第21卷第1期
2012年2月

激光生物学报
ACTA LASER BIOLOGY SINICA

Vol. 21 No. 1
Feb. 2012

doi:10.3969/j.issn.1007-7146.2012.01.014

基于支持向量机的皮肤自体荧光光谱分类方法在糖尿病筛查中的应用*

张 龙[1]，张元志[1]，王贻坤[1]，朱 灵[1*]，余 锋[2]，张 弓[3]，刘 勇[1]，王 安[1]

（1.中国科学院安徽光学精密机械研究所，安徽 合肥 230031；2.安徽医科大学附属省立医院，安徽 合肥 230001；
3.加拿大温尼伯大学，马尼托巴 温尼伯 R3T6A5）

摘 要：将63例II型糖尿病患者以及140例正常人皮肤的自体荧光光谱分为训练集和测试集两类，针对常用的四种核函数，运用交叉验证、网格寻优法计算最优分类参数，然后结合训练集建模并对测试集分类，结果显示使用径向基核函数时分类效果相对最佳。在此基础上，构建了一种基于线性核函数与径向基核函数的混合核函数，该核函数对人体皮肤自体荧光光谱的分类效果较之于径向基核函数更优，其分类正确率为82.61%，敏感性为69.57%，特异性为95.65%。研究结果表明支持向量机可用于人体皮肤自体荧光光谱的分类，有助于提高糖尿病筛查的正确率。

关键词：支持向量机；荧光光谱；糖尿病

中图分类号：R446

文献标识码：A

文章编号：1007-7146(2012)01-0065-06

Classification of Human Skin Autofluorescence Spectrum Based on Support Vector Machines for Diagnosis of Diabetes

ZHANG Long[1], *ZHANG Yuanzhi*[1], *WANG Yikun*[1], *ZHU Ling*[1*], *YU Feng*[2], *ZHANG Gong*[3], *LIU Yong*[1], *WANG An*[1]

(1. Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, Anhui, China;
2. The Affiliated Provincial Hospital of Anhui Medical University, Hefei 230001, Anhui, China;
3. University of Winnipeg, Winnipeg R3T6A5, Manitoba, Canada)

Abstract: Skin autofluorescence spectrum of 63 type II diabetics and 140 normal subjects were divided into training set and testing set. According to the four commonly used kernel functions in SVM, cross validation and grid-searching were used to calculate the best parameters for classification. Mode was set up using training set and then verified by testing set. The test result indicated that the best choice for classification is radical basis function. A kind of mixed kernel function based on liner kernel function and radical basis function was built and the result of classification was better than using radical basis function. Its accuracy, sensitivity, and specificity were 82.61%, 69.57% and 95.65% respectively. The result proved that SVM is suitable for the classification of human skin autofluorescence spectrum and conduces to improve the accuracy of diagnosis of diabetes.

* **收稿日期**：2011-09-25；**修回日期**：2012-01-12

基金项目：中国科学院知识创新工程青年人才领域专项前沿项目资助课题(O83RC11124)

作者简介：张龙(1979－)，男，安徽人，副研究员，主要从事生物医学光子学研究。(电话)0551－5592128；(电子邮箱)zhanglong@aiofm.ac.cn

* **通讯作者**：朱灵(1982－)，男，安徽人，助理研究员，博士，主要从事生物医学光子学研究。(电话)0551-5592128；(电子邮箱)zhul@aiofm.ac.cn

Key words: support vector machines; fluorescence spectrum; diabetes mellitus

0 引言

晚期糖基化终末产物(Advanced Glycation End products,AGEs)是指在非酶促条件下,蛋白质、氨基酸、脂类或核酸等大分子物质的游离氨基与还原糖的醛基经过缩合、重排、裂解、氧化修饰后产生的一组稳定的终末产物[1]。临床研究表明,人体皮肤组织中的AGEs含量与糖尿病密切相关,有效检测人体皮肤组织中AGEs水平,可以预测受试者患糖尿病及其并发症的潜在风险[2],对糖尿病的预防和筛查具有重要的现实意义。AGEs具有自发荧光特性,其激发波长为300-420 nm,发射荧光波长为420-600 nm[3]。因此,可以通过检测受试者皮肤组织内AGEs的荧光光谱来反映其在皮肤组织中的水平,进而达到筛查糖尿病的效果。

支持向量机(Support Vector Machines,SVM)是一种基于统计学习理论的机器学习方法。2003年,Lin[4]等将SVM用于在体荧光光谱的分类,说明了SVM在用于在体荧光光谱分类上的优势,并运用主成分分析法优化SVM算法,简化了SVM运算过程;2007年,Widjaja[5]等利用SVM和近红外喇曼光谱对结肠组织进行分类,探讨了SVM种类及核函数类型对分类效果的影响;2009年,李建更[6]等利用SVM对胃癌分型标志基因进行提取,在提取之前使用多种降维方法对数据进行处理,给出了详细的数据预处理思想。

本文使用自行设计的AGEs荧光光谱检测装置采集皮肤组织自体荧光光谱,并首次将支持向量机算法应用于AGEs荧光光谱分类,以达到筛查糖尿病的效果。文章主要针对常用的四种核函数,运用交叉验证、网格寻优法计算最优分类参数,然后结合训练集建模并对测试集分类,找出最合适的核函数。在此基础上,运用线性核函数以及径向基核函数构造混合核函数,进一步优化分类模型得到更优的分类效果,给AGEs荧光光谱的分类提供了新的方法,提高了糖尿病筛查的正确率。

1 实验算法与数据处理

1.1 支持向量机算法

支持向量机是建立在统计学习理论和结构风险最小化原理基础上的机器学习方法,它在解决小样本、非线性及高维模式识别中具有许多特有的优势[7]。理论上,支持向量机能够实现线性可分问题的最优分类。对于非线性可分问题,支持向量机则是通过一个非线性映射,将非线性可分问题映射到高维的线性可分空间,然后按照线性可分问题求解。这里的非线性映射由一个内积函数即核函数定义[8],常用的核函数有以下几种:

(1)线性核函数(linear kernel) $K(x_i,x)=x_i\cdot y$

(2)多项式核函数(polynomial kernel) $K(x_i,x)=(y(x_i,x)+m)^d, d=1,2\cdots\cdots$

(3)径向基核函数(radical basis function, RBF) $K(x_i,x)=\exp(-\|x_i-x\|^2/2\delta^2$

(4)Sigmoid核函数(sigmoid tanh) $K(x_i,x)=\tanh(y(x_i\cdot x)+m)$

核函数类型和相关参数的选取,对分类效果有直接影响,对于具体实验数据需要根据实际情况选择恰当的核函数。除此之外,惩罚参数c对SVM的分类效果也有较大影响。惩罚参数表征了对训练误差和泛化能力的折中,参数c的选取可以理解为在特定的子空间中调节经验风险与置信风险的比例,以期获得最好的泛化推广能力。目前核函数和参数的选择主要依据相关经验或者采用交叉验证、网格寻优法[9]。

1.2 病例来源与数据处理

本文使用自行设计的AGEs荧光光谱检测装置[10]在安徽省立医院进行受试者皮肤荧光光谱采集,采集过程征得被测者得同意。测试的主要指标如下:

测试对象:安徽省立医院203例测试者,其中63例为II型糖尿病患者,140例为正常人。

测试位置:受试者前臂内侧(避开血管、疤痕、苔藓样硬化斑以及畸形皮肤位置)。

激发波长:370 nm

发射波长:420-600 nm,平均采样间隔0.25 nm,共采集718个特征点。

文中程序均在Matlab软件中实现,数据处理主要基于林智仁教授开发的Libsvm工具包[11]。核函数参数γ在Matlab程序中用参数g表示,且经验证对于本文数据参数m的取值对分类效果影响不大,故m取0;另为避免过拟合,参数d取3已经足够。

支持向量机分类流程如图1所示,包括训练集和测试集选取、数据预处理、参数寻优、SVM建模、分

类判断以及分类正确率求解六个步骤。训练集和测试集选取,即将测得的数据分为两个部分,一份作为训练集,一份作为测试集。文中训练集包含 40 个糖尿病患者与 117 个正常人,测试集包含 23 个糖尿病患者与 23 个正常人。数据预处理包括归一化处理以及主成分分析,归一化处理的目的是避免奇异样本数据影响并加快程序收敛速度;主成分分析法则是为了降低数据维数,简化 SVM 运算。数据处理完后结合所选核函数进行参数寻优,并根据训练集与测试集分别进行建模与分类测试,最后求解分类正确率。

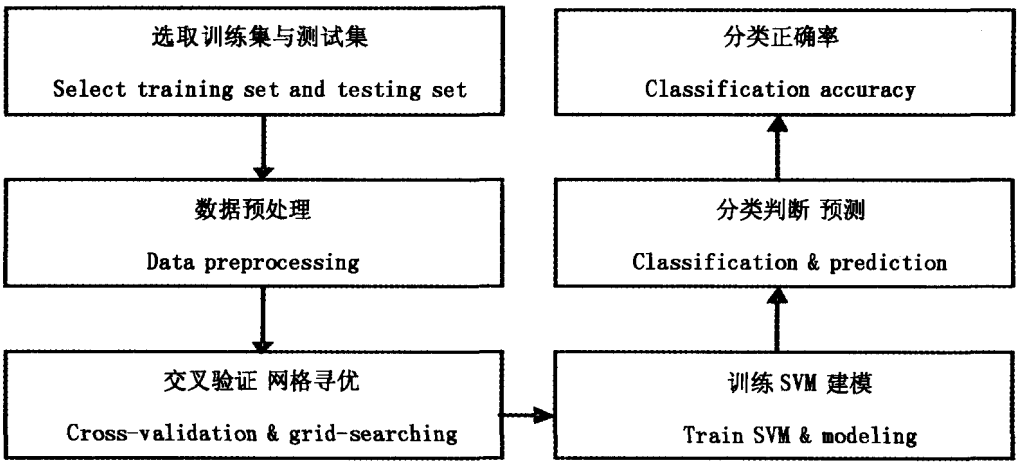

图 1　SVM 分类流程

Fig. 1　Flow chart of SVM

2　实验结果与讨论

本文主要针对线性、多项式、RBF 以及 sigmoid 四种常用核函数进行交叉验证、网格寻优。图 2 和图 3 分别针对线性核函数和径向基核函数,研究分类正确率与相关参数的关系。

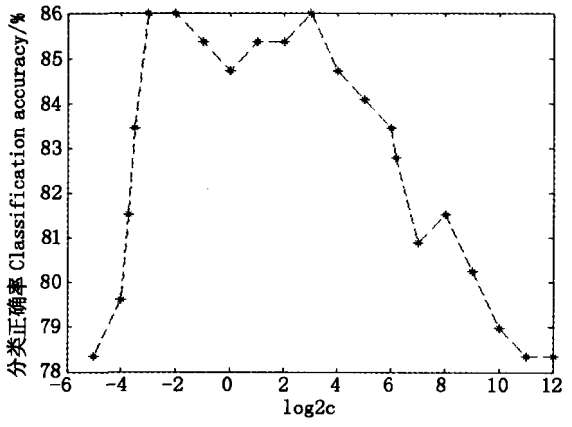

图 2　采用线性核函数时,惩罚参数 c 与分类正确率的关系

Fig. 2　Dependence of classification accuracy on parameter C for a linear SVM2

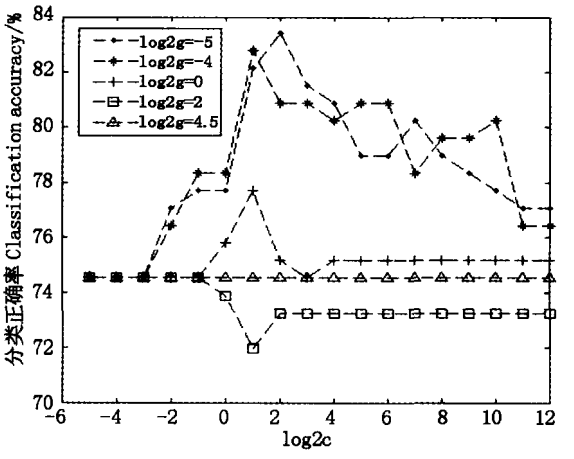

图 3　采用径向基核函数时,在不同参数 g 的情况下,惩罚参数 c 与分类正确率的关系

Fig. 3　Dependence of classification accuracy on parameters C and g for a RBF SVM

当交叉验证平均分类正确率相同时,惩罚参数 c 一般越小越好。因为 c 值过大,会导致过学习问题,降低 SVM 泛化推广能力。因此,文中采用线性核函数时,最优参数 c 取 0.125;采用多项式核函数时,最优参数 c 取 0.0625,g 取 2。根据交叉验证、网格寻优的结果,分别采用线性、多项式、RBF 以及 sigmoid

核函数时，最优的分类参数以及交叉验证平均正确率如表1所示。

表1显示采用线性、多项式及RBF核函数时，交叉验证平均分类正确率都在80%以上，只有采用sigmoid核函数进行交叉验证时，平均正确分类率较低。结合最优参数以及训练集构建分类模型，并代入测试集验证四种核函数分类性能，结果如表2所示。

通过对比可以看出多项式核函数在用于训练集分类时正确率最高，但在用于测试集的分类时正确率反而是最低的，明显存在过学习问题，鉴于它在分类过程中的不稳定性，所以文中排除多项式核函数。而sigmoid核函数对本文数据分类效果太差，也不宜采用。另外，采用RBF核函数对训练集分类时正确率比线性核函数要低，但在对测试集进行分类时其分类正确率要高于线性核函数，即泛化推广能力较好。故得出结论：采用RBF核函数对AGEs荧光光谱分类时效果相对最佳，分类正确率为78.26%，敏感性为60.87%，特异性为95.96%。

目前除了以上四种常用核函数以外，还可以自定义核函数或者线性组合两种核函数生成混合核函数[12]。考虑到采用线性核函数以及径向基核函数对本文数据分类时效果较好，文中通过线性组合线性核函数与RBF核函数建立新的分类模型。混合核函数的公式如下：

$$K_{mix} = \lambda K_{line} + (1 - \lambda) K_{RBF}$$

参数λ的取值范围为(0,1)，经验证文中λ适取0.8，参数g取1，惩罚参数c取2，然后结合训练集建立分类模型，再代入测试集验证，其分类效果与RBF核函数对比如下：

表1 四种核函数的最优参数

Tab. 1 Classic parameters of four different kernels

	线性核函数 Linear kernel	多项式核函数 Polynomial kernel	径向基核函数 RBF sigmoid	核函数 Sigmoid tanh
最优参数 Parameter	c=0.125	c=0.0625 g=2	c=4 g=0.03125	c=0.125 g=0.125
交叉验证正确率(%) Cross-validation classification accuracy	85.99	84.08	83.44	76.43

表2 四种核函数的分类性能

Tab. 2 Results of classification with four different kernels

	线性核函数 Linear kernel	多项式核函数 Polynomial kernel	径向基核函数 RBF sigmoid	核函数 Sigmoid tanh
训练集正确率(%) The training set classification accuracy	86.62	92.99	85.99	75.80
敏感性(%) Sensitivity	65	75	60	15
特异性(%) Specificity	94.02	99.15	94.87	96.58
测试集正确率(%) The testing set classification accuracy	76.09	67.39	78.26	52.17%
敏感性(%) Sensitivity	60.87	69.57	60.87	4.35
特异性(%) Specificity	91.3	65.22	95.96	100

表3　RBF核函数与混合核函数分类效果

Tab. 3　Results of classification with RBF kernel and mixture kernel

	分类正确率(%) Classification accuracy	敏感性(%) Sensitivity	特异性(%) Specificity
径向基核函数 RBF	78.26	60.87	95.96
混合核函数 Mixture kernel	82.61	69.57	95.65

表3显示采用混合核函数时的分类正确率相对于RBF核函数提高到了82.61%,此时敏感性为69.57%,特异性为95.65%。而传统空腹血糖法在空腹血糖值为5.5 mmol/L时,对应的敏感性和特异性分别为77.8%和77.5%,显然,基于支持向量机的皮肤自体荧光光谱分类法在用于糖尿病筛查时已经不输于传统的空腹血糖法。文中影响支持向量机对AGEs荧光光谱分类效果的因素主要包括:实验中使用的荧光光谱受皮肤吸收、散射的影响,光谱校正[13]技术有待进一步发展;数据预处理时,对荧光光谱信息进行了归一化以及主成分分析,虽然减小了奇异样本数据的干扰并加快了收敛速度,但不可避免会丢失部分信息;另外在网格寻优时,为保证足够大的寻优范围,本文取各参数的对数每次步进1,导致部分数据没有测试到,限制了最优参数的选取,从而影响最终的分类效果。随着组织荧光校正技术的发展,归一化函数的优化以及寻优过程的细化,以上影响都能被减小甚至消除,因此SVM在AGEs荧光光谱分类、糖尿病筛查中还有很大的发展空间。

在现有技术条件下,使用SVM对AGEs荧光光谱进行分类时,通常可采用RBF核函数,如果此时分类正确率还不够,可以考虑线性组合线性核函数与RBF核函数,构造新的核函数;核函数参数以及惩罚参数的选取,一般采用交叉验证、网格寻优法,而为保证足够大的寻优范围,一般都是寻找相关参数的对数与分类正确率之间的关系,然而这样会导致一些参数没有测试到,找出的参数可能还不是最优,此时可以考虑先进行大范围内的寻优,再从结果中取出分类正确率较大的部分,然后减小寻优参数的步进,进行精确寻优。另外给不同类别数据设置不同的惩罚参数,可以在敏感性和特异性之间进行取舍,满足筛查和诊断需求。

3　结论

本文将支持向量机用于人体皮肤自体荧光光谱分类,针对常用的四种核函数,进行寻优、建模和分类预测,分析结果显示使用径向基核函数时分类效果相对最佳,其分类正确率为78.26%,敏感性为60.87%,特异性为95.96%。此外通过构建基于混合核函数的分类模型,得到更好的分类效果,其分类正确率为82.61%,敏感性为69.57%,特异性为95.65%,与传统空腹血糖法分类效果相当。研究结果表明支持向量机可用于人体皮肤自体荧光光谱的分类,有助于提高糖尿病筛查的正确率。

参考文献

[1] 孙缅恩, 杜冠华. 晚期糖基化终产物的病理意义及其机制[J]. 中国药理学通报, 2002, 18(3):246-249.
SUN Mianen, DU Guanhua. Pathological significance and mechanism of advanced glycation end products [J]. Chinese Pharmacological Bulletin, 2002, 18(3):246-249.

[2] GENUTH S, SUN W, CLEARY P, *et al.* Glycation and carboxymethyllysine levels in skin collagen predict the risk of future 10-year progression of diabetic retinopathy and nephropathy in the diabetes control and complications trial and epidemiology of diabetes interventions and complications participants with type 1 diabetes [J]. Diabetes, 2005, 54(11):3103-3111.

[3] LUTGERS H L, GRAAFF R, LINKS T P. Skin autofluorescence as a noninvasive marker of vascular damage in patients with type 2 diabetes [J]. Diabetes Care, 2006, 29(12):2654-2659.

[4] WUMEI LIN, XIN YUAN. Classification of *in vivo* autofluorescence spectra using support vector machines[J]. Journal of Biomedical Optics, 2004, 9(1):180-186.

[5] EFFENDI WIDJAJA, WEI ZHANG, ZHIWEI HUANG, *et al.* Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines [J]. International Journal of Oncology, 2008, 32:653-662.

[6] 李建更, 李萍, 严志, 等. 基于机器学习方法的胃癌分型标志基因提取[J]. 中国生物医学工程学报, 2009, 28(4):554-560.
LI Jiangeng, LI Ping, YAN Zhi, *et al.* Selection of gastric cancer subgroups marker genes based on machine learning methods [J]. Chinese Journal of Biomedical Engineering, 2009, 28(4):554-560.

[7] 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):2-10.
DING Shifei, QI Bingjuan, TAN Hongyan. An overview on theory and algorithm of aupport vector machines [J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1):2-10.

[8] S K MAJUMDEER, N GHOSH, P K GUPTA. Support vector machines for optical diagnosis of cancer [J]. Journal of Biomedical Optics, 10(2), 024034.

[9] 奉国和. SVM分类核函数及参数选择比较[J]. 计算机工程与应用, 2011, 47(3):123-124.
FENG Guohe. Parameter optimizing for support vector machines classification [J]. Computer Engineering and Applications, 2011, 47(3):123-124.

[10] 杨三梅, 余锋, 王贻坤, 等. 晚期糖基化终末产物荧光光谱检测法在糖尿病筛查中的应用研究[J]. 激光生物学报, 2011, 20(1):116-119.
YANG Sanmei, YU Feng, WANG Yikun, *et al*. Study on the application of fluorescence spectrum of advanced glycation end products to diagnosis of diabetes mellitus [J]. Acta Laser Biology Sinica, 2011, 20(1):116-119.

[11] CHIH-CHUNG CHANG, CHIH-JEN LIN. LIBSVM: a libsvm for support vector machines[R]. Department of Computer Science: National Taiwan University, 2011.

[12] SMITS G F, JORDAAN E M. Improved SVM regression using mixtures of kernel [C]. Proceeding of the 2002 International Joint Conference on Neural Networks. Hawaii: IEEE, 2002, 2785-2790.

[13] ROBERT S BRADLEY, MAUREEN S THOMILEY. A review of attenuation correction techniques for tissue fluorescence [J]. Journal of the Royal Society Interface, 2006, 3:1-13.