分离过程,且只能获得单一质荷比信息,所以有很多成分无法区分,其谱图中的同一质荷比信号可能对应多个VOCs成分。具有高分辨能力的PTR-TOF MS虽可进一步分离谱图信号,但仍不能分辨同分异构体,例如,二甲苯和乙苯,丙酮和丙醛等。另外,利用质子转移反应这种化学软电离方式得到的VOCs质谱峰,也会存在多个碎片离子峰。例如,乙酸乙酯($C_4H_8O_2$)与H_3O^+发生质子转移反应后,谱图中会有m/z 61、43、89三种质荷比信号,它们的比例随着漂移管内E/N比值的大小而变化。所以,PTR-MS中的每个质荷比信号可能对应多个VOC,而每个VOC在PTR-MS产生的谱图中可能对应多个质荷比信号,但目标成分通常是痕量的,极有可能被干扰离子掩盖。此外,同一VOC产生的不同质荷比信号,其内在的比例关系可能干扰统计分析算法在一些筛查标志物实验中的应用。这些问题将给PTR-MS用于VOCs的检测分析带来诸多困难,尤其是对未知混合物的分析。

4 总结与展望

PTR-MS在VOCs检测方面有着独特的优势,随着其在多个领域的广泛应用,数据处理方法面临着越来越多的问题和难点。本文总结了符合PTR-MS仪器特点的数据预处理方法和机器学习方法在PTR-MS谱图分析上的应用。在数据预处理方面,重点描述了无需标定的浓度计算方法;在数据分析方面,概括了不同算法的特点和不足,并归纳了不同算法的典型应用。

本课题组在PTR-MS仪器研制和应用方面做了一些研究并取得了阶段性成果。目前,本课题组自主研制的一台PTR-MS仪器整机已搭建完成,正处于性能参数调试阶段。在呼气检测和食品领域进行了应用研究,取得了一定的成果[71]。李子晓等[72]对呼气成分分析中湿度和CO_2的影响进行了分析。申丹宁等[73]用PTR-MS检测了不同品种和同品种不同产地橙汁的顶空挥发性气体,通过PCA区分了不同品种和产地的橙汁,并用费舍尔判别法建立了橙汁品种和产地的鉴别模型。郭冰清等[74]利用PTR-MS对肺癌患者呼气中特异性VOCs进行研究,建立了标准的临床试验方案,利用PTR-MS对40名肺癌患者、32名健康志愿者呼出气体中的VOCs进行检测,并进一步采用秩和分析、人工神经网络、随机森林、支持向量机和二元Logistic回归对全部呼气数据进行数据挖掘,发现了3种高可靠性的呼气特征生物标记物,并建立了相应的分类模型。结果表明,利用支持向量机建立的分类模型灵敏度为99.2%,特异性为98.5%,可对未知人群的患癌情况进行早期预判。

随着大数据时代的到来,不同仪器平台的整合、不同样品的数据融合是未来趋势。在这种背景下,PTR-MS会面临更多的挑战。合理的数据预处理技术以及机器学习方法,将会对数据分析起到越来越重要的作用,使PTR-MS技术的应用更加广泛。

参考文献:

[1] LINDINGER W, HANSEL A, JORDAN A. On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) medical applications, food control and environmental research[J]. International Journal of Mass Spectrometry & Ion Processes, 1998, 173(3): 191-241.

[2] BLAKE R S, MONKS P S, ELLIS A M. Proton-transfer reaction mass spectrometry[J]. Chemical Reviews, 2009, 109(3): 861.

[3] De L C B, AMANN A, ALKATEB H, et al. A review of the volatiles from the healthy human body[J]. Journal of Breath Research, 2014, 8

(1): 014001.

[4] CAPPELLIN L, BIASIOLI F, GRANITTO P M, et al. On data analysis in PTR-TOF-MS: from raw spectra to data mining[J]. Sensors & Actuators B Chemical, 2011, 155(1): 183-190.

[5] MÜLLER M, GRAUS M, RUUSKANEN T M, et al. First eddy covariance flux measurements by PTR-TOF[J]. Atmospheric Measurement Techniques, 2010, 3(2): 387-395.

[6] 张长水,杨强. 机器学习及其应用[M]. 北京:清华大学出版社,2013.

[7] 柯朝甫,张涛,武晓岩,等. 代谢组学数据分析的统计学方法[J]. 中国卫生统计,2014,31(2):357-359.
KE Chaofu, ZHANG Tao, WU Xiaoyan, et al. Statistical methods for metabolomics data analysis[J]. Chinese Journal of Health Stastistics, 2014, 31(2): 357-359(in Chinese).

[8] 蔡延亮. 统计机器学习方法在蛋白质组学中的应用[D]. 北京:北京大学,2011.

[9] FAZZINI M. Multidimensional statistical analysis of PTR-MS breath samples: a test study on irradiation detection[J]. International Journal of Mass Spectrometry, 2010, 295(1/2):13-20.

[10] HANSEL A, JORDAN A, HOLZINGER R, et al. Proton transfer reaction mass spectrometry: on-line trace gas analysis at the ppb level[J]. International Journal of Mass Spectrometry & Ion Processes, 1995, s149-150(150): 609-619.

[11] HANSEL A, MÄRK T D. Foreword[J]. International Journal of Mass Spectrometry, 2004, 239(2/3): Ⅶ-Ⅷ.

[12] THEKEDAR B. Investigations on the use of breath gas analysis with proton transfer reaction mass spectrometry (PTR-MS) for a non-invasive method of early lung cancer detection[J]. Acta Geodaetica Et Cartographic Sinica, 2009.

[13] STEEGHS M M L, SIKKENS C, CRESPO E, et al. Development of a proton-transfer reaction ion trap mass spectrometer: online detection and analysis of volatile organic compounds[J]. International Journal of Mass Spectrometry, 2007, 262(1/2): 16-24.

[14] MÜLLER M, MIKOVINY T, FEIL S, et al. A compact PTR-ToF-MS instrument for airborne measurements of VOCs at high spatio-temporal resolution[J]. Atmospheric Measurement Techniques Discussions, 2014, 7(6): 5 533-5 558.

[15] HO T J, KUO C H, WANG S Y, et al. True ion pick (TIPick): a denoising and peak picking algorithm to extract ion signals from liquid chromatography/mass spectrometry data[J]. Journal of Mass Spectrometry Jms, 2013, 48(2): 234.

[16] NILS H, MATTHIAS K, HEIKO N, et al. Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets[J]. Bmc Bioinformatics, 2012, 13(1): 1-20.

[17] 信号处理的小波导引:稀疏方法[M]. 北京:机械工业出版社,2003.

[18] HAYWARD S, HEWITT C N, SARTIN J H, et al. Performance characteristics and applications of a proton transfer reaction-mass spectrometer for measuring volatile organic compounds in ambient air[J]. Environmental Science & Technology, 2002, 36(7): 1 554-1 560.

[19] HAMILTON J D. Time series analysis[M]. Time Series Analysis, 1994: 401-409.

[20] KANTZ H. Nonlinear time series analysis[M]. Beijing: World Publishing Corportation, 2015.

[21] STURM B L. A wavelet tour of signal processing[M]. Academic Press, 1999: 83-85.

[22] GOUW J A D, GOLDAN P D, WARNEKE C, et al. Validation of proton transfer reaction-mass spectrometry (PTR-MS) measurements of gas-phase organic compounds in the atmosphere during the New England Air Quality Study (NEAQS) in 2002[J]. Journal of Geophysical Research Atmospheres, 2003, 72(108): 4 682.

[23] TANI A, HAYWARD S, HANSEL A, et al. Effect of water vapour pressure on monoterpene measurements using proton transfer reaction-mass spectrometry (PTR-MS)[J]. International Journal of Mass Spectrometry, 2004, 239(2/3): 161-169.

[24] SINHA V, CUSTER T G, KLUEPFEL T, et al. The effect of relative humidity on the detection of pyrrole by PTR-MS for OH reactivity measurements[J]. International Journal of Mass Spectrometry, 2009, 282(3): 108-111.

[25] De G J, WARNEKE C. Measurements of volatile organic compounds in the earth's atmosphere using proton-transfer-reaction mass spectrometry [J]. Mass Spectrometry Reviews, 2007, 26(2):

223-257.

[26] KEMPER P. Gas phase ion-molecule reaction rate constants through 1986[J]. International Journal of Mass Spectrometry & Ion Processes, 1988, 84(3): R17-R18.

[27] ZHAO J, ZHANG R. Proton transfer reaction rate constants between hydronium ion (H_3O^+) and volatile organic compounds[J]. Atmospheric Environment, 2004, 38(14): 2 177-2 185.

[28] KECK L, OEH U, HOESCHEN C, et al. Corrected equation for the concentrations in the drift tube of a proton transfer reaction-mass spectrometer (PTR-MS)[J]. International Journal of Mass Spectrometry, 2007, 264(1): 92-95.

[29] MITCHELL T M. Machine learning[M]. Beijing: China Machine Press, 2003.

[30] 王雪松,程玉虎. 机器学习理论、方法及应用[M]. 北京:科学出版社,2009.

[31] BROADHURST D I, KELL D B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments[J]. Metabolomics, 2006, 2(4): 171-196.

[32] 梅长林,周家良. 实用统计方法[M]. 北京:科学出版社,2005.

[33] BOSCAINI E, MIKOVINY T, WISTHALER A, et al. Characterization of wine with PTR-MS [J]. International Journal of Mass Spectrometry, 2004, 239(2): 215-219.

[34] FARNETI B, CRISTESCU S M, COSTA G, et al. Rapid tomato volatile profiling by using proton-transfer reaction mass spectrometry (PTR-MS)[J]. Journal of Food Science, 2012, 77(5): C551.

[35] 张丹丹,韦航,邱晓红,等. 基于质子转移反应-飞行时间质谱快速鉴别不同产地闽北水仙茶[J]. 分析化学,2017,45(6):914-921.
ZHANG Dandan, WEI Hang, QIU Xiaohong, et al. Rapid identification of Shuixian tea in northern Fujian from different regions by proton transfer reaction-time of flight-mass spectrometry [J]. Chinese Journal of Analytical Chemistry, 2017, 45(6): 914-921(in Chinese).

[36] SMVAN R, FRASNELLI J, CARBONELL L. Volatile flavour retention in food technology and during consumption: juice and custard examples [J]. Food Chemistry, 2008, 106(4): 1 385-1 392.

[37] LUTHRIA D L, LIN L Z, ROBBINS R J, et al. Discriminating between cultivars and treatments of broccoli using mass spectral fingerprinting and analysis of variance-principal component analysis [J]. Journal of Agricultural & Food Chemistry, 2008, 56(21): 9 819-9 827.

[38] GASPERI F, GALLERANI G, BOSCHETTI A, et al. The mozzarella cheese flavour profile: a comparison between judge panel analysis and proton transfer reaction mass spectrometry[J]. Journal of the Science of Food & Agriculture, 2001, 81(3): 357-363.

[39] CAMPBELL-SILLS H, CAPOZZI V, ROMANO A, et al. Advances in wine analysis by PTR-TOF-MS: optimization of the method and discrimination of wines from different geographical origins and fermented with different malolactic starters[J]. International Journal of Mass Spectrometry, 2016, (397/398): 42-51.

[40] DASZYKOWSKI M, KACZMAREK K, HEYDEN Y V, et al. Robust statistics in data analysis-A review: Basic concepts[J]. Chemometrics & Intelligent Laboratory Systems, 2007, 85(2): 203-219.

[41] APREA E, BIASIOLI F, CARLIN S, et al. Rapid white truffle headspace analysis by proton transfer reaction mass spectrometry and comparison with solid-phase microextraction coupled with gas chromatography/mass spectrometry[J]. Rapid Communications in Mass Spectrometry Rcm, 2007, 21(16): 2 564-2 572.

[42] YENER S, ROMANO A, CAPPELLIN L, et al. Tracing coffee origin by direct injection headspace analysis with PTR/SRI-MS[J]. Food Research International, 2015, 69: 235-243.

[43] RUTH S M V, ROZIJN M, KOOT A, et al. Authentication of feeding fats: classification of animal fats, fish oils and recycled cooking oils [J]. Animal Feed Science & Technology, 2010, 155(1): 65-73.

[44] SCHUHFRIED E, SÁNCHEZ D P J, BOBBA M, et al. Classification of 7 monofloral honey varieties by PTR-TOF-MS direct headspace analysis and chemometrics[J]. Talanta, 2016, 147: 213-219.

[45] DASZYKOWSKI M, KACZMAREK K, HEYDEN Y V, et al. Robust statistics in data analy-

sis-a review: basic concepts[J]. Chemometrics & Intelligent Laboratory Systems, 2007, 85(2): 203-219.

[46] 方开泰,潘恩沛. 聚类分析[M]. 北京:地质出版社,1982.

[47] WEBB A R, COPSEY K D. Statistical pattern recognition, third edition[J]. John Wiley & Sons Inc, 2002, 265(2): 183-190.

[48] SÁNCHEZLÓPEZ J A, ZIMMERMANN R, YERETZIAN C. Insight into the time-resolved extraction of aroma compounds during espresso coffee preparation: online monitoring by PTR-TOF-MS[J]. Analytical Chemistry, 2014, 86(23): 11 696-11 704.

[49] CIESA F, HÖLLER I, GUERRA W, et al. Chemodiversity in the fingerprint analysis of volatile organic compounds (VOCs) of 35 old and 7 modern apple cultivars determined by proton-transfer-reaction mass spectrometry (PTR-MS) in two different seasons[J]. Chemistry & Biodiversity, 2015, 12(5): 800.

[50] POZO-BAYÓN M, REINECCIUS G A. Monitoring changes in the volatile profile of cheese crackers during storage using GC-MS and PTR-MS[J]. Flavour & Fragrance Journal, 2010, 24(3): 133-139.

[51] BRO R, PAPALEXAKIS E E, ACAR E, et al. Coclustering-a useful tool for chemometrics[J]. Journal of Chemometrics, 2012, 26(6): 256-263.

[52] MADEIRA S C, OLIVEIRA A L. Biclustering algorithms for biological data analysis: a survey [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2004, 1(1): 24.

[53] 许国旺. 代谢组学:方法与应用[M]. 北京:科学出版社,2008.

[54] WANG C, KONG H, GUAN Y, et al. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis[J]. Analytical Chemistry, 2005, 77(13): 4 108-4 116.

[55] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京:国防工业出版社,1999.

[56] van RUTH S M, KOOT A, AKKERMANS W, et al. Butter and butter oil classification by PTR-MS[J]. European Food Research & Technology, 2008, 227(1): 307-317.

[57] NOOSHIN A, JENNIFER C, ALEX K, et al. Geographical origin classification of olive oils by PTR-MS[J]. Food Chemistry, 2008, 108(1): 374-383.

[58] MAÇATELLI M, AKKERMANS W, KOOT A, et al. Verification of the geographical origin of European butters using PTR-MS[J]. Journal of Food Composition & Analysis, 2009, 22(2): 169-175.

[59] APREA E, CAPPELLIN L, GASPERI F, et al. Application of PTR-TOF-MS to investigate metabolites in exhaled breath of patients affected by coeliac disease under gluten free diet[J]. Journal of Chromatography B, 2014, 966: 208-213.

[60] RUTH S M V, VILLEGAS B, AKKERMANS W, et al. Prediction of the identity of fats and oils by their fatty acid, triacylglycerol and volatile compositions using PLS-DA[J]. Food Chemistry, 2010, 118(4): 948-955.

[61] TRYGG J, WOLD S. Orthogonal projections to latent structures (O-PLS)[J]. Journal of Chemometrics, 2010, 16(3): 119-128.

[62] HAN J W, KAMBER M, PEI J, et al. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2012.

[63] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[64] KISTLER M, SZYMCZAK W, FEDRIGO M, et al. Effects of diet-matrix on volatile organic compounds in breath in diet-induced obese mice [J]. Journal of Breath Research, 2014, 8(1): 016004.

[65] KARPIEVITCH Y V, HILL E G, LECLERC A P, et al. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++[J]. Plos One, 2009, 4(9): e7087.

[66] GRANITTO P M, BIASIOLI F, FURLANELLO C, et al. Efficient feature selection for PTR-MS fingerprinting of agroindustrial products[C]. International Conference on Artificial Neural Networks. Springer-Verlag, 2008: 42-51.

[67] GRANITTO P M, FURLANELLO C, BIASIOLI F, et al. Recursive feature elimination with random forest for PTR-MS analysis of agroindus-

trial products[J]. Chemometrics & Intelligent Laboratory Systems, 2006, 83(2): 83-90.

[68] 赵学玨,李维康,杜康,等. 质子转移反应质谱在呼气检测领域的研究进展[J]. 生物医学工程学杂志,2015,(6):1 374-1 379.
ZHAO Xuehong, LI Weikang, DU Kang, et al. Research progress of proton transfer reaction mass spectrometry in the field of breathing gas detection[J]. Journal of Biomedical Engineering, 2015, (6): 1 374-1 379(in Chinese).

[69] THEKEDAR B. Investigations on the use of breath gas analysis with proton transfer reaction mass spectrometry (PTR-MS) for a non-invasive method of early lung cancer detection[J]. Acta Geodaetica Et Cartographic Sinica, 2009.

[70] B C. Non-invasive diagnosis of human diseases by combining breath analysis and neural network modeling[D]. Madrid: Universidad Complutense, 2016.

[71] 李维康. PTR-MS 关键部件的研制及呼气丙酮测量方法的研究[D]. 天津:天津大学,2014.

[72] 申丹宁,赵学玨,孙运,等. 质子转移反应质谱在食品挥发性有机物检测分析中的应用[J]. 食品科学,2017,38(23):289-297.
SHEN Danning, ZHAO Xuehong, SUN Yun, et al. Application of proton transfer reaction mass spectrometer in analyze of bolatile organic compounds in food[J]. Food Science, 2017, 38(23): 289-297(in Chinese).

[73] 李子晓,赵学玨,李维康,等. 质子转移反应质谱法测量呼气丙酮的影响因素分析[J]. 质谱学报,2016,37(4):351-358.
LI Zixiao, ZHAO Xuehong, LI Weikang, et al. Analysis of influence factors for the breath acetone measurement by proton transfer reaction mass spectrometry[J]. Journal of Chinese Mass Spectrometry Society, 2016, 37(4): 351-358(in Chinese).

[74] 郭冰清. 基于 PTR-MS 技术的肺癌患者呼出气体中检测分析方法的研究[D]. 天津:天津大学,2017.