文章编号：1005-5630(2020)02-0026-06 **DOI**: 10.3969/j.issn.1005-5630.2020.02.005

基于 PCA 和 SVM 算法的肝癌细胞显微后向散射光谱分类

王 成[1]，史继毅[1]，郑 刚[1]，项华中[1]，陈明慧[1]，张大伟[2,3]
（1. 上海理工大学 生物医学光学与视光学研究所，上海 200093；
2. 上海理工大学 上海市现代光学系统重点实验室，上海 200093；
3. 上海理工大学 教育部光学仪器与系统工程研究中心，上海 200093）

摘要：为了实现对肝癌的早期实时和在体探测，基于前期搭建的光纤共聚焦后向散射(FCBS)光谱仪获取肝癌细胞的显微后向散射光谱，分别使用主成分分析(PCA)和支持向量机(SVM)两种算法，对获得的正常肝细胞株(L02)、低转移潜能肝癌细胞株(MHCC97-L)和高转移潜能肝癌细胞株(HCCLM3)三种细胞的后向散射光谱进行分类。使用 PCA 对获得的三种细胞光谱数据进行降维分析，得到的前两个主成分综合了全部信息的 95.4%，由主成分 1 和主成分 2 的得分图可以观察到，三种细胞在直观上有明显的区分；对同一数据集选取 69 例对象通过 SVM 机器学习算法训练分类模型，随机抽取 50 例作为训练集，19 例作为预测集，最终分类的准确度达到了 94.7%。实验结果表明：使用光纤共聚焦后向散射(FCBS)光谱仪获取的细胞显微后向散射光谱可以分别通过 PCA 和 SVM 对不同转移潜能的肝癌细胞进行自动分类，这将为研究活检提供必要的检测手段。

关键词：后向散射光谱；细胞分类；主成分分析；支持向量机；肝癌细胞
中图分类号：TP 751 **文献标志码**：A

Classification of micro-backscattering spectra of liver cancer cell based on PCA and SVM algorithm

WANG Cheng[1], SHI Jiyi[1], ZHENG Gang[1], XIANG Huazhong[1], CHEN Minghui[1], ZHANG Dawei[2,3]
(1. Institute of Biomedical Optics & Optometry, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Shanghai Key Laboratory of Modern Optical System, University of Shanghai for Science and Technology, Shanghai 200093, China;
3. Engineering Research Center of Optical Instruments and Systems, Ministry of Education, University of Shanghai for Science and Technology, Shanghai 200093, China)

收稿日期：2019-05-14
基金项目：国家自然科学基金(61775140)
作者简介：王 成(1977—)，男，副教授，研究方向为生物医学光学。E-mail：c.wang@usst.edu.cn
通信作者：张大伟(1977—)，男，教授，研究方向为微纳光学和光学生物传感器。E-mail：dwzhang@usst.edu.cn

Abstract: In order to realize the clinical detection of hepatocellular carcinoma (HCC) in vivo, real time and earlier, a normal liver cell line L02, a low-metastatic-potential hepatocellular carcinoma cell line MHCC97-L and a high-metastatic-potential hepatocellular carcinoma cell line HCCLM3 were measured, respectively, based on the established fiber confocal back scattering micro-spectrometer (FCBS). The principal component analysis (PCA) and the support vector machine (SVM) algorithm were used to classify the acquired spectrums, respectively. The PCA was used to study the spectrum in wavelength range of 500−900 nm. The first two of the principal components have taken 95.4% of the whole information; therefore, the three kinds of cell distribution were distinguished obviously on the scores diagram of principal component. 69 object data were chosen randomly to train the SVM classification model. 50 sets of these data were used as training sets and 19 sets were used as testing sets. The classification accuracy of the model has reached 94.7%. These results have indicated that the back-scattering micro-spectra of cells measured by fiber confocal back scattering micro-spectrometer (FCBS) combined PCA or SVM could classify liver cancer cells with different metastatic potential automatically. This will provide the necessary testing tools for the research of hepatocellular carcinoma cell in vivo and real time.

Keywords: back-scattering spectrum；cell classification；principal component analysis；support vector machine；hepatocellular carcinoma cell

引　言

肝癌分为原发性和继发性两大类，其中原发性肝细胞癌（hepatocellular carcinoma，HCC）是全世界排名第 6 位的常见恶性肿瘤，其致死率世界排名第 3 位[1]。目前，手术切除仍然是临床肝癌治疗的主要手段，但即使是彻底性切除，5 年内转移复发的概率仍高达 60%～70%[2]，转移复发已成为提高肝癌生存率的瓶颈[3]。临床的转移是一个不断复制筛选的过程，这种筛选使有转移潜能的细胞数量增多。在筛选过程中，出现了一些具有不同转移潜能的癌细胞，转移潜能大的癌细胞和转移潜能小的癌细胞在形态结构和基因表达上有明显的差异性[4]。对不同转移潜能的肝癌细胞的检测，对抑制和预测癌症的转移复发有重要的临床意义。

目前，肝癌检测的金标准仍然是病理分析，但是，随着计算机技术和生物医学技术的发展和应用，出现了许多新的检测手段，例如：超声探测根据回声的不同可以检测组织病变情况；CT 扫描和磁共振成像可以对肿瘤进行快速检测。但这些手段不能呈现细胞水平的图像，而光学检测以其非侵入、非接触的优点为临床诊断和活体细胞的研究提供了更有利的工具[5]。内窥式激光共聚焦显微镜[6]、内窥式光学层析成像（OCT）[7]可以在细胞水平区分正常组织和癌变组织，这两种光学技术在无标记条件下，都是基于光散射的探测。细胞的复杂结构是光散射的主要来源，可以采用散射显微光谱对细胞进行识别[8]。本文使用前期搭建的光纤共焦后向散射（fiber confocal back-scattering, FCBS）显微光谱检测了正常肝细胞和不同转移潜能的肝癌细胞的后向散射显微光谱，再分别结合统计数据分析和机器学习算法对其进行自动分类、识别。

1　实验方法

1.1　主成分分析法

主成分分析（PCA）是一种常用的数据分析方法，目的是对原始数据进行降维，把原本相关性很高的变量转换为较少几个彼此互相独立不相干的变量，这几个变量包含了原始数据的主要信息，称为主成分[9]。由于所采集的光谱数据采样间隔是纳米或亚纳米，一条光谱曲线是维数较高的二维数据。PCA 可以把高维数据映射到低维

空间，解决因为维数过多计算量大的弊端。

PCA 的算法步骤如下：

设光谱数据为 $\{\alpha_1,\alpha_2,\alpha_3,\cdots,\alpha_n\}$，组成样本矩阵 $\boldsymbol{\alpha}$，对其中每一个样本数据标准化处理：

$$\alpha(\lambda)=\frac{\alpha_i-\alpha_{\min}}{\alpha_{\max}-\alpha_{\min}} \tag{1}$$

式中：$\alpha(\lambda)$ 为归一化后的光谱数据；$\alpha_{\max}$ 和 $\alpha_{\min}$ 分别为光谱数据每一个样本的最大值和最小值。计算样本矩阵 $\boldsymbol{\alpha}$ 的协方差矩阵：

$$\boldsymbol{c}=\mathrm{cov}(\alpha)=\begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \tag{2}$$

将 $\boldsymbol{c}$ 的特征值从大到小排列，组成特征值矩阵，取前 k 个特征值，一般累积贡献率要大于 85%[10]。

1.2　支持向量机

支持向量机（support vector machine，SVM）在解决小样本、非线性和高维数据方面具有特殊的优势。SVM 主要用于定性分类、定量回归和预测，近年来在生物医学、图像处理、模式识别等方面得到了广泛的应用[11]。

SVM 的目标是寻找到一个最优超平面使得每类样本与超平面之间的间隔最大，从而对样本实现分类。分类超平面表示为：

$$\boldsymbol{\omega}^{\mathrm{T}}x+b=0 \tag{3}$$

式中：$\boldsymbol{\omega}$ 为超平面的法向量；b 为常数，表示数据直线拟合的截距。每类数据到最优超平面的几何间隔为 $1/\|\omega\|$，由此寻找最优超平面，等价于最小化 $1/2\|\omega\|^2$。为了解决最小化问题，引入 Lagrange 函数，转化为对偶问题：

$$\mathcal{L}(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{n}\alpha_i\left(y_i\left(\omega^{\mathrm{T}}x_i+b\right)-1\right) \tag{4}$$

式中 $\alpha_i>0$ 为 Lagrange 乘数。求解对偶问题，首先固定 α，让 $\mathcal{L}$ 关于 $\boldsymbol{\omega}$ 和 b 最小化，分别对 $\boldsymbol{\omega}$ 和 b 求偏导数使其等于零,代入后得到：

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)$$

$$\text{s.t.}\quad \sum_{i=1}^{n}\alpha_i y_i=0\quad \alpha_i\geqslant 0,\ i=1,\cdots,n \tag{5}$$

对于线性不可分的样本，SVM 的方法是把输入向量投射到更高维度，在更高维度实现线性可分，找寻最优分类超平面。

虽然 SVM 是一个二分类器，对于多分类，SVM 的实现方式有两种：一对多或一对一。本文所用工具包为台湾林智仁教授撰写的 libsvm 库[12]，libsvm 库使用的是多对多的方法。即给定 m 个类，每两个类训练一个二分类器，总分类器个数为 $m(m-1)/2$ 个，每个分类器对样本进行投票，以最终票数结果作为分类结果[13]。

2　实验装置和材料

2.1　实验装置和样品

本实验采用已在早期的文献中报道的 FCBS 光谱仪[8]，主要的检测原理是结合了光纤共焦显微成像和弹性散射光谱技术，可以同时提供单个细胞的背散射光谱和图像。整个系统如图 1 所示，由宽带光源、准直镜、光纤耦合器、光学探头、光谱仪和主控电脑组成。在 400～1 000 nm 范围内响应良好，光谱分辨率为 4 nm。

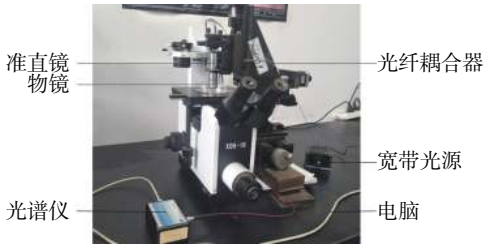
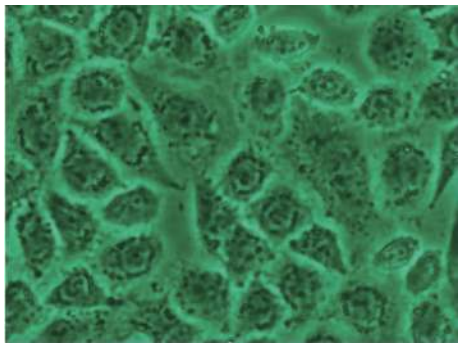


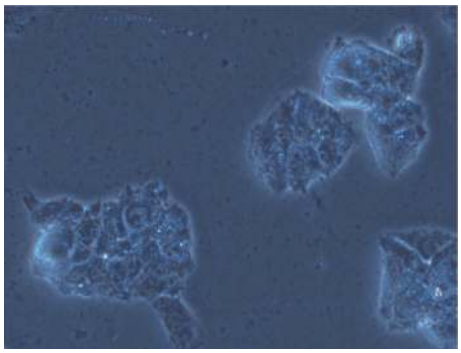
图 1　FCBS 实验装置

Fig. 1　The experimental device of fiber confocal back scattering micro-spectrometer

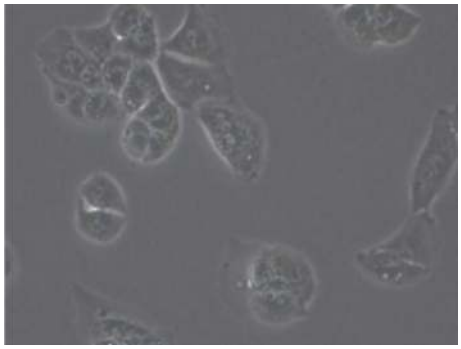
实验所用的人正常肝细胞株（L02）、低转移潜能肝癌细胞株（MHCC97-L）和高转移潜能肝癌细胞株（HCCLM3）样本共 69 例，如图 2 所示，均由复旦大学附属中山医院肝癌研究所提供。细胞均在含 10% 胎牛血清的 RPMI-1640 培养基中生长，培养条件为 37 ℃、CO_2 体积分数为 5% 的细菌培养箱中培养。所有细胞植株均在直径为 35 mm 的培养皿中以低传代数（$<10^5$ 个细胞/mL）培养。培养时间大约为 12 h，最终植株稳定粘附在培养皿上。



(a) 正常肝细胞株



(b) 低转移潜能肝癌细胞株



(c) 高转移潜能肝癌细胞株

图 2　20×显微镜下的细胞图片

Fig. 2　Cell pictures under 20× microscope

2.2　实验方法和步骤

FCBS 基于共焦理论，利用光纤耦合器的单模光纤将光源照射到样品，并在焦点处接收回来的散射光。光纤端面同时作为点光源和点探测，实现了共焦探测。把带有贴壁细胞的培养皿抛却培养液后放到物镜下，在如图 1 所示的系统中，宽带光源经过光纤耦合器耦合到准直镜，经物镜照射到样品表面，同时细胞的后向散射光被准直镜接收，再经光纤耦合器耦合到光谱仪上，获得细胞的后向散射光谱。光谱仪接收到的光谱数据传输到电脑进行处理分析。FCBS 光谱仪带有一个观察部分，用来确保检测的是单个细胞[8]。

3　实验结果

3.1　光谱分析

图 3 为光谱仪所采集到的三种细胞的平均后向散射光谱数据，分别为正常肝细胞株（L02）、低转移潜能肝细胞株（MHCC97-L）和高转移潜能肝细胞株（HCCLM3）。其中横坐标为波长；纵坐标是相对后向散射光强度 S，即相对于硅片的后向散射光谱。实验测量了可见光到近红外波段，即 450～1 000 nm 波长。

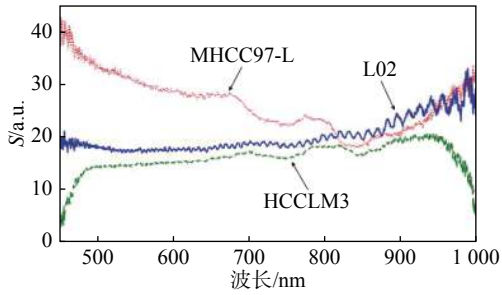


图 3　三种细胞典型光谱

Fig. 3　Typical spectra of three kinds of cells

由图 3 可见，三种细胞在整体曲线趋势上区别很明显。在 500～800 nm 波长范围，低转移潜能肝癌细胞株的散射光强明显高于另外两种细胞，可能是由于细胞癌变后体积变大，相对核仁较少，细胞内部结构分布的不均匀提高了散射系数，导致后向散射光强增大。正常肝细胞株和高转移潜能肝癌细胞株的光谱曲线趋势在 500～900 nm 波长范围有一定的相似性，但高转移肝癌细胞没有如正常肝细胞一样的周期性的变化，且散射光强要比正常肝细胞低，可能是高转移潜能肝癌细胞核仁增多，内部结构的不对称已经破坏了细胞质和细胞核边界所形成的峰，高核质比使细胞对照射光的吸收更强，散射光强变小。

3.2　PCA 降维结果

从图 3 光谱曲线图上看，由于系统误差，可以看到光谱曲线在开始和结尾处有明显的噪声影

响，因此选取 500～950 nm 波段的光谱进行 PCA 分析。对原始光谱数据进行平滑、标准化等预处理后进行主成分分析。从式(2)中协方差矩阵 c 找到一个正交矩阵 p，满足 $p^{T}cp=\lambda$，得到特征值矩阵 λ 后降序排列，这个特征值矩阵就是主成分的贡献率。三种细胞共焦后向散射光谱的主成分贡献率如表 1 所示，由于前两个主成分的累积贡献率已经达到 95.4%，所以前两个主成分已经可以表示原始光谱的主要信息。

表 1　前 8 个主成分贡献率及其累积贡献率
Tab. 1　Contribution rate and the cumulative contribution rate of 8 principal components

主成分	贡献率/%	累积贡献率/%
PC1	79.699 1	79.699 1
PC2	15.775 8	95.444 9
PC3	3.242 5	98.687 4
PC4	0.198 2	98.885 6
PC5	0.128 2	99.013 8
PC6	0.109 7	99.123 5
PC7	0.092 6	99.216 1
PC8	0.085 5	99.301 6

图 4 表示三种肝细胞共 69 个样本的主成分 1、2 的得分图，其中正常肝细胞株 19 组，低转移潜能肝癌细胞株 20 组，高转移潜能肝癌细胞株 30 组。

从图 4 主成分 1 和主成分 2 的得分图可以观察到，三种细胞具有明显的区分。正常肝细胞株(L02)分布在 PC1 的−20 与 20 之间，PC2 的−10 与 30 之间。低转移潜能肝癌细胞株(MHCC97-L)主要分布在 PC1 的正半轴和 PC2 的负半轴。高转移潜能肝癌细胞株(HCCLM3)主要分布在 PC1 的负半轴，PC2 的−20 与−10 之间。

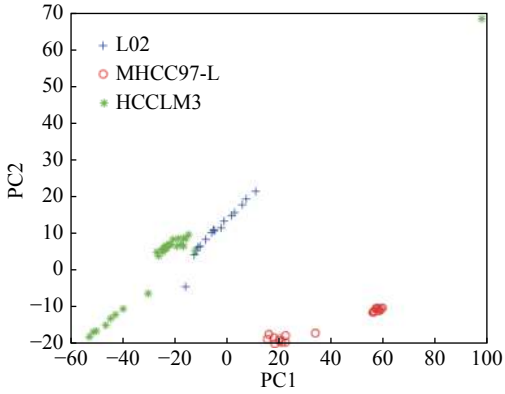


图 4　主成分 1、2 得分图
Fig. 4　Two principal component score of three kinds of cells

3.3　SVM 的模型预测结果

对 FCBS 光谱仪获得的三种细胞后向散射光谱，随机选取其中的 50 组数据(正常肝细胞株 15 组，低转移潜能肝细胞株 15 组，高转移潜能肝细胞株 20 组)作为训练集，首先对数据进行归一化预处理，建立属性矩阵和标签，训练得到模型并对剩余的 19 例样本预测集进行预测，由于光谱数据的特征比较多，选用线性核。预测结果如表 2 所示。

表 2　SVM 预测结果
Tab. 2　SVM prediction results of samples

样本序号	实际物体	预测结果	样本序号	实际物体	预测结果
1	MHCC97-L	MHCC97-L	11	HCCLM3	HCCLM3
2	L02	L02	12	HCCLM3	HCCLM3
3	HCCLM3	HCCLM3	13	HCCLM3	HCCLM3
4	MHCC97-L	MHCC97-L	14	HCCLM3	HCCLM3
5	HCCLM3	HCCLM3	15	MHCC97-L	MHCC97-L
6	HCCLM3	HCCLM3	16	L02	L02
7	HCCLM3	HCCLM3	17	HCCLM3	HCCLM3
8	L02	HCCLM3	18	HCCLM3	HCCLM3
9	MHCC97-L	MHCC97-L	19	L02	L02
10	MHCC97-L	MHCC97-L			

对预测样本预测结果如图 5 所示，由表 2 及图 5 可见，样本序号 8 预测结果错误，其他预测结果准确，整体分类准确率为 94.7%，具有较高的正确率。说明 SVM 可以进行细胞光谱的分类识别。

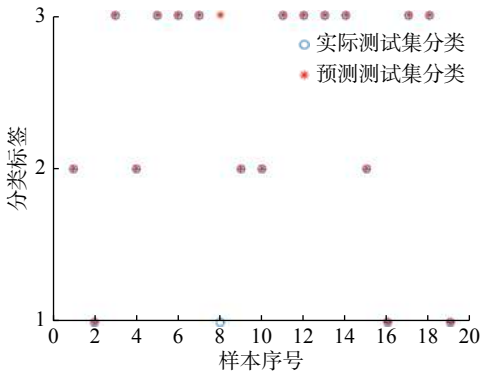


图 5　测试集预测结果图

Fig. 5　Prediction results of testing set

4　讨论与结论

肝癌转移复发的过程中会不断复制筛选出具有不同转移潜能的肝癌细胞，本文基于光纤共焦后向散射光谱系统，采集正常肝细胞株、低转移潜能肝癌细胞株和高转移潜能肝癌细胞株的后向散射显微光谱数据，分别采用 PCA 和 SVM 两种不同的算法对光谱数据进行了自动分类研究。

实验结果显示，PCA 的前两个主成分的累积贡献率达到 95.4%，因此前两个主成分已经包含了原始光谱数据的大部分信息。从图 4 中我们可以看到，三种细胞在主成分 1 和主成分 2 的得分图上的分布较为规律，三种细胞有明显的区分。支持向量机是以统计学习理论为基础的算法，对小样本情况具有非常好的分类效果。本研究中使用 SVM 对不同转移潜能的肝癌细胞训练分类模型并进行预测，准确率达到了 94.7%。

实验证明 FCBS 光谱仪分别结合 PCA 和 SVM 可以对肝癌转移侵袭时不同转移潜能的细胞实现快速、准确的分类。为了进一步提高光谱识别和分类的精度，未来需要进一步增加样本量，建立不同细胞的标准光谱数据库，优化智能识别算法，为临床预测和抑制原发性肝细胞癌的转移侵袭提供有效的检测手段。

参考文献：

[1] 沈玮博, 谭宏涛, 姜洪池. 肝细胞癌抗血管治疗: 现状及新视野 [J]. 中国医刊, 2015, 50(5): 25 – 27, 28.

[2] 樊嘉, 王征. 原发性肝癌综合治疗新进展 [J]. 国际消化病杂志, 2013, 33(2): 73 – 74, 92.

[3] 田慧. HIF-1α 影响肝细胞肝癌转移能力的作用机制研究 [D]. 上海: 复旦大学, 2014.

[4] 张萌, 孙岩, 田伟军. RNA 干扰沉默 ADM 基因对胰腺癌细胞转移潜能影响研究 [J]. 中华肿瘤防治杂志, 2015, 22(19): 1534 – 1539.

[5] WANG C, LI Y, LIU G, et al. Fiber confocal back-scattering micro-spectral analysis for single cell[J]. Technology in Cancer Research & Treatment, 2011, 10(5): 457 – 463.

[6] SAKASHITA M, INOUE H, KASHIDA H, et al. Virtual histology of colorectal lesions using laser-scanning confocal microscopy[J]. Endoscopy, 2003, 35(12): 1033 – 1038.

[7] PAHLEVANINEZHAD H, KHORASANINEJAD M, HUANG Y W, et al. Nano-optic endoscope for high-resolution optical coherence tomography in vivo[J]. Nature Photonics, 2018, 12(9): 540 – 547.

[8] WANG C, GUO X D, FANG B Y, et al. Study of back-scattering microspectrum for stomach cells at single-cell scale[J]. Journal of Biomedical Optics, 2010, 15(4): 040505.

[9] 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.

[10] 杨静, 王成, 谢成颖, 等. 基于主成分分析和反向传播神经网络的肝癌细胞后向散射显微光谱判别 [J]. 生物医学工程学杂志, 2017, 34(2): 246 – 252.

[11] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述 [J]. 电子科技大学学报, 2011, 40(1): 1 – 10.

[12] 刘莹. 图像纹理的特征提取和分类方法研究 [D]. 武汉: 华中科技大学, 2013.

[13] 崔萌, 张春雷. LIBSVM, LIBLINEAR, SVM~(muticlass) 比较研究 [J]. 电子技术, 2015(6): 1 – 5.

（编辑：张　磊）