第40卷，第11期
2020年11月

光谱学与光谱分析
Spectroscopy and Spectral Analysis

Vol. 40, No. 11, pp3457-3462
November, 2020

基于神经网络的拉曼光谱波长选择方法

沈东旭，洪明坚*，董家林

重庆大学大数据与软件学院，重庆 401331

摘 要 血液鉴别对于检验检疫、刑侦以及动物保护领域具有非常重要的意义，传统的的血液鉴别方法在鉴别的过程中存在分析周期长、对血液样本造成损害等缺点。而拉曼光谱可以通过分析与入射光频率不同的散射光谱得到分子振动、转动方面的信息，进而得到物质的组成成分，并且具有零污染非接触的特点，为血液的无损鉴别提供了可能，但是在拉曼光谱中，各个波长点之间存在严重的多重共线性，直接使用全光谱进行建模会增加模型的复杂性和降低模型的稳定性。针对拉曼光谱的特点，提出了一种基于神经网络的波长选择方法。该方法利用神经网络学习到各个波长点对校正模型的贡献权重，并将权重的均值作为阈值，去除权重低于阈值的波长点，以达到波长选择的目的。为了更容易确定筛选的阈值，在权重学习的过程中加入了稀疏约束，极大的减少了用于筛选的波长点。利用动物与人血清的拉曼光谱数据集对所提方法进行了验证，实验结果表明，利用该方法得到的光谱建立的校正模型，相比于全光谱数据在分类准确率和AUC值都有一定的提升，人工神经网络(NN)的准确率达到了94.495%，AUC值达到了0.9850，偏最小二乘(PLS-DA)的准确率达到了92.661%，AUC值达到了0.9760。与传统的波长选择方法UVE相比，该方法选择的波长点更少，仅选择了42个波长点用于建模，而且得到的校正模型的分类准确率和AUC值更高，证明该波长选择方法能有效的筛选出对建模有贡献的波长点，提高了模型的分类准确率和稳定性，为血液的无损鉴别提供了可能，具有一定的实用价值。

关键词 光谱分析；拉曼光谱；波长选择；神经网络；校正模型

中图分类号： O657.37 **文献标识码：** A **DOI：** 10.3964/j.issn.1000-0593(2020)11-3457-06

引 言

血液中包含了大量的遗传信息和基因信息，在出入境的检验检疫、刑侦以及动物保护等领域常常需要对未知的血液样本进行物种鉴别，所以对于血液鉴别的研究有非常重要的意义。传统的血液鉴别方法包括沉淀反应、凝聚反应、酶免疫分析、等电聚焦、高效液相色谱仪等[1]。这些技术都存在分析周期长、对血液样品会造成损害等缺点，但是许多领域要求在检测过程中尽可能的保证样品的完整，尤其是刑事侦测方面，需要利用有限的样品做进一步的分析。

拉曼光谱分析法是通过对与入射光频率不同的散射光谱进行分析以得到分子振动、转动方面的信息，进而得到物质的组成成分，它具有零污染、非接触等特点，为血液的无损鉴别提供了可能。到目前为止，已经有大量实验证明利用拉曼光谱能够区分不同物种的血液[2]。

在拉曼光谱中，光谱信息和预测成分之间在某些波段不存在特定的相关关系，除此之外，各个波长点之间也存在严重的多重共线性[3]。如果采用全光谱数据进行建模，不仅会增加校正模型的复杂性，而且其中一些噪声变量还会降低模型的预测能力。文献[4-9]中提到利用选定的波长点或波长区间而不是全光谱数据进行建模可以有效的提高校正模型的性能。因此，研究波长选择算法对于简化校正模型和提高校正模型的性能具有重要的实际意义。

现有的波长选择方法大致可以分为三类。第一类是利用与校正模型相关的统计信息对波长点或波长区间进行评估，一般是对各个波长点或波长区间逐一考察，决定该剔除哪一些波长点或波长区间，典型方法有区间偏最小二乘法(IPLS)[5]、移动窗口偏最小二乘法(MWPLS)[6]、SVP(stability and variable permutation)[7]等。第二类是根据各个波长点的回归系数、协相关系数等一些指标进行排序，然后选择剔除靠后的那些波长点，例如偏最小二乘无信息剔除法

收稿日期： 2019-09-25，**修订日期：** 2020-01-06

基金项目： 国家重点研发计划项目(2018YFF01011204)资助

作者简介： 沈东旭，1995年生，重庆大学大数据与软件学院硕士研究生 e-mail：Dongxu1995@outlook.com

* 通讯联系人 e-mail：hmj@cqu.edu.cn

(UVE-PLS)[8]、竞争自适应重加权抽样法(CARS)[4]和变量置换总体分析法[9]等。第三类是将波长选择看作全局优化问题，利用智能优化算法对波长点进行选择，遗传算法(GA)、粒子群优化算法(PSO)、模拟退火算法(SA)等算法均属于此类。

本文提出了一种基于神经网络的拉曼光谱波长选择方法。该方法利用神经网络的自学习能力，自适应得到各个波长点对校正模型的贡献权重，然后将权重的均值作为阈值，去掉权重低于阈值的波长点，从而达到波长选择的目的。利用该方法对人与动物血清的拉曼光谱进行波长选择，将波长选择后的光谱进行分类实验，取得了较好的实验结果。

1 波长选择网络

拉曼光谱中包含大量与建模无关的波长点，本文利用神经网络学习到各个波长点对校正模型的贡献权重，通过学得的权重筛选出与校正模型相关的波长点用于建模，有效的提升了校正模型的性能。波长选择网络的结构如图 1 所示，为了方便描述，将这个网络结构命名为 WSNet(wavelength selection network)。WSNet 分为波长选择和校正模型两个部分，波长选择模块(红色虚线框中的结构)由 FC1 层和 FC2 层两个全连接层构成，校正模型由 FC3 层、FC4 层两个全连接层和输出层(OUTPUT 层)构成。

FC1 层的作用是对输入的光谱数据进行降维，以防止误差反向传播时波长选择模块的网络结构梯度更新过大。FC1 层的变换过程可以表示为

$$\begin{cases} x_{\mathrm{FC1}} = A_{\mathrm{FC}}(W_{\mathrm{FC1}}x + b_{\mathrm{FC1}}) \\ A_{\mathrm{FC1}}(x) = \mathrm{ReLU}(x) = \max(x,0) \end{cases} \tag{1}$$

W_{FC1} 为 FC1 层的权值矩阵，b_{FC1} 为 FC1 层的偏置向量，$A_{\mathrm{FC1}}(\cdot)$ 为 FC1 层的激活函数，保证权重大于零，x_{FC} 表示 FC1 层的输出。FC1 层的神经元个数为 C/r，C 为输入的波长点个数，$r(r>1)$ 为降维倍数。

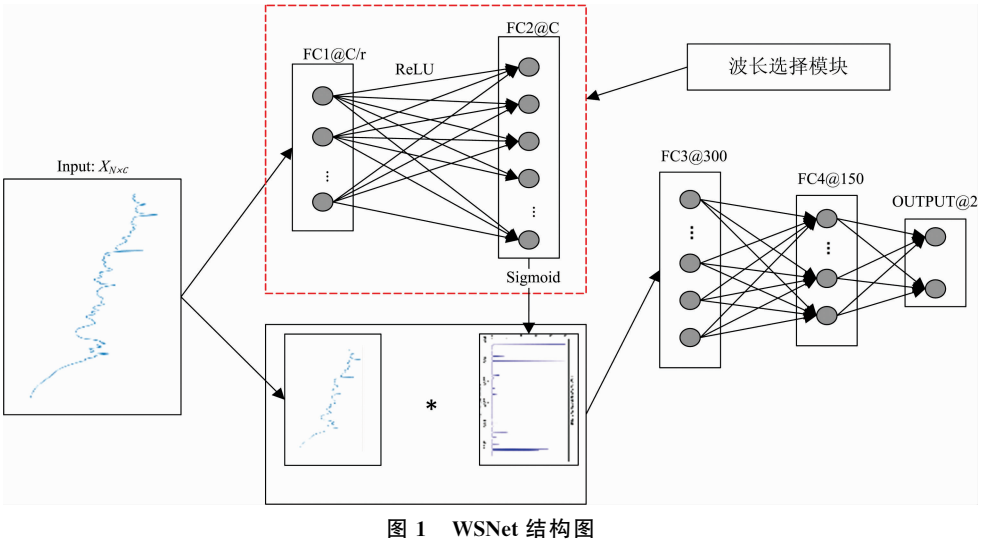

图 1 WSNet 结构图

Fig. 1 WSNet structure chart

FC2 层利用降维后的数据学习出各个波长点对建模的贡献权重(s)

$$\begin{cases} s = A_{\mathrm{FC}}(W_{\mathrm{FC}}x_{\mathrm{FC}} + b_{\mathrm{FC2}}) \\ A_{\mathrm{FC2}}(x) = \mathrm{sigmoid}(x) = \dfrac{1}{1+\mathrm{e}^{-x}} \end{cases} \tag{2}$$

W_{FC} 为 FC2 层的权值矩阵，b_{FC2} 为 FC2 层的偏置向量，$A_{\mathrm{FC2}}(\cdot)$ 为 FC2 层的激活函数。FC2 层神经元个数与输入的波长点个数相同均为 C。FC2 层学习到的各个波长点的权重相差较小，不容易筛选。考虑到光谱中对建模有贡献的波长点往往较少，因此对权重增加稀疏约束。因此构造如下的损失函数 $J(s)$

$$J(s) = \mathrm{Loss}(\cdot) + \lambda \| s \|_1 \tag{3}$$

Loss(·)是原损失函数，λ 是控制稀疏程度的参数。

WSNet 网络选择具有以下优点：

(1)对学得的波长点权重增加稀疏约束，更容易选择与校正模型相关的波长点用于建模：通过网络学习到的各个波长点的权重差异较小，不容易选取需要的波长点，而且在拉曼光谱中对建模有贡献的波长点在全光谱中是稀疏的，因此对权重增加了稀疏约束，极大的减少了用于选择的波长点，更容易确定筛选的阈值，选出与建模相关的波长点。

(2)波长选择模块计算权值的方式与注意力机制[10]的功能相吻合，能计算出各个波长点对校正模型的贡献：注意力机制可以通过扫描全局获取重点关注的目标区域，在目标区域投入更多的资源，获取关注目标的更多细节。波长选择网络利用注意力机制，能够获取到对模型有贡献的波长点，能有效的提高模型的性能。

2 实验部分

2.1 光谱采集

实验选用人、比格犬、新西兰兔三种动物的血清作为实验样本，其中包括采集于重庆市西南医院的人血清样本 110 例和采集于重庆市中药研究院的动物血清样本 216 例(其中犬 116 例、兔 100 例)，共计 326 个样本，人与动物血清样本

清单如表 1 所示。

表 1　血清样本清单

Table 1　Serum sample list

血清样本	来源	取样方式	样本数
人血清	重庆西南医院	无菌取样	110
比格犬血清	重庆中医研究院	无菌取样	116
新西兰兔血清	重庆中医研究院	无菌取样	100

采集的血清样本均通过取沉淀后的血液样本的上层清液获得，采用乙二胺乙酸二钠(EDTA)抗凝管盛放。每次实验均采用移液管取 2 μL 血清样本置于石英载玻片上，然后利用光谱仪进行测量，为了减少实验误差，每个样本都要测量 4 次，取平均值作为最终测量值。

实验仪器主要包括计算机、光谱仪、石英载玻片、移液管、抗凝管等。光谱仪是采用小型拉曼光谱仪海洋 IDR-Micro-785，光谱的激发波长为 785 mm，分辨率 4 cm^{-1}，扫描范围为 200～2 000 cm^{-1}，激发功率为 18.8 mW，曝光时间为 5 s。

虽然人、犬、兔的血液中都包含了蛋白质、葡萄糖、色氨酸和血红素等物质[10]，使得这三种动物血清的拉曼光谱具有相似的谱峰(如图 2 所示)，但是不同物种的遗传基因不同，导致了血清中的某些成分的浓度和化学特性是不同的[12]，使其在对应的谱峰上存在微弱的差异，可以利用这些差异对人和动物的血清进行区分。

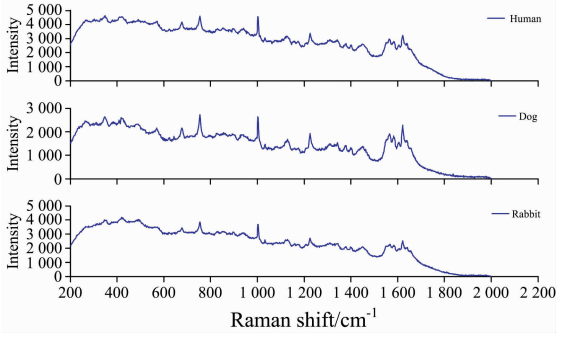

图 2　测得的人、比格犬、新西兰兔的拉曼光谱

Fig. 2　Raman spectra of human, beagle and New Zealand rabbit were measured

2.2　方法

原始光谱在 200～630 cm^{-1} 波段受石英载玻片的影响，在 1 710 cm^{-1} 波段之后无明显的波峰，所以实验选择了 598～1 718 cm^{-1} 波段共 800 个波长点进行分析。从图 2 中可以看出样本之间的光强差异很大，为了消除样本之间的光强差异，每条光谱都除以该光谱中最大的值，将光强归一化到 0～1 之间；然后利用移动平均平滑法去除光谱中的噪声，滑动窗口大小为 9。

波长选择模块中输入的光谱样本波长点个数 $C=800$，降维倍数 $r=5$，所以第一个全连接层(FC1 层)的神经元个数为 160 个，第二个全连接层(FC2 层)的神经元个数与输入的

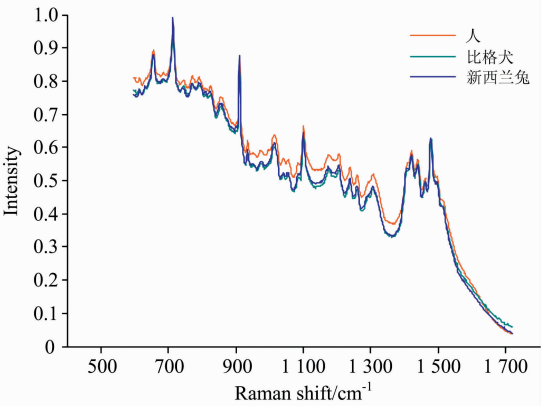

图 3　人、犬、兔血清的拉曼光谱归一化后的对比

Fig. 3　Raman spectra of human serum, dog serum and rabbit serum after normalization

波长点个数相同均为 $C=800$，神经网络校正模型中的两个全连接层(FC3 和 FC4 层)的神经元个数分别为 300 个和 150 个，OUTPUT 层有两个神经元，为二分类。本实验所采用的目标函数是交叉熵，计算公式如式(4)

$$J(s)=-\frac{1}{N}\sum_{i=1}^{N}[y_i\ln\hat{y}_i+(1-y_i)\ln(1-\hat{y}_i)]+\lambda\|s\|_1 \tag{4}$$

式(4)中，y_i 为第 i 个样本的测量值，$\hat{y}_i$ 为第 i 个样本的预测值，N 为输入的样本数量。利用随机优化方法自适应矩估计算法(Adam 算法)进行求解，网络的学习率设置为 0.001，最大的迭代次数为 100 次。WSNet 训练过程中的损失值和分类准确率如图 4 所示，可以很明显的看出，随着迭代次数的增加，损失函数逐渐收敛，分类准确率逐渐趋近于 1。本文将权重的均值作为阈值，计算公式如式(5)

$$\bar{s}=\frac{1}{N\times C}\sum_{i=1}^{N}\sum_{j=1}^{C}s_{ij} \tag{5}$$

式(5)中，s_{ij} 第 i 个样本的第 j 个波长点处的权重。

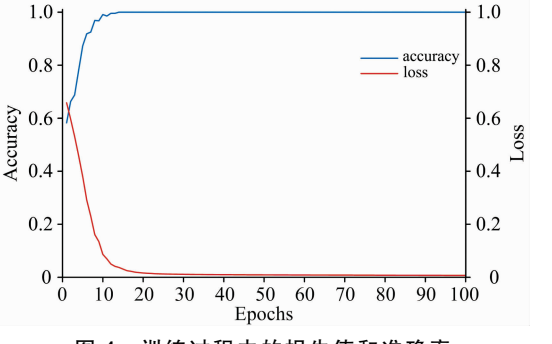

图 4　训练过程中的损失值和准确率

Fig. 4　Loss value and accuracy during model training

降维倍数 r 是网络结构中非常重要的一个参数，r 过大或过小都不能使校正模型达到最佳性能，如图 5 所示，当 $r=5$ 时校正模型的分类准确率最高，在 $r<5$ 时，校正模型在校正集和测试集的的分类准确率都较低；在 $r>5$ 时，校正集上的分类准确率基本不变，但是测试集上的分类准确率略有下

降，因此本文选取的降维倍数为 5。

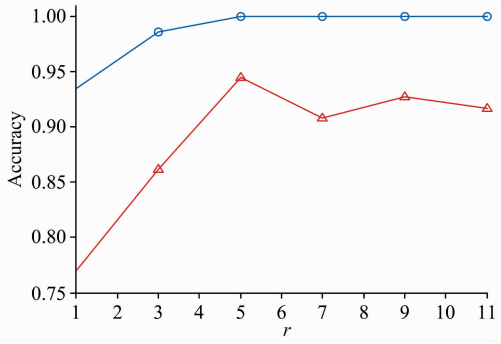

图 5 不同的降维倍数 r 在测试集和校正集上的分类准确率

Fig. 5 Classification accuracy of different dimensionality reductions on test sets and calibration sets

3 结果与讨论

在实验的过程中，将 326 例光谱样本按比例取出 1/3(共计 109 例，37 例人、39 例犬、33 例兔)作为测试集对校正模型的性能进行测试，余下的 2/3(共计 217 例，73 例人、77 例犬、67 例兔)作为校正集用于校正模型的训练。将利用 WSNet 得到的光谱与全光谱数据分别使用神经网络(neural network，NN)和 PLS-DA 建模进行对比，同时将 WSNet 与经典的波长选择算法 UVE(uninformative variables elimination)进行对比，并使用分类准确率、ROC 曲线(receiver operating characteristic curve)和 AUC(area under curve)对这些校正模型的性能进行评价。ROC 曲线的横坐标为假正率(FPR)，纵坐标为真正率(TPR)，TPR 和 FPR 的计算公式如式(6)

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{TN + FP} \tag{6}$$

式(6)中，TP 为真正例，FP 为假正例，FN 为假反例，TN 为真反例。ROC 曲线所包围的面积称之为 AUC，AUC 的值越大，证明模型的分类精度越高。AUC 的计算公式如式(7)

$$AUC = \sum_{i=1}^{N-1}(TPR_{i+1} + TPR_i)(FPR_{i+1} - FPR_i) \tag{7}$$

式(7)中，N 为样本的数量。

图 6(a)给出了 WSNet 学得的光谱权重，只有 109 个波长点的权重不为零。利用式(5)计算出一个阈值，将权重小

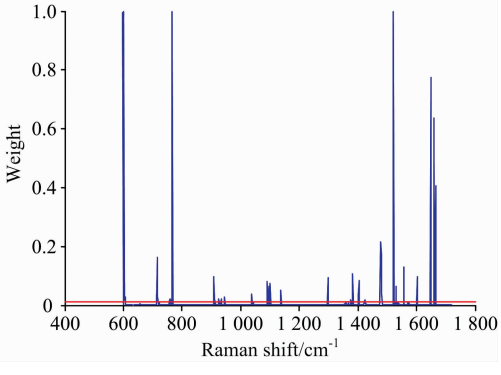

图 6(a) 光谱权重

Fig. 6(a) Spectral wavelength point weight

于阈值的波长点也去掉。最终，WSNet 选择了 42 个波长点用于建模，如图 6(b)所示。

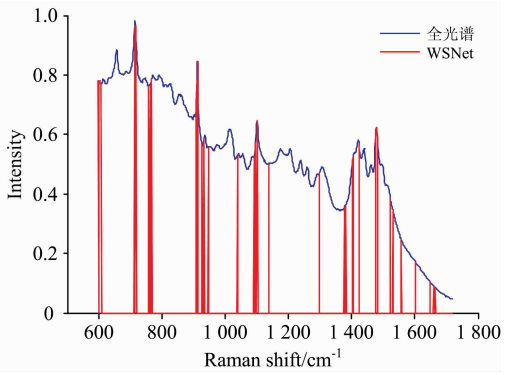

图 6(b) 拉曼光谱波长选择前后谱线对比图

Fig. 6(b) Raman spectrum wavelength selection before and after comparison chart

将波长选择前后的光谱利用 PLS-DA 和 NN 进行建模，实验结果表明，四种校正模型在校正集上都取得了非常好的效果，如表 2 所示，NN 和 WSNet-NN 均未错分，分类精度都达到了 100%，PLS-DA 和 WSNet-PLS-DA 分别错分了 2 例和 1 例，分类精度也分别达到了 99.539%和 99.078%。在测试集上，WSNet-NN 分类精度优于其他三种模型，仅错分了 6 例，分类精度最高为 94.495%；PLS-DA 错分了 10 例，分类精度为 90.826%；NN 和 WSNet-PLS-DA 均错分了 8 例，分类精度相同为 92.661%。从上述的实验结果可以看出，利用 WSNet 对光谱进行波长选择有利于提升校正模型的性能。

表 2 分类结果对比

Table 2 Comparison of classification results

建模方法	校正集准确率/%	测试集准确率/%
PLS-DA	98.156	90.826
NN	100.000	92.661
WSNet-NN	100.000	94.495
WSNet-PLS-DA	99.078	92.661

PLS-DA，WSNet-PLS-DA，NN 和 WSNet-NN 在测试集上的 ROC 曲线如图 7 所示。通过式(7)计算可以得到这四种校正模型的 AUC 值，PLS-DA 和 WSNet-PLS-DA 的 AUC 值分别为 0.976 0 和 0.932 8，NN 和 WSNet-NN 的 AUC 值分别为 0.985 0 和 0.970 3，从这四种校正模型的 AUC 值可以看出，利用波长选择后的光谱进行建模能提升校正模型的性能，让其具有更优的鉴别能力。

分别使用 WSNet 和 UVE 对光谱进行波长选择，如图 8 所示，WSNet 选择的波长点个数远少于 UVE。在稀疏约束的作用之下，WSNet 去掉了绝大部分的波长点，只选择了 42 个波长点，而 UVE 却选择了 486 个波长点。将 WSNet 和 UVE 波长选择后的光谱分别用于建立 PLS-DA 校正模型，分类结果如表 3 所示，WSNet-PLS-DA 在校正集和测试集上

的分类准确率分别为 99.078% 和 92.661%，UVE-PLS-DA 在校正集和测试集上的分类准确率分别为 98.156% 和 91.743%。WSNet-PLS-DA 在校正集和测试集的分类准确率相比于 UVE-PLS-DA 都要高一些。

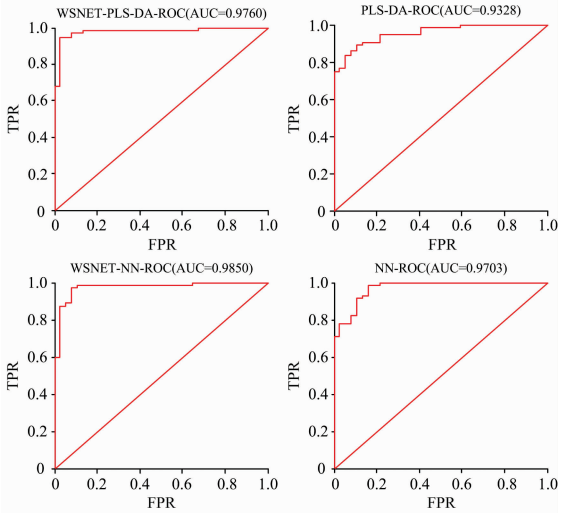

图 7　PLS-DA 和 NN 血液鉴别的 ROC 曲线

Fig. 7　ROC curve of PLS-DA and NN classification

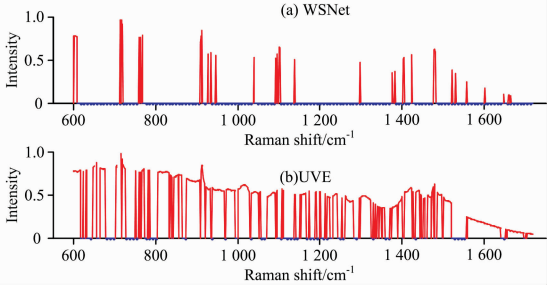

图 8　(a)Raman-WSNet 波长选择后的光谱；(b)UVE 波长选择后的光谱

红色部分是保留的波长点

Fig. 8　(a) the spectrum after Raman-WSNet wavelength selection and (b) the spectrum after UVE wavelength selection

The red part is reserved wavelength point

表 3　Raman-WSNet-PLS-DA 和 UVE-PLS-DA 分类结果对比

Table 3　Comparison of Raman-WSNet-PLS-DA and UVE-PLS-DA classification results

建模方法	校正集分类结果/%	测试集分类结果/%
WSNet-PLS-DA	99.078	92.661
UVE-PLS-DA	98.156	91.743

WSNet-PLS-DA 和 UVE-PLS-DA 在测试集上的 ROC 曲线如图 9 所示。根据 AUC 的计算公式(7)计算可以得到各自的 AUC 值分别为 0.976 0 和 0.965 8。从中可以看出，相比于 UVE，WSNet 不仅能去掉更多的波长点，而且对校正模型的性能提升也更为显著。

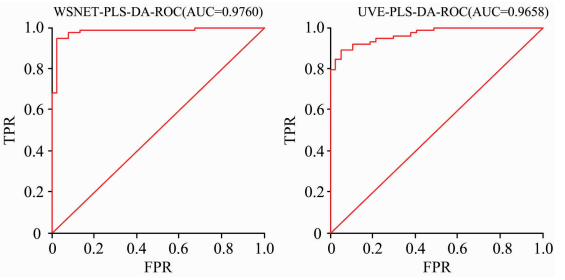

图 9　Raman-WSNet-PLS-DA 和 UVE-PLS-DA ROC 曲线对比

Fig. 9　Comparison of Raman-WSNet-PLS-DA and UVE-PLS-DA ROC curves

拉曼光谱的各个波长点之间存在严重的多重共线性，采用全拉曼光谱数据用于 NN 和 PLS-DA 模型的构建，得到的校正模型性能不够好。利用 WSNet 对拉曼光谱进行处理，去除了光谱中对建模有影响和无贡献的波长点，对于校正模型的性能有显著的提升。WSNet 与经典的波长选择算法 UVE 相比也具有明显的优势，UVE 算法的核心思想是在原光谱中引入人工随机变量，将随机变量和光谱一起进行建模，将原光谱中各个波长点的回归系数与随机变量回归系数的平均值进行比较，如果原光谱波长点的回归系数低于人工随机变量回归系数的平均值，就认为该变量是噪声变量，应该被剔除。由于 UVE 算法引入随机变量，导致该算法存在一些随机因素，使得每次筛选的结果都不相同，但都表现出很强的一致性。UVE 在剔除波长点的过程中参考的是人工随机变量，即随机噪声，所以剔除的仅仅是影响建模的噪声波长点，对建模既没有贡献也没有影响的波长点并不会剔除。WSNet 相比于 UVE 就要“激进”许多。WSNet 由于是通过神经网络进行波长选择的，同样也存在一些随机因素，WSNet 是通过神经网络自学习能力自适应得到每个波长点的权重，从而利用学得的权重对波长点进行筛选。同时在权重学习的过程中加入了稀疏约束，所以学得权重非常稀疏，利用稀疏的权重对波长点进行筛选会变得更加容易。通过上面的实验可以看出，UVE 所选择的波长点比较多，它并没有筛选掉太多的波长点，是一种比较“保守”的波长选择算法。相比之下，WSNet 所选择的波长点较少，而且建立的校正模型性能更好，证明了 WSNet 是有效的。

4　结　论

利用神经网络的学习能力提出了一种基于神经网络的拉曼光谱稀疏波长选择的新方法——WSNet。它能自适应的选择拉曼光谱中包含信息的波长点，从而建立更加高效的校正模型。将波长选择前后的拉曼光谱利用同一种方法建模进行对比，进一步验证了该方法有利于校正模型分类精度的提升，最高达到了 94.495%。同时还与 UVE 波长选择算法进行了对比，实验结果表明 WSNet 在波长点选择的个数和校正模型的分类精度方面都优于 UVE 算法，精度达到了 92.661%。

References

[1] Doty K C, Lednev I K. Forensic Science International, 2018, 282: 204.
[2] Bian H, Wang P, Wang N, et al. Biomedical Optics Express, 2018, 9(8): 3512.
[3] Martens H, Naes T, Naes T. Multivariate Calibration. John Wiley & Sons, 1992.
[4] LI Pao, ZHOU Jun, JIANG Li-wen, et al(李 跑, 周 骏, 蒋立文, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(5): 1428.
[5] Duarte L M, Paschoal D, Izumi C M S, et al. Food Research International, 2017, 99: 106.
[6] Wang S H, Zhao Y, Hu, R, et al. Chinese Journal of Analytical Chemistry, 2019, 47(4): e19034.
[7] Chen J, Yang C, Zhu H, et al. Chemometrics and Intelligent Laboratory Systems, 2018, 182: 188.
[8] Centner V, Massart D L, de Noord O E, et al. Analytical Chemistry, 1996, 68(21): 3851.
[9] Bin J, Ai F, Fan W, et al. Chemometrics and Intelligent Laboratory Systems, 2016, 158: 1.
[10] Choi H, Cho K, Bengio Y. Neurocomputing, 2018, 284: 171.
[11] Premasiri W R, Lee J C, Ziegler L D. The Journal of Physical Chemistry B, 2012, 116(31): 9376.
[12] Araújo M C U, Saldanha T C B, Galvao R K H, et al. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65.

Raman Spectrum Wavelength Selection Method Based on Neural Network

SHEN Dong-xu, HONG Ming-jian*, DONG Jia-lin

School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China

Abstract Blood identification is crucial for the field of inspection and quarantine, criminal investigation and animal protection. Traditional blood identification methods have shortcomings such as long analysis period and damaging to blood samples during the identification process. Raman spectroscopy can obtain molecular vibration and rotation information by analyzing the scattering spectrum different from the incident light frequency, and obtain the composition of the material. Moreover, it has the characteristics of zero pollution and non-contact, which provides the possibility of non-destructive identification of blood. However, there is serious multicollinearity between each wavelength point in Raman spectrum, and directing the use of full-spectrum for modeling will increase the complexity of the model and reduce the stability of the model. According to the characteristics of Raman spectroscopy, this paper proposes a wavelength selection based on neural network. The method uses the neural network to learn the contribution weight of each wavelength point to the correction model, and uses the mean value of the weight as the threshold value to remove the wavelength point whose weight is lower than the threshold value, so as to achieve the purpose of wavelength selection. In order to make it easier to determine the threshold of the screening, sparse constraints are added to the weight learning process, which greatly reduces the wavelength points used for screening. The proposed method was validated by Raman spectroscopy datasets of animal and human serum. The experimental results show that the model established by the wavelength selection using this method has a certain improvement in classification accuracy compared and AUC value with the full spectrum, the accuracy of artificial neural network (NN) reached 94.495% and AUC value reached 0.985 0. The accuracy of PLS-DA reached 92.661%, and AUC value reached 0.976 0. Compared with the traditional wavelength selection method UVE, the method selects fewer wavelength points, and only 42 wavelength points have been selected for modeling, the classification accuracy of the calibration model and AUC value is high, and accuracy reached 92.661%, and AUC value reached 0.976 0. It is proved that the wavelength selection method can effectively screen out the wavelength points contributing to the modeling, which provide a possibility for non-destructive identification of blood, which has certain practical value.

Keywords Spectral analysis; Raman spectrum; Wavelength selection; Neural network; Calibration model

* Corresponding author

(Received Sep. 25, 2019; accepted Jan. 6, 2020)