文章编号:1004-4213(2011)11-1641-5

基于 PCA 和 PNN 的高甘油三脂血清荧光光谱识别

李鹏[1],周建民[1],赵志敏[2]
(1 华东交通大学 机电工程学院,南昌 330013)
(2 南京航空航天大学 理学院,南京 210016)

摘 要:基于主成分分析和概率神经网络,提出了一种有效识别高甘油三脂血清荧光光谱的新方法.研究测量了正常和高甘油三脂血清在 290 nm 和 350 nm 激发光下产生的荧光光谱,并分别以 3 种采样间隔(1 nm、2 nm 和 5 nm)提取荧光强度作为样品的初始特征;利用主成分分析法对初始特征进行分析,以累积可信度大于 95%的主成分作为样品特征;构建了 4 层概率神经网络,并分析了平滑系数和采样间隔对识别效果的影响.实验结果表明,当采样间隔采用 5 nm,平滑系数位于 0.26~0.92区间时,正常和高甘油三脂血清样品的识别率可达到 95%和 100%.

关键词:甘油三脂;荧光光谱;主成分分析;概率神经网络
中图分类号:O433.4 **文献标识码**:A **doi**:10.3788/gzxb20114011.1641

0 引言

血液中甘油三脂含量高于正常标准所导致的血脂异常称为高甘油三脂血症,它是诱发冠心病、心肌梗塞、高血压和糖尿病的直接原因[1].目前,在临床医学上,通常采用酶化法测定血脂常规指标(血清甘油三脂)来判定甘油三脂是否异常,该方法对仪器、试剂、标准液和操作均有较高要求[2].随着光谱技术的发展,利用荧光光谱灵敏度高和选择性强等特点[3-4],对血液成分异常进行识别现已逐渐成为近年来的研究热点[5].然而,目前关于血液成分异常的研究主要以识别血液中血糖[6]或重金属(如,硒、铜和锌)[7]等成分的浓度异常为主,而对于甘油三脂异常的研究很少,其主要原因是由于正常和高甘油三脂血清荧光光谱的相互混叠增加了识别的难度.因而,研究有效的识别方法以提高识别率具有重要的意义.

针对因光谱混叠造成识别率较低的问题,本文首先采用主成分分析(Principal Component Analysis,PCA)对正常和高甘油三脂血清荧光光谱的特征进行了提取,有效减少了光谱混叠对识别率的影响,降低了特征向量的维数;其次,为避免因引入主成分分析导致的频繁的网络训练过程,采用概率神经网络(Probabilistic Neural Networks,PNN)实现了对正常和高甘油三脂血清荧光光谱的有效识别;最后,通过分析平滑系数和 3 种采样间隔对识别率的影响,最终获得了理想的识别效果.

1 光谱实验部分

1.1 样品来源与制备

在南京航空航天大学校医院的配合下,经受试者同意,实验采集了 56 位空腹成年男性的血液,并测试其生化指标用于制备样品.根据成年男性甘油三脂浓度正常标准 0.45~1.7 mmol/L,所采集样品中 28 例为正常样品,另外 28 例为高甘油三脂样品(甘油三脂浓度范围为 1.75~5.62 mmol/L).实验采用纯净水进行混合稀释处理后,将 56 例样品分为 A、B、C 和 D 四组,其中:1)A、B 组样品各 8 个(按浓度由低到高进行编号:A_1~A_8 和 B_1~B_8),分别为正常和高甘油三脂样品,组成训练集;2)C、D 组样品各 20 个(按浓度由低到高进行编号:C_1~C_{20} 和 D_1~D_{20}),分别为正常和高甘油三脂样品,组成预测集.所有样品在分配至训练集和预测集时,样品甘油三脂的浓度分别均匀分布在正常标准和高甘油三脂标准的范围内.此外,由于概率神经网络与有导师神经网络不同,不需通过不断学习修改网络权值,因此选择 16 例训练样品和 40 例预测样品,已可满足识别的需要.

1.2 仪器及光谱数据获取

实验仪器选用日本岛津公司生产的 RF-5301PC 荧光分光光度计.在室温条件下,采用荧光光度计测量各组样品的荧光光谱,测量时用比色皿

基金项目:国家自然科学基金(No.61065002)和华东交通大学博士启动基金(No.09102005)资助
第一作者:李鹏(1976—),男,讲师,博士,主要研究方向为光谱分析与光信息处理.Email:ecjtulipeng@126.com
收稿日期:2011-06-20;**修回日期**:2011-08-20

取 3 mL 样品进行测试，根据高甘油三脂血清样品激发光谱的峰值波长，激发波长(λ_{EX})选用 290 nm 和 350 nm，扫描间隔 1 nm，采用中速自动扫描. 为去除高频随机噪音、基线漂移、样品不均匀和光散射等影响，实验首先对各样品分别进行了 3 次测量，取均值；然后，采用滑动平均滤波法(窗口大小为 5)，对光谱曲线进行平滑处理.

图 1 和图 2 为分别采用 290 nm 和 350 nm 波长激发光测得的 A，B 两组样品的荧光光谱图. 由于系统误差会导致光谱曲线的首、尾端存在较大误差，因此实验分别选取 300～750 nm(λ_{EX} = 290 nm)和 400～670 nm(λ_{EX} = 350 nm)波段的光谱用于分析. 图中，按照特征峰值由大到小的顺序，标注了各光谱曲线所对应的样品编号.

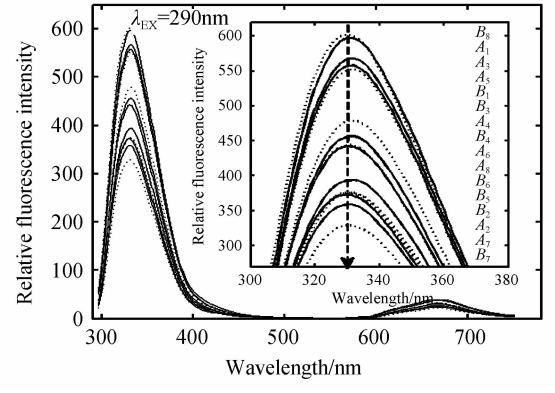

图 1 A、B 组样品的荧光光谱(激发光波长为 290 nm)

Fig. 1 Fluorescence spectra of A, B group samples (λ_{EX} = 290 nm)

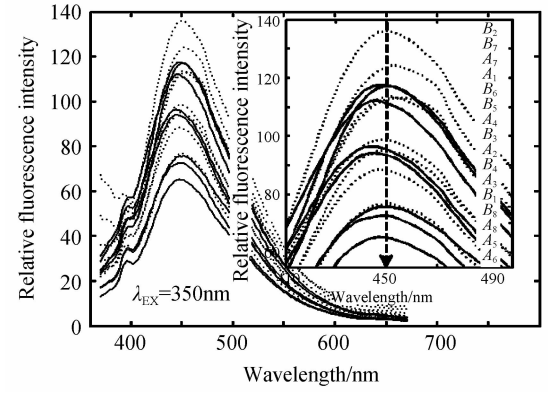

图 2 A、B 组样品的荧光光谱(激发光波长为 350 nm)

Fig. 2 Fluorescence spectra of A, B group samples (λ_{EX} = 350 nm)

由图 1 和图 2 可知：采用波长为 290 nm 和 350 nm激发光所获得的荧光光谱，各样品的特征波长分别为 330 nm 和 450 nm，特征峰值与甘油三脂的浓度没有明显的相关性，且各样品的光谱曲线形状基本相同，该现象与文献[8]一致. 实验结果表明：正常(A 组)样品与高甘油三脂(B 组)样品的光谱出现了混叠现象，要获得满意的识别效果，必须对光谱数据进行分析和处理.

2 光谱初始特征的主成分提取

光谱分析中，与选择单个或几个波长点的荧光强度作为光谱特征相比，分析整个光谱可提供更加丰富的特征信息. 然而，考虑到各样品光谱出现的混叠现象，若直接以整个光谱的荧光强度作为光谱特征用于识别，各类特征之间将包含大量的相关(或混叠)信息，而且采用高维特征进行识别，也会增加识别的难度[9].

因此，为减小光谱混叠对识别率的影响，首先需要以一定采样方式提取全光谱的荧光强度作为光谱的初始特征；然后，在获取足够光谱特征信息的基础上，采用主成分分析法[10]对光谱的初始特征进行提取和降维处理. 具体步骤如下[11]：

1)选择预测样品与训练集样品，建立初始特征矩阵 X. 分别以不同采样间隔(1 nm，2 nm 和 5 nm)抽取预测样品与训练集(A、B 组)样品的荧光强度，经标准化后，建立初始特征矩阵 X 为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n_1} & x_{n_2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

式中，n 为样品个数 17(包括：1 个预测样品和 16 个训练样品)；p 为初始特征维数，其大小由采样间隔决定，当采样间隔取 1 nm，2 nm 和 5 nm 时，p 分别为 722、362 和 146.

2)计算初始特征矩阵 X 的相关系数矩阵 S 为

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{2}$$

式中，$s_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$

3)计算相关系数矩阵 S 的特征值 λ_i(i=1，2，…，p)(其中，特征值 λ_i 按由大到小顺序排列)，及相应的正交化单位特征向量 $T_i=[t_{i1},t_{i2},\cdots,t_{ip}]$. 并计算累计贡献率 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$，确定贡献率大于 95% 的主成分个数 m.

4)计算主成分：$R_i=XT_i'$(i=1，2，…，m)，式中 R_i 为 n 维向量，表示 n 个样品的第 i 个主成分(即，n 个样品的第 i 个特征).

3 概率神经网络识别

为减少光谱混叠对识别率的影响，在对每一个

预测样品进行识别前，均需对该样品和全部训练样品的初始特征进行主成分分析，从而获得用于识别的样品特征. 因此，在识别不同预测样品时，特征值将会发生变化. 概率神经网络[12]所具有的特殊拓扑结构，允许网络根据实际需要修改和变更样品特征，从而避免了因特征值变化而引起的频繁的网络训练过程，与其它神经网络相比，这大大提升了样品的识别速度[13].

概率神经网络是基于贝叶斯分类规则与概率密度函数估计方法发展而来的一种并行分类算法. 其贝叶斯决策的依据是概率密度函数的无参估计.

图 3 为本实验所采用的概率神经网络拓扑结构示意图，该网络属于 4 层网络结构，包括：输入、模式、累加和输出层.

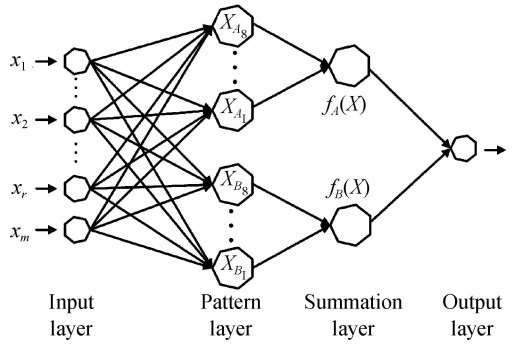

图 3　概率神经网络的拓扑结构示意图

Fig. 3　Probabilistic neural network structure

当 A 组(正常)与 B 组(高甘油三脂)两类训练集样品放入模式层之后，用于识别的神经网络就已训练完成. 预测样品的特征矢量由输入层放入神经网络后，模式层与累加层将计算出该样品在每一模式类的概率密度，具有最大值的那一类将被认为是当前预测样品的模式类. 其中，概率密度的计算公式为[14]

$$f_A(X)=\frac{1}{(2\pi)^{\frac{m}{2}}\sigma^m N_A}\cdot \sum_{i=1}^{N_A}\exp\left[\frac{-(X-X_{Ai})^{\mathrm{T}}(X-X_{Ai})}{2\sigma^2}\right] \tag{3}$$

式中，X 为预测样品的 m 维特征向量；$f_A(X)$ 为 X 属于 A 类的概率密度函数；X_{Aj} 为属于 A 类的第 j 个训练样品的特征向量；N_A 为 A 类中训练样品个数 8；σ 为平滑系数.

4　识别结果与分析

本文通过实验分析了平滑系数和采样间隔对识别效果的影响. 实验采用了 3 种不同采样间隔(1 nm，2 nm 和 5 nm)抽取各样品的荧光强度作为初始特征，通过主成分分析和概率神经网络，以 A，B 两组样品作为训练集样品，依次对 C、D 两组预测集中 2×20 个样品进行识别，并分析了不同平滑系数对识别率的影响，实验结果如图 4. 图中横坐标为平滑系数，纵坐标为识别率，图(a)、(b)和(c)分别为 1 nm，2 nm 和 5 nm 采样间隔下的识别结果.

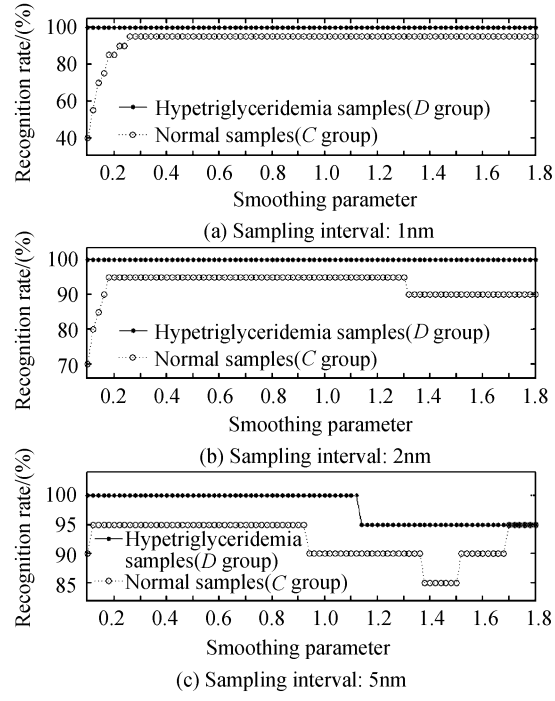

图 4　平滑系数和采样间隔对识别率的影响

Fig. 4　Recognition rate with different smoothing parameter and sampling interval

实验结果表明：

1)当采样间隔为 1 nm 时(图 4(a))，随着平滑系数的增加，高甘油三脂样品的识别率一直保持在 100%，而正常样品的识别率出现了逐渐提高的趋势，并最终稳定在 95%；

2)当采样间隔为 2 nm 时(图 4(b))，随着平滑系数的增加，高甘油三脂样品的识别率同样保持在 100%，而正常样品的识别率则出现了先升高后降低的现象，其最高识别率为 95%；

3)当采样间隔为 5 nm 时(图 4(c))，随着平滑系数的增加，高甘油三脂样品的识别率由 100%降为 95%，而正常样品识别率的波动范围为 95%～85%；

4)三种采样间隔，识别率下降的位置均位于平滑系数变化区间的两端，这是由于概率神经网络的平滑系数 σ 代表高斯核的标准偏差，较小的平滑系数将导致密度函数的估计出现尖峰，而较大的平滑系数又会使密度函数的估计过于圆滑，因此降低了识别的效果；

5)当平滑系数位于 0.26～0.92 区间时，三种采样方式的正常和高甘油三脂样品的识别率相同，分别为 95%和 100%，平均识别率可达到 97.5%；

6)实验结果表明,选取不同采样间隔(1 nm,2 nm 和 5 nm)所需的识别时间分别为 56.7 ms、7.7 ms和 2.4 ms,造成这一差异的原因在于采样间隔的选择会直接影响主成分分析过程.显然,选择 5 nm采样间隔可在保证识别率的前提下,获得更快的识别速度.

5 结论

针对因正常和高甘油三脂血清荧光光谱混叠,而造成高甘油三脂血清识别率较低的问题,实验分别采用 290 nm 和 350 nm 激发光激发正常和高甘油三脂血清,并以不同采样间隔的荧光强度作为样品的初始特征;通过对初始特征的主成分分析,成功地提取了样品的特征向量;并在此基础上构建了概率神经网络模型.实验分析了不同平滑系数和采样间隔对识别效果的影响,最终结果表明,当采样间隔采用 5 nm,平滑系数位于 0.26～0.92 区间时,正常和高甘油三脂血清的识别率分别为 95%和 100%.本文所述的识别方法可为高甘油三脂血症的检测提供一种有效的新途径.

参考文献

[1] SONG Li-ya, YANG Yao-rong. Significance of ratio between triglycerides and HDL for AMI patients [J]. *Journal of Medical Forum*, 2005, **26**(18): 22-23.
宋丽雅,杨耀荣. 急性心肌梗死与脑梗塞患者血脂变化探讨[J]. 医药论坛杂志, 2005, **26**(18): 22-23.

[2] DONG Li-na, SU Yao-dong, SHEN Yu-huan, *et al*. Determination of triglycerides in human serum by reversed-phase high-performance liquid chromatography[J]. *Journal of Tongji University (Medical Science)*, 2004, **25**(2): 104-107.
董莉娜,苏耀东,沈玉桓,等. 反相高效液相色谱法测定人血清中甘油三酯含量[J]. 同济大学学报(医学版), 2004, **25**(2): 104-107.

[3] LI Peng, ZHAO Zhi-min, HONG Xiao-qin. Quantum efficiency evaluation for photoinitiators based on spectral analysis[J]. *Acta Photonica Sinica*, 2009, **38**(11): 2817-2819.
李鹏,赵志敏,洪小芹. 基于光谱分析方法的光引发剂量子效率评估[J]. 光子学报, 2009, **38**(11): 2817-2819.

[4] YANG Yun, YANG Ai-ling. Synchronous fluorescence spectra of standard PAHs and their mixed solutions[J]. *Acta Photonica Sinica*, 2010, **39**(11): 1976-1981.
杨云,杨爱玲. 标准芳烃及其混合溶液的同步荧光光谱分析[J]. 光子学报, 2010, **39**(11): 1976-1981.

[5] LAN Xiu-feng, LIU Jian-gang, LIU Ying, *et al*. Spectroscopy research on cholesterol in hypercholesterolemia serum [J]. *Spectroscopy and Spectral Analysis*, 2006, **26**(3): 467-470.
兰秀风,刘建刚,刘莹,等. 高胆固醇血症血清内胆固醇的光谱学研究[J]. 光谱学与光谱分析, 2006, **26**(3): 467-470.

[6] LING Ming-sheng, QIAN Zhi-yu, LIANG Chao-ying. Research on blood glucose concentration monitoring by fluorescence spectrum [J]. *Chinese Journal of Quantum Electronics*, 2007, **24**(5): 635-639.
凌明胜,钱志余,梁超英. 血糖浓度荧光光谱检测研究[J]. 量子电子学报, 2007, **24**(5): 635-639.

[7] HE Bang-ping, LI Dong-fang, MA Jian-wei, *et al*. Determination of trace copper and zinc in hypertension complicated with hyperlipemia by atomic absorption spectrophotometry [J]. *Spectroscopy and Spectral Analysis*, 2004, **24**(6): 741-743.
何邦平,李东方,马建伟,等. 原子吸收光谱法测定高血压合并高血脂症患者血清铜和锌[J]. 光谱学与光谱分析, 2004, **24**(6): 741-743.

[8] WANG Le-xin. Exploration and research on spectral characteristic of human blood [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2009: 73.
王乐新. 人体血样的光谱特征探索与研究[D]. 南京:南京航空航天大学, 2009: 73.

[9] WANG Ying, GUO Lei, LIANG Nan. A dimensionality reduction method based on KPCA with optimized sample set for hyperspectral image[J]. *Acta Photonica Sinica*, 2011, **40**(6): 847-851.
王瀛,郭雷,梁楠. 基于优选样本的KPCA高光谱图像降维方法[J]. 光子学报, 2011, **40**(6): 847-851.

[10] SU Ling-hua, YI Tong-sheng, WAN Jian-wei. Compression of hyperspectral image based on independent component analysis[J]. *Acta Photonica Sinica*, 2008, **37**(5): 973-976.
苏令华,衣同胜,万建伟. 基于独立分量分析的高光谱图像压缩[J]. 光子学报, 2008, **37**(5): 973-976.

[11] CHEN Xiu-li, WANG Gui-wen, TAO Zhan-hua, *et al*. Raman spectral discrimination of thalassemia erythrocytes based on PCA arithmetic and BP network model[J]. *Chinese Journal of Lasers*, 2009, **36**(9): 2448-2554.
陈秀丽,王桂文,陶站华,等. 基于 PCA 和 BP 网络的地中海贫血红细胞拉曼光谱判别[J]. 中国激光, 2009, **36**(9): 2448-2554.

[12] RAMAKRISHNAN S, IBRAHIEM M M. Comparative study between traditional and modified probabilistic neural networks[J]. *Telecommunication Systems*, 2009, **40**(1-2): 67-74.

[13] INAN G, ELIF D U. Implementing wavelet/probabilistic neural networks for Doppler ultrasound blood flow signals [J]. *Expert Systems with Applications*, 2007, **33**(1): 162-170.

[14] LI P. Structural damage localization using probabilistic neural networks [J]. *Mathematical and Computer Modelling*, 2011, **54**(3-4): 965-969.

Fluorescence Spectra Recognition of Hypertriglyceridemia Serum Using Principal Component Analysis and Probabilistic Neural Networks

LI Peng[1], ZHOU Jian-min[1], ZHAO Zhi-min[2]

(1 *School of Mechanical and Electronical Engineering, East China Jiaotong University, Nanchang* 330013, *China*)

(2 *College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*)

Abstract: A novel method for recognizing fluorescence spectra of hypertriglyceridemia serum was presented based on principal component analysis and probabilistic neural networks. Firstly, two sorts of fluorescence spectra of normal and hypertriglyceridemia serum were measured at 290 nm and 350 nm excitation. And initial feature vectors were obtained from fluorescence intensities at intervals of 1 nm, 2 nm and 5 nm respectively. Secondly, principal component analysis was used to distill initial feature vectors and establish new sample′s feature vectors according to the cumulate reliabilities (>95%). Finally, the probabilistic neural network was designed. Recognition rates with different smoothing parameter and sampling interval were studied. Results show that recognition rates of the normal and hypertriglyceridemia serum are 95% and 100% respectively, when the sampling internal is 5 nm and the smoothing parameter is in range of 0.26～0.92.

Key words: Triglyceride; Fluorescence spectroscopy; Principal component analysis; Probabilistic neural networks