基于拉曼光谱特征的生物组织识别方法

郑家文，杨唐文

北京交通大学信息科学研究所，北京 100044

摘要 利用生物分子独特的拉曼光谱特征进行生物组织分类研究。利用研制的拉曼探针穿刺生物组织，获得生物分子的拉曼光谱信号数据，并对数据进行基线校正和滤波预处理；利用主成分分析法，提取拉曼光谱数据的关键特征；通过反向传播(BP)神经网络算法对这些特征进行组织分类；利用动物组织样品上采集的拉曼光谱数据，进行自动分类实验研究。结果表明，BP 神经网络能够实现不同生物组织的分类，且准确率达到 95%。

关键词 光谱学；拉曼光谱；生物组织；主成分分析法；BP 神经网络

中图分类号 O657.3 **文献标识码** A

doi：10.3788/LOP54.053001

Classification Method of Biological Tissues Based on Raman Spectrum Features

Zheng Jiawen, Yang Tangwen

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, *China*

Abstract Biological tissues are identified based on unique features of their Raman spectra. Raman signal data of biological tissues is acquired by a self-designed Raman probe, and preprocessed to rectify the baseline through filtering noises and stray light. The principal component analysis method is used to extract the critical Raman signal features of biological tissues, and then a back propagation (BP) neural network algorithm is used to classify the biological tissues by using these features. Automatic classification is implemented with the Raman spectrum data from the animal tissue phantoms. Experimental results show that the BP neural network is efficient to identify different animal tissues, and the accuracy rate reaches nearly 95%.

Key words spectroscopy; Raman spectroscopy; biological tissues; principal component analysis; back propagation neural network

OCIS codes 300.6450; 300.6390; 170.6935

1 引 言

拉曼光谱检测技术是近几年快速发展起来的一种基于拉曼散射效应进行分子结构和化学键分析的方法。不同生物分子表现出变化的拉曼频移、强度等特征信息，提供从分子层面区分生物组织的证据。目前，利用拉曼光谱数据进行病变和非病变组织的判定及其边界划分已经成为一个研究热点[1-3]。拉曼光谱技术在生物领域的研究主要集中在组织结构分析或其成分鉴定(比如脂类、蛋白质、糖类、水等)以及类型判别[4-6]。与其他检测手段相比，拉曼光谱对水分子不敏感，对生物组织检测效果更好[7-8]。

然而，生物组织拉曼光谱数据特征维数高，信息量大，若不进行特征提取，难以找出生物组织类型的关键特征[9-10]。主成分分析法是一种基于统计学的数据挖掘技术，在数据特征提取中广泛应用。该方法通过降

收稿日期：2016-12-19；**收到修改稿日期**：2017-01-04

基金项目：国家自然科学基金(61375109，61273356)、机器人学国家重点实验室基金(2013-O12)

作者简介：郑家文(1991—)，男，硕士研究生，主要从事拉曼光谱学、模式识别方面的研究。

E-mail：14120374@bjtu.edu.cn

导师简介：杨唐文(1971—)，男，博士，副教授，主要从事信号与信息处理、机器人学方面的研究。

E-mail：twyang@bjtu.edu.cn(通信联系人)

维将多个具有相关性的变量转化为少数几个相互无关独立的主成分变量。这些主成分在不同波数上的载荷系数反映了拉曼光谱的主要特征[10-12]。提取特征之后,则可进行样本分类,实现生物组织类型鉴定。生物组织中的蛋白质、脂肪、糖类、无机盐和水等的比例随生物种类变化,生物分子大小也不同。因此,利用提取的拉曼光谱数据特征是可以识别不同生物组织的,但数据特征太多导致生物成分检测和组织识别极其复杂。常用的分类方法有线性判别分析[13-14]、支持向量机[15-16]和神经网络[17-18]等算法,其中,神经网络算法作为一种强大的分类学习方法,能够实现输入与输出之间的高度非线性映射。反向传播(BP)神经网络学习速度快,效率高,适应大样本数据分类,适用于本文研究。

为了更方便地采集体内器官组织的拉曼信号,进行活体原位检测,本文研究中将拉曼探头做成探针形式,刺入组织内部以更好地保证组织固有生理特性,并利用八种动物组织样本采集的拉曼光谱数据,进行生物组织的拉曼光谱分类方法研究。

2 数据采集与预处理

2.1 采集设备

生物组织拉曼光谱数据采集系统主要由光谱仪(波长范围 785~1100 nm,分辨率 0.035~0.6 nm)、激光源(波长 785 nm,功率范围 0~650 mW)、探头(波长 785 nm,拉曼频移范围 175~4000 cm^{-1},焦距7.5 mm)、探针(长度 10 cm,外径 2.8 mm)和计算机组成。激光源产生波长为 785 nm 的激光,经探针的激发光路传递至生物组织,照射在生物分子上,引发光子与生物分子碰撞,发生散射,其中部分散射光(拉曼散射光)经收集光路传递回光谱仪,进行信号转换后,生成的数据被计算机读取。

2.2 采集实验

为了实现对不同生物组织的分类和鉴定,需要采集大量的生物组织拉曼光谱数据样本。考虑到动物组织中的蛋白质、脂肪、无机盐等成分和比例是不同的,比如蛋白质,它是一类由 C、H、O、N、S、P 等元素组成的含氮生物高分子,分子量大,结构复杂,没有固定的化学式,因而常被用作实验素材。选用猪、牛、羊、鸡、鸭等动物组织样本进行拉曼数据采集和鉴定实验。采集过程中使用拉曼探针穿刺到生物组织内部,收集拉曼散射信号,避免样本在外部检测受到环境污染。在多种动物组织样本中不同的位置采集数据,获得了 215 组拉曼光谱数据。

2.3 数据预处理

拉曼光谱检测受到荧光背景的干扰,同时由于本身稳定性的限制(激光、光谱仪),会产生背景噪声,出现基线漂移,对拉曼光谱分析结果会产生很大的影响。所以,在进一步对数据处理分析前,有必要对采集到的拉曼光谱数据进行基线校正及平滑处理,消除基线漂移和噪声带来的影响。在光谱软件中利用自适应迭代惩罚最小二乘方法对采集的拉曼光谱数据进行基线校正,并使用 Savitzky-Golay 滤波函数对数据进行噪声去除。图 1 所示为八组样本拉曼光谱数据预处理前后的结果。可以看出,经过预处理后的拉曼光谱图强度峰值特征获得显著增强。同时可以发现这些动物组织的拉曼光谱图上特征差异并不明显,差异较小,肉眼几乎难以分辨。事实上,在样本数据量很大时,同类样本个体也呈现些微差异。因此采用 BP 神经网络算法对主成分分析提取的样本数据特征进行分类和生物组织鉴定研究。

3 拉曼光谱数据特征提取

生物组织的拉曼光谱特征具有独特性,是组织分析和鉴定的重要依据。通常,生物组织的特征拉曼峰集中在 800~3200 cm^{-1} 之间,这些拉曼峰值差异(比如强度、频移和宽度等)很难通过指认特征拉曼峰化学键的归属进行分析和鉴定。采用主成分分析法先提取采集的拉曼光谱数据的特征,再进行鉴定分析。

主成分分析法的基本原理如下。设 $x_1, x_2, \cdots, x_n$ 为原变量,$y_1, y_2, \cdots, y_k$($k \leqslant n$)为新变量,每一个新变量是 n 个原始变量的线性组合。新变量 $y_1, y_2, \cdots, y_k$ 分别为原变量 $x_1, x_2, \cdots, x_n$ 的第 1,第 2,…,第 k 主成分。其核心思想是通过将原变量线性组合为新变量实现对原始数据特征的降维。所以,主成分既要包含原始数据的丰富信息,又能够更直观地反映出原始数据的特征,且所含的信息互不重叠。

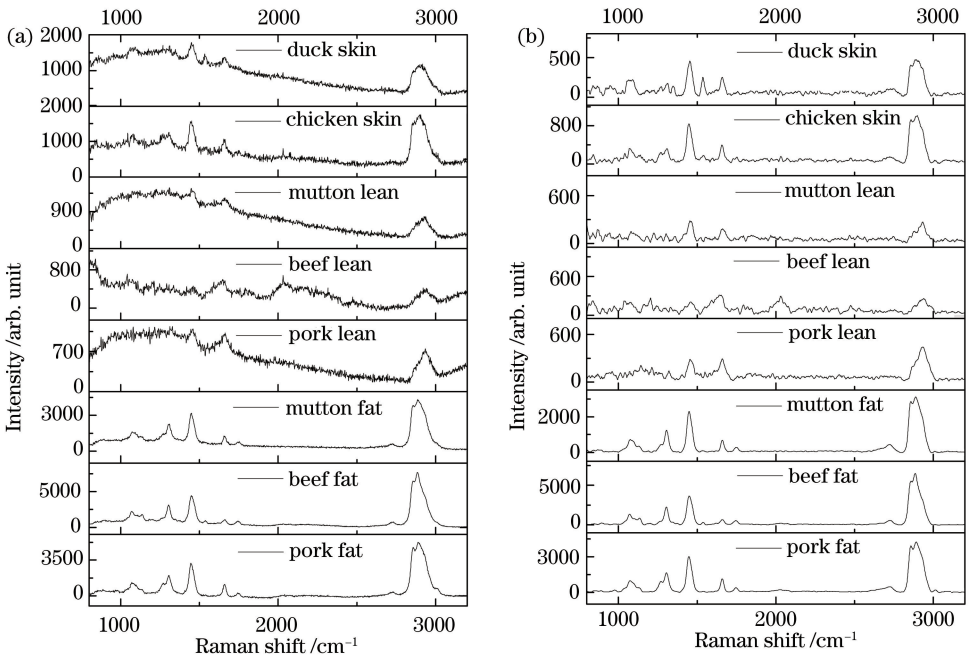

图 1 生物组织拉曼光谱图。(a)预处理前;(b)预处理后

Fig. 1 Raman spectra of biological tissues. (a) Before processing; (b) after processing

对于含有 n 个变量的数据,一般不会取 n 个主成分,而是取其中的几个,使这几个主成分的累积贡献率达到 85%~95%。一般来说,第一个主成分的方差值最大,即它所含信息量在所有数据中起主导作用,主成分越靠后,其方差贡献率越小,重要性也越小。对获得的 150 组动物组织拉曼光谱数据进行主成分分析,得到拉曼光谱数据主成分。图 2 所示为其中前 10 个主成分的贡献率,发现第一主成分 PCA1 的贡献率达到了 97.56%。也就是说,第一主成分几乎包含了所有的拉曼光谱信息。因此,利用该主成分提取拉曼光谱数据特征。

图 3 所示为第一主成分 PCA1 在整个拉曼频移范围的载荷图。横坐标为拉曼频移,纵坐标表示各拉曼频移变量对于主成分 PCA1 的载荷值,即各个拉曼频移变量与主成分 PCA1 的相关性。当设定载荷阈值 Y 时,可以提取高度在阈值 Y 以上的峰值所对应的拉曼频移变量。这些拉曼频移与主成分 PCA1 的相关性较强,即被定义为拉曼特征频移,阈值 Y 决定了特征频移的数量。

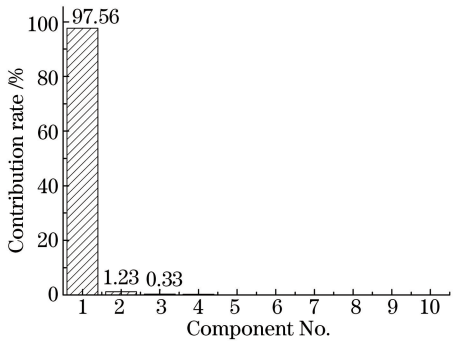

图 2 主成分贡献率分布图

Fig. 2 Contribution rate of the first ten principal components

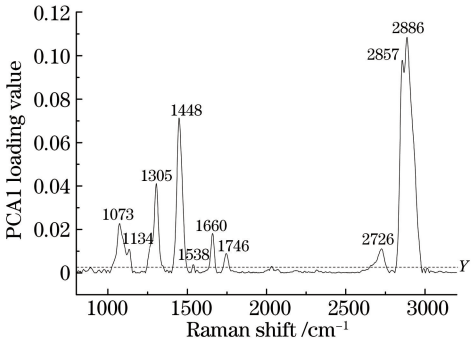

图 3 主成分 PCA1 载荷系数

Fig. 3 Loading factor of PCA1

4 拉曼光谱特征分类

BP 神经网络是一种较为流行的多层前馈神经网络,主要基于信号正向传递以及误差逆向传播原理进行网络模型训练。在信号正向传递过程中,输入层接受样本输入数据,通过隐含层处理后,经过输出层输出。如果预测输出与实际输出之间的误差达不到所设定的目标误差时,则转入误差逆向传播过程,根据误差调整

网络的权值与阈值,进而使预测输出不断逼近实际输出。BP 神经网络具有自组织、自学习、强大的容错能力以及分类能力等特点,更适合多类多样本的分类应用。

数据经过主成分分析后,提取到拉曼光谱的特征频移,作为下一步进行神经网络生物组织分类的输入数据。

将特征频移对应的拉曼强度值进行归一化处理,作为网络输入,输出则为动物组织样本的类别。共有八种类别,分别用数字 0~7 表示猪肉、牛肉和羊肉中的肥肉和瘦肉,及鸡肉和鸭肉表皮。所以需要设置四个输出节点,编码[0 0 0 0]代表类别 0,[0 0 0 1]代表类别 1,…,依次类推到[0 1 1 1]代表类别 7。此外,考虑到神经网络学习率设置偏小,虽容易保证网络收敛,但是训练学习收敛较慢。相反,设置偏大则有可能使网络训练不收敛,影响最终识别效果。为保证系统的稳定性,设置该值为 0.01。神经网络的隐含层节点数选择根据为

$$t < \sqrt{(t_1 + t_2)} + \alpha, \tag{1}$$

式中 t 为隐含层的节点数,t_1 为输入层节点数,t_2 为输出层节点数,α 为 0~10 之间的常数。结合(1)式,多次调整隐含层节点数,发现隐含层节点的个数对于识别率的影响并不大,隐含层节点个数为 12。

改变载荷阈值 Y,提取的特征拉曼频移数量会发生较大变化。针对不同载荷阈值下提取的特征频移训练一个 BP 神经网络,训练数据有 150 组,建立生物组织分类模型,在此基础上,对 65 组拉曼光谱数据进行生物组织分类测试。表 1 列出了分类结果,并从分类准确率、均方误差(MSE)及运行时间三个方面比较特征变量数对单隐层和双隐层 BP 神经网络性能的影响。训练迭代设定为 10000 次,目标误差设为 0.001。

表 1 BP 神经网络不同特征输入下的分类结果

Table 1 Classification results with various feature inputs to BP neural networks

Y	Feature number	BP neural network with single hidden layer				BP neural network with double hidden layers			
		Accuracy[1] /%	Accuracy[2] /%	MSE	Time /s	Accuracy[1] /%	Accuracy[2] /%	MSE	Time /s
0.0010	41	100.000	90.7692	0.0010	24	100.000	90.7692	0.0010	13
0.0020	17	100.000	89.2308	0.0010	51	100.000	95.3846	0.0010	17
0.0025	14	100.000	89.2308	0.0010	55	100.000	84.6154	0.0010	18
0.0030	12	100.000	73.8462	0.0029	133	100.000	83.0769	0.0010	26
0.0040	9	100.000	67.6923	0.0027	133	100.000	75.3846	0.0010	40
0.0100	8	100.000	72.3077	0.0027	133	100.000	70.7692	0.0010	37
0.0150	6	99.333	67.6923	0.0067	136	100.000	72.3077	0.0010	51
0.0200	5	91.333	66.1538	0.0295	139	97.333	67.6923	0.0067	157
0.0300	4	85.333	66.1538	0.0378	133	92.000	63.0769	0.0172	156
0.0500	3	67.333	49.2308	0.0671	133	72.000	44.6154	0.0599	152

[1] For the training data, [2] for the test data

从表 1 可以看出,对单隐层神经网络而言,特征变量数在 8 个以上时,对训练数据的分类准确率能够达到 100%,而对于测试数据,网络的分类准确率随着特征变量数减少而逐渐降低,这说明了 BP 神经网络逐渐进入过学习状态。当特征变量数在 8 个以下时,神经网络对训练和测试数据的分类准确率均呈下降趋势,这说明特征变量减少到一定程度时,神经网络的分类精度降低。另外,随着特征变量个数的减少,网络训练需要的时间反而更长,且均方误差变大。对于双隐层神经网络,结果整体上与单隐层类似。不同的是,当达到目标误差时,双隐层神经网络的训练时间小于单隐层;未达到目标误差时,训练时间却大于单隐层神经网络。而且与单隐层相比,双隐层 BP 神经网络分类准确率更高。

考虑到双隐层神经网络的泛化能力和预测精度优势,选用一个双隐层的 BP 神经网络进行分类研究。根据前面的分析,随着阈值 Y 的增大,主成分分析提取的特征频移数减少,分类准确率降低。阈值 Y 取 0.002,得到 17 个特征频移,由此确定了一个 17-12-12-4 的双隐层 BP 神经网络结构。从表 1 可以看到该神经网络的分类准确率达到 95%。

5 结 论

利用主成分分析和 BP 神经网络算法对生物组织样本的拉曼光谱数据进行特征提取和分类研究。先利

用研制的拉曼探针获取动物组织样本的大量拉曼光谱数据,并进行基线校正和滤波预处理。然后,通过主成分分析法提取拉曼光谱数据特征,分别作为训练和测试数据输入到一个 BP 神经网络进行组织分类。主成分分析法的载荷阈值选取很大程度上决定了特征拉曼频移的数量,同时也确定了 BP 神经网络的输入节点数。自动分类实验结果表明所提出的方法能够有效实现生物组织的分类,准确率达到 95%。相比于传统复杂的生物化学分析,本文提出的方法能快速可靠地完成生物组织在体原位鉴定,具有潜在的应用价值。

参 考 文 献

[1] Iping Petterson I E, Day J C C, Fullwood L M, *et al*. Characterisation of a fibre optic Raman probe within a hypodermic needle[J]. Analytical and Bioanalytical Chemistry, 2015, 407(27): 8311-8320.

[2] Knipfer C, Motz J, Adler W, *et al*. Raman difference spectroscopy: a non-invasive method for identification of oral squamous cell carcinoma[J]. Biomedical Optics Express, 2014, 5(9): 3252-3265.

[3] Chen Rong, Li Yongzeng, Feng Shangyuan, *et al*. Advances in Raman spectroscopy for human tissue[J]. Laser & Optoelectronics Progress, 2008, 45(1): 16-23.
陈 荣, 李永增, 冯尚源, 等. 人体组织拉曼光谱研究进展[J]. 激光与光电子学进展, 2008, 45(1): 16-23.

[4] Tintchev F, Wackerbarth H, Kuhlmann U, *et al*. Molecular effects of high-pressure processing on food studied by resonance Raman[J]. Annals of the New York Academy of Sciences, 2010, 1189(1): 34-42.

[5] Xu Bin, Lin Manman, Yao Huilu, *et al*. Measurement of hemoglobin concentration of single red blood cell using Raman spectroscopy[J]. Chinese J Lasers, 2016, 43(1): 0115003.
徐 斌, 林漫漫, 姚辉璐, 等. 拉曼光谱技术测量单个红细胞的血红蛋白浓度[J]. 中国激光, 2016, 43(1): 0115003.

[6] Qin Zhaojun, Tao Zhanhua, Liao Wei, *et al*. Raman spectral profiles of PHB synthesis influenced by different nitrogen sources[J]. Acta Optica Sinica, 2016, 36(4): 0417001.
覃赵军, 陶站华, 廖 威, 等. 氮源影响 PHB 合成代谢的拉曼光谱分析[J]. 光学学报, 2016, 36(4): 0417001.

[7] Depciuch J, Kaznowska E, Zawlik I, *et al*. Application of Raman spectroscopy and infrared spectroscopy in the identification of breast cancer[J]. Applied Spectroscopy, 2016, 70(2): 251-263.

[8] Xiong Xingchuang, Fang Xiang, Ouyang Zheng, *et al*. Artificial neural networks method of classification and identification for mass spectrometry imaging data of biological tissue[J]. Chinese Journal of Analytical Chemistry, 2012, 40(1): 43-49.
熊行创, 方 向, 欧阳证, 等. 基于人工神经网络的生物组织质谱成像分类与识别方法[J]. 分析化学, 2012, 40(1): 43-49.

[9] Ye Z, Auner G. Principal component analysis approach for biomedical sample identification[C]. IEEE International Conference on Systems, Man and Cybernetics, 2004, 2: 1348-1353.

[10] Chen Yang, Zhang Taining, Guo Peng, *et al*. Quantitative analysis for nonlinear fluorescent spectra based on principal component analysis[J]. Acta Optica Sinica, 2009, 29(5): 1285-1291.
陈 扬, 张太宁, 郭 澎, 等. 基于主成分分析的复杂光谱定量分析方法的研究[J]. 光学学报, 2009, 29(5): 1285-1291.

[11] Sowoidnich K, Kronfeldt H D. Shifted excitation Raman difference spectroscopy at multiple wavelengths for *in-situ* meat species differentiation[J]. Applied Physics B, 2012, 108(4): 975-982.

[12] Niu Liyuan, Lin Manman, Li Xue, *et al*. Raman spectroscopic analysis of single white blood cell of DM mouse *in vivo*[J]. Laser & Optoelectronics Progress, 2012, 49(6): 063001.
牛丽媛, 林漫漫, 李 雪, 等. 活体糖尿病小鼠中单个白细胞的拉曼光谱分析[J]. 激光与光电子学进展, 2012, 49(6): 063001.

[13] Boyaci I H, Uysal R S, Temiz T, *et al*. A rapid method for determination of the origin of meat and meat products based on the extracted fat spectra by using of Raman spectroscopy and chemometric method[J]. European Food Research and Technology, 2014, 238(5): 845-852.

[14] Duraipandian S, Zheng W, Ng J, *et al*. Near-infrared-excited confocal Raman spectroscopy advances *in vivo* diagnosis of cervical precancer[J]. Journal of Biomedical Optics, 2013, 18(6): 067007.

[15] Chen G, Hu H, Chen R, *et al*. Statistical classification based on SVM for Raman spectra discrimination of nasopharyngeal carcinoma cell[C]. International Conference on Biomedical Engineering and Informatics, 2012: 1000-1003.

[16] Guo Libin. Raman spectral analysis of cancer tissue based on support vector machines[D]. Fuzhou: Fujian Normal University, 2015.
郭利斌. 基于支持向量机算法的癌症组织拉曼光谱数据分析[D]. 福州: 福建师范大学, 2015.

[17] Xia Dailin, Meng Hongxia, Zhang Yangde, *et al*. Study of the methods of wavelet feature extraction and neural network classification of fluorescence spectra to improve the diagnostic rate of colonic earlier stage cancer[J]. Spectroscopy and Spectral Analysis, 2006, 26(11): 2076-2079.
夏代林, 孟红霞, 张阳德, 等. 提高结肠早癌诊断率的荧光光谱小波特征提取与神经网络分类方法研究[J]. 光谱学与光谱分析, 2006, 26(11): 2076-2079.

[18] Kuang Ainong, Sun Li, Zhou Jing. Application of BP neural network in discrimination of Raman spectra[J]. Laser Journal, 2016, 37(4): 25-28.
况爱农, 孙 丽, 周 静. BP 神经网络在拉曼光谱判别中的应用[J]. 激光杂志, 2016, 37(4): 25-28.