doi：10. 3969/j. issn. 1004 – 4957. 2020. 11. 015

基于近红外光谱技术的电子烟油烟碱含量快速检测研究

杨双艳[1]，周　瑾[1]，沈彦文[1]，杨紫刚[1]，费　宇[2*]，张四伟[3]

（1. 云南巴菰生物科技有限公司，云南　昆明　650000；2. 云南财经大学　统计与数学学院，云南　昆明　650000；3. 云南省烟草公司文山州公司，云南　文山　663000）

摘　要：烟碱是电子烟烟油中的主要成分，其含量决定了电子烟油的风味口感及产品的安全性。为了提高电子烟油烟碱含量的测量效率，该文采用近红外光谱技术和极限学习机回归（ELMR）建立了电子烟油烟碱含量的定量预测模型。实验结果表明：相比于传统的主成分回归（PCR）和偏最小二乘回归（PLSR）模型，所建立的 ELMR 预测模型的决定系数 R^2 为 0. 926 2，远高于 PCR 预测模型的 0. 859 0 和 PLSR 预测模型的 0. 860 4；同时，使用 ELMR 模型的预测均方根误差（RMSEP）为 0. 026 8，小于 PCR 预测模型的 0. 043 1 和 PLSR 预测模型的 0. 040 9。以上结果说明该文所建立的近红外光谱定量模型能够应用于烟碱含量的快速准确测量，为实现电子烟油烟碱含量的实时在线监测和其它质量参数的快速测量奠定了良好的基础。

关键词：电子烟油；烟碱含量；近红外光谱；极限学习机；快速检测

中图分类号：O657. 3；TS41　　**文献标识码**：A　　**文章编号**：1004 – 4957(2020)11 – 1411 – 05

Rapid Determination of Nicotine Content of E-cigarette Liquid Based on Near-infrared Spectroscopy Technology

YANG Shuang-yan[1], ZHOU Jin[1], SHEN Yan-wen[1], YANG Zi-gang[1], FEI Yu[2*], ZHANG Si-wei[3]

(1. Yunnan Tobacco Biological Technology Co., Ltd, Kunming 650000, China; 2. School of Stastics and Mathematics, Yunnan University of Finance and Economics, Kunming 650000, China; 3. Wenshan Branch Company, Yunnan Tobacco Company, Wenshan 663000, China)

Abstract: Nicotine is the most important component in E-cigarette liquid, whose content determines the flavor and the safety of the product. In order to improve the detecting efficiency of the nicotine content, a novel near-infrared spectroscopy (NIR) combined with extreme learning machine regression (ELMR) algorithm was adopted to establish an NIR – ELMR prediction model for nicotine content in E-cigarette liquid. The experimental results showed that, compared with traditional partial least squares regression (PLSR) model and principal component regression (PCR) model, the NIR – ELMR model was much better with a determination coefficient (R^2) of 0. 926 2, which was higher than 0. 859 0 for PCR prediction model and 0. 860 4 for PLSR prediction model. Besides, the root mean square error of prediction (RMSEP) for NIR – ELMR model was 0. 026 8, which was smaller than 0. 043 1 for PCR model and 0. 040 9 for PLSR model. The above results indicated the established model could be applied to the rapid and accurate determination of the nicotine content of E-cigarette liquid, which lay a foundation for the online analysis of nicotine content and the rapid determination of other quality parameters.

Key words: E-cigarette liquid; nicotine content; near-infrared spectroscopy (NIR); extreme learning machine (ELM); rapid determination

电子烟在传递尼古丁的过程中不需要对烟草进行燃烧，相比传统香烟更加安全且具有更少的有害成分，因此逐渐成为传统香烟新的替代品[1]。烟碱作为电子烟烟油中最主要的成分，其含量决定了电子烟油的风味口感及产品的安全性，一些国家和地区相继将电子烟烟油中的烟碱纳入监管范围。目前，

收稿日期：2020 – 03 – 10；**修回日期**：2020 – 04 – 12

基金项目：国家自然科学基金资助项目(11971421)

* **通讯作者**：费　宇，博士，教授，研究方向：统计理论与方法、应用统计分析，E – mail：feiyukm@ aliyun. com

对电子烟烟油中烟碱的检测大多参考卷烟烟草的检测方法，主要采用气相色谱法和液相色谱法进行测定，但这些方法存在检测时间长、样品预处理繁琐、费用高、对操作人员要求高等缺点。因此，研究开发一种准确、快速、无损的检测方法获得电子烟油的烟碱指标对于控制电子烟油的品质和工艺具有重大意义。

近红外光谱(NIR)分析技术具有简便、快速、前处理简单、对样品无破坏性、无污染并可多组分同时测定等优点[2]，在农业[3-4]、石油[5-6]、烟草[7-9]等领域有着广泛应用，但目前尚未见采用近红外光谱对电子烟油进行检测的研究。电子烟油中有机组分的化学和物理信息在近红外光谱中均有体现，因此近红外光谱非常适合对电子烟油进行分析检测。

为了解决反向传播算法(Backward probagation)学习效率低、参数设定繁琐的问题，2004 年 Huang 等[10]提出极限学习机(Extreme learning machine，ELM)算法，并发表于当年的 IEEE 国际交互会议(IEEE International Joint Conference)。ELM 是一类基于前馈神经网络(Feedforward neuron network)的机器学习算法，其主要特点是隐含层节点参数可以随机或人为给定且不需要调整，学习过程仅需计算输出权重。ELM 具有学习效率高和泛化能力强的优点，被广泛应用于分类[11]、回归[12]、聚类[13]、特征学习[14]等问题中，但尚未见应用于电子烟油近红外光谱分析的相关研究。

本文以近红外光谱分析技术为基础，结合 ELM 算法对电子烟油的近红外光谱数据和烟碱指标进行定量建模。与现有检测方法相比，本文所提出的方法具有快速准确、绿色无损等优点，能够实现电子烟油烟碱指标的快速准确测量，为电子烟油重要理化指标的实时在线监测和其它质量参数的快速测量奠定了良好的基础。

1 极限学习机算法的基本理论

极限学习机与传统的梯度下降学习算法相比具有较大优势：(1)随机给定隐含层的连接权值，训练过程不需要迭代调整，计算速度非常快；(2)传统的梯度下降算法容易陷入局部极小，而 ELM 算法由于求解输出权值最小二乘解的过程是一个凸优化问题，因此不会陷入局部最优；(3)参数选择简单，只需选择合适的隐含层结点便可获得良好的性能，而传统的梯度下降算法，如 BP 网络等，需要选择合适的学习率、训练步长等，选择不当会影响网络的泛化性。

对于一个单隐层神经网络，假设有个任意的样本($\boldsymbol{t}_i$，$\boldsymbol{X}_i$)，其中：

$$\boldsymbol{X}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^{\boldsymbol{T}} \in \boldsymbol{R}^n \tag{1}$$

$$\boldsymbol{t}_i = [t_{i1}, t_{i2}, \cdots, t_{in}]^{\boldsymbol{T}} \in \boldsymbol{R}^m \tag{2}$$

对于一个有 N 个隐层节点的单隐层神经网络可以表示为：

$$\sum_{i=1}^{L} = \boldsymbol{\beta}_i g(\boldsymbol{W}_i \cdot \boldsymbol{X}_j + \boldsymbol{b}_i) \quad = \boldsymbol{o}_j,\ j = 1, \cdots, N \tag{3}$$

其中，$g(x)$为激活函数，$\boldsymbol{W}_i = [w_{i,1}, w_{i,2}, \cdots, w_{i,n}]^{\boldsymbol{T}}$ 为输入权重，$\boldsymbol{\beta}_i$ 为输出权重，$\boldsymbol{b}_i$ 是第 i 个隐层单元的偏置。$\boldsymbol{W}_i \cdot \boldsymbol{X}_j$ 表示 $\boldsymbol{W}_i$ 和 $\boldsymbol{X}_j$ 的内积。

单隐层神经网络学习的目标是使输出的误差最小，可以表示为：

$$\sum_{j=1}^{N} \| \boldsymbol{o}_j - \boldsymbol{t}_j \| = 0 \tag{4}$$

即存在$\boldsymbol{\beta}_i$，$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$，使得：

$$\sum_{i=1}^{L} \boldsymbol{\beta}_i g(\boldsymbol{W}_i \cdot \boldsymbol{X}_j + \boldsymbol{b}_i) = \boldsymbol{o}_j,\ j = 1, \cdots, N \tag{5}$$

可以矩阵表示为：

$$\boldsymbol{H\beta} = \boldsymbol{T} \tag{6}$$

其中，$\boldsymbol{H}$ 是隐层节点的输出，$\boldsymbol{\beta}$ 为输出权重，$\boldsymbol{T}$ 为期望输出。

$$\boldsymbol{H}(W_1, \cdots, W_L, b_1, \cdots, b_L, X_1, \cdots, X_N) = \begin{bmatrix} g(W_1 \cdot X_1 + b_1) & \cdots & g(W_L \cdot X_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(W_1 \cdot X_N + b_1) & \cdots & g(W_L \cdot X_N + b_L) \end{bmatrix}_{N \times L} \tag{7}$$

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_1^T\\ \vdots\\ \beta_L^T\end{bmatrix}_{L\times m},\quad \boldsymbol{T}=\begin{bmatrix}T_1^T\\ \vdots\\ T_N^T\end{bmatrix}_{N\times m} \tag{8}$$

为了能够训练单隐层神经网络，希望得到$\hat{\boldsymbol{\beta}}_i$，$\hat{\boldsymbol{W}}_i$ 和 $\hat{\boldsymbol{b}}_i$，使得：

$$\left\| \boldsymbol{H}\left(\hat{\boldsymbol{W}}_i,\ \hat{\boldsymbol{b}}_i\right)\hat{\boldsymbol{\beta}}_i-\boldsymbol{T}\right\| =\min_{W,b,\beta}\left\| \boldsymbol{H}(\boldsymbol{W}_i,\ \boldsymbol{b}_i)\boldsymbol{\beta}_i-\boldsymbol{T}\right\| \tag{9}$$

其中，$i=1$，…，L，这等价于最小化损失函数：

$$\boldsymbol{E}=\sum_{j=1}^{N}\left(\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{W}_i\cdot\boldsymbol{X}_j+\boldsymbol{b}_i)-\boldsymbol{t}_j\right)^2 \tag{10}$$

传统的一些基于梯度下降法的算法，可以用来求解式(10)中的问题，但是基本的基于梯度的学习算法需要在迭代的过程中调整所有参数。而在 ELM 算法中，一旦输入权重 $\boldsymbol{W}_i$ 和隐层的偏置 $\boldsymbol{b}_i$ 被随机确定，隐层的输出矩阵 $\boldsymbol{H}$ 就被唯一确定。训练单隐层神经网络可以转化为求解一个线性系统 $\boldsymbol{H\beta}=\boldsymbol{T}$。并且输出权重可以被确定：

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{-1}\boldsymbol{T} \tag{11}$$

其中，$\boldsymbol{H}^{-1}$是矩阵 $\boldsymbol{H}$ 的 Moore－Penrose 广义逆。且可证明求得的解$\hat{\boldsymbol{\beta}}$ 的范数最小且唯一。

2　实验部分

2.1　仪器与样本

样本的近红外光谱采集使用 Antaris 傅里叶变换近红外光谱仪(Thermo Nicolet，USA)，配有透射检测器，采样系统和 Result、TQ Analyst 等数据处理软件；实验样本由云南巴菰生物科技有限公司提供，共 70 个样本。实验过程中，按照样本烟碱含量从低到高均匀分布的原则选取 40 个样本作为训练样本，30 个样本作为测试样本；使用气相色谱仪/氢火焰离子化检测器获取电子烟油的烟碱含量，训练样本的烟碱含量范围为 1～60 mg/g，平均值为 27.98 mg/g，标准差为 15.96；测试样本的烟碱含量范围为 3～52 mg/g，平均值为 27.37 mg/g，标准差为 14.80。实验样本的详细信息见表 1。

表 1　实验样本的详细信息

Table 1　Detail information of experimental samples

Year	Name	Sample	Dimension of spectroscopy	Vaule of nicotine (mg/g)	Number of samples	Average vaule (mg/g)	Standard deviation
2018	E-cigarette liquid	Training	1 550	1～60	40	27.98	15.96
2018	E-cigarette liquid	Testing	1 550	3～52	30	27.37	14.80

2.2　近红外光谱采集

近红外光谱仪的相关参数设置：光谱采集模式为透射模型，数据格式为 Absorbance，扫描次数为 32，分辨率为 4 cm^{-1}，光纤透射式探头光程为 2 mm，以空气为参比，光谱扫描范围为 4 000～10 000 cm^{-1}。将烟油样本滴入石英皿中，每个样本重复采样 3 次，取 3 次光谱的平均值作为样本的最终光谱。实验样本的原始近红外光谱数据如图 1 所示。

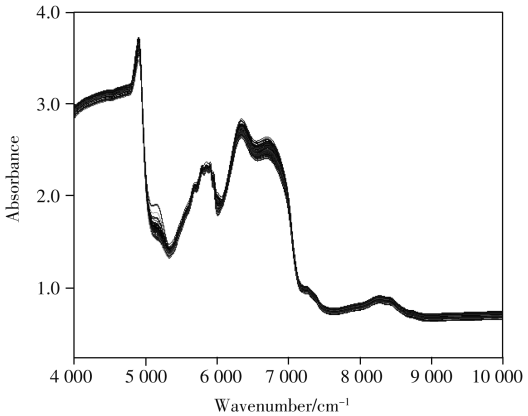

图 1　实验样本的原始近红外光谱数据

Fig. 1　Original NIR data of the samples

2.3　数据处理方法与模型性能评价指标

首先对采集的电子烟油的近红外光谱数据进行预处理操作，并选择合适的波段，分别采用主成分回归(Principal component regression，PCR)[15]、偏最小二乘回归(Partial least squares regression，PLSR)[16]和极限学习机回归(Extreme learning machine regression，ELMR)建立近红外光谱数据与烟碱含量之间的定量校正模型。使用决定系数

(R^2)、校正均方根误差(Root mean square error of calibration，RMSEC)、预测均方根误差(Root mean square error of prediction，RMSEP)为指标优化建模参数，用以考察模型性能，以上参数的计算方法见文献[17－18]。

3 结果与讨论

对近红外光谱数据进行分析和处理时，其中一个重要步骤是对光谱数据进行预处理操作。光谱的预处理操作能够降低或消除非目标因素对光谱信息的影响，通过对其进行适当的数学操作，能够最大程度去除冗余信息，从而更利于从复杂光谱中提取有效信息，在一定程度上提高校正模型的稳健性。本文通过多元散射校正和 Savitzky－Golay 一阶导数(窗口大小为 5，3 次多项式)滤波的方法进行预处理操作，处理结果如图 2 所示。可以看出，经过预处理的光谱图像有效消除了光谱的基线漂移现象。从图 2 还可以看出，光谱的吸收波长区间主要集中在 4 492～7 864 cm^{-1}。因此，随后将主要使用此波长区间对电子烟油的近红外光谱数据与样本的烟碱含量进行定量建模。

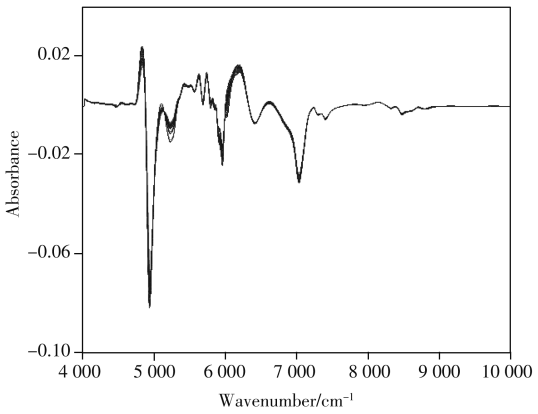

图 2 原始光谱经过多元散射校正和 Savitzky－Golay 一阶导数(窗口大小为 5，3 次多项式)滤波后的预处理结果

Fig. 2 Pretreatment result of spectral data by means of using multiplicative scatter correction and Savitzky－Golay first derivative with a 5-point window and three polynomial order

分别采用 PCR、PLSR 和 ELMR 建立近红外光谱数据和传统化学方法测量所获得的烟碱含量之间的定量校正模型，并以 R^2、RMSEC、RMSEP 为指标优化建模参数，建模结果和测试结果分别如表 2 和表 3 所示。其中，使用 PCR 和 PLSR 进行光谱建模时，首先对光谱数据进行主成分降维处理，所选用的主成分数为 5。设置 ELM 算法的隐含层神经元数为 30，以 Sigmoidal 函数为隐含层神经元激励函数。由 ELM 算法的基本理论得知，输入权重 $\boldsymbol{W}_i$ 和隐层的偏置 $\boldsymbol{b}_i$ 将会在训练过程中随机确定，不需人工设定。

由表 2 可以看出，使用 ELMR 算法所建立校正集模型的 R^2 为 0.950 0，远高于 PCR 和 PLSR 算法；同时，ELMR 算法的 RMSEC 为 0.014 9，远低于 PCR 和 PLSR 算法。表 3 显示，在预测方面，ELMR 算法预测模型的 R^2 为 0.926 2，远高于 PCR 和 PLSR 算法；同时，使用 ELMR 算法的 RMSEP 为 0.026 8，远低于 PCR 和 PLSR 算法。因此，ELMR 算法在建模效果和预测结果方面，都取得了最高的决定系数和最小的均方根误差。上述结果证明，采用近红外光谱技术快速测定电子烟油的烟碱含量时，使用 ELMR 算法建立的模型性能优于经典的 PCR 和 PLSR 算法。相对于传统方法，ELMR 提高了训练集的数据利用率，具有更好的范化性能和更高的回归预测精度，算法的预测精度高，泛化能力强，不容易出现过拟合倾向。

表 2 不同建模方法的烟碱训练结果

Table 2 Training results of nicotine using different modeling methods

Experimental samples	Number of samples	Parameter	Range of wavelength (cm^{-1})	PCR (principal component 5)		PLSR (principal component 5)		ELMR (hidden neuron 30)	
				R^2	RMSEC	R^2	RMSEC	R^2	RMSEC
E-cigarette liquid	40	Nicotine	4 492～7 864	0.858 8	0.043 3	0.877 2	0.042 0	0.950 0	0.014 9

表 3 不同建模方法测试样本的预测结果

Table 3 Prediction results of testing samples using different modeling methods

Experimental samples	Number of samples	Parameter	Range of wavelength (cm^{-1})	PCR (principal component 5)		PLSR (principal component 5)		ELMR (hidden neuron 30)	
				R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
E-cigarette liquid	30	Nicotine	4 492～7 864	0.859 0	0.043 1	0.860 4	0.040 9	0.926 2	0.026 8

4　结　论

本文以近红外光谱分析技术为基础，结合极限学习机算法对电子烟油进行近红外光谱定量建模。与现有检测方法相比，本文所提出的检测方法具有快速准确、绿色无损等优点，能够实现电子烟油烟碱含量的快速准确测量，为电子烟油烟碱含量的实时在线监测和其它质量参数的快速测量奠定了良好的基础。

参考文献：

[1] Cobb N K, Abrams D B. *New Engl. J. Med.*, **2011**, 365(3): 193 – 195.
[2] Chu X L, Shi Y Y, Chen P, Li J Y, Xu Y P. *J. Instrum. Anal.* (褚小立, 史云颖, 陈瀑, 李敬岩, 许育鹏. 分析测试学报), **2019**, 38(5): 603 – 611.
[3] Kovalenko I V, Rippke G R, Hurburgh C R. *J. Am. Oil Chem. Soc.*, **2006**, 83(5): 421 – 427.
[4] Das B, Sahoo R N, Pargal S, Krishna G, Verma R, Chinnusamy V, Sehgal V K, Gupta V K, Dash S K, Swain P. *Spectrochim. Acta A*, **2018**, 192: 41 – 51.
[5] Balabin R M, Safieva R Z, Lomakina E I. *Microchem. J.*, **2011**, 98(1): 121 – 128.
[6] Jiang L L, Luo M F, Zhang Y, Yu X J, Kong W W, Liu F. *Spectrosc. Spectral Anal.* (蒋璐璐, 骆美富, 张瑜, 余心杰, 孔汶汶, 刘飞. 光谱学与光谱分析), **2014**, 34(1): 64 – 68.
[7] Song X Z, Lai Y Q, Li Z H, Zheng B, Li Q Q, Wu L J, Zhang L D, Xiong Y M, Min S G. *J. Anal. Sci.* (宋相中, 赖衍清, 李祖红, 郑波, 李倩倩, 吴丽君, 张录达, 熊艳梅, 闵顺耕. 分析科学学报), **2014**, 30(3): 327 – 331.
[8] Zhang J Q, Liu W J, Yang Y M. *J. Braz. Chem. Soc.*, **2019**, 30(9): 1927 – 1932.
[9] Zhang J Q, Liu W J, Zhang H H, Hou Y, Yang P P, Li C Y, Yang Y M, Li M. *J. Near Infrared Spectrosc.*, **2018**, 26(2): 101 – 105.
[10] Huang G B, Zhu Q Y, Siew C K. *Neurocomputing*, **2006**, 70(1/3): 489 – 501.
[11] Heeswijk M V, Miche Y, Oja E, Lendasse A. *Neurocomputing*, **2011**, 74(16): 2430 – 2437.
[12] Jin Y, Li J, Lang C Y, Ruan Q Q. *Multidim. Syst. Signal Process.*, **2017**, 28(3): 905 – 920.
[13] Liu G H, Jiang H, Xiao X H, Zhang D J, Mei C L, Ding Y H. *Spectrosc. Spectral Anal.* (刘国海, 江辉, 肖夏宏, 张东娟, 梅从立, 丁煜函. 光谱学与光谱分析), **2012**, 32(4): 970 – 973.
[14] Jin Y, Cao J W, Wang Y Z, Zhi R C. *Multimed. Tools Appl.*, **2016**, 75(19): 11831 – 11846.
[15] Fang Y, Park J I, Jeong Y S, Jeong M K, Baek S H, Cho H W. *Ann. Operat. Res.*, **2011**, 190(1): 3 – 15.
[16] Chen Q S, Zhao J W, Liu M H, Cai J R, Liu J H. *J. Pharm. Biomed. Anal.*, **2008**, 46(3): 568 – 573.
[17] Blanco M, Gozález Bañó R, Bertran E. *Talanta*, **2002**, 56(1): 203 – 212.
[18] Xiang D, Berry J, Buntz S, Gargiulo P, Cheney J, Joshi Y, Wabuyele B, Wu H, Haned M, Hussain A S, Khan M A. *J. Pharm. Sci.*, **2009**, 98(3): 1155 – 1166.

（责任编辑：盛文彦）