

利用机器学习方法对神经肌肉罕见病 DMD 进行分类预测

章鸣媛^{1,2}, 陈 瑛², 沈 瑛³, 马军山¹

(1. 上海理工大学 光电信息与计算机工程学院, 上海 200093;

2. 上海杉达学院 信息科学与技术学院, 上海 201209; 3. 上海交通大学医学院附属新华医院, 上海 200092)

摘要: 为早期诊断和检测神经肌肉罕见病——杜兴氏肌营养不良(DMD), 设计了一组分类预测试验. 首先, 利用小波变换对 DMD 患者组和健康对照组的磁共振图像(MRI)进行小波分解; 其次, 从所得的分解图像中提取出若干纹理特征参数并进行降维处理; 最后, 再基于这些纹理特征参数, 利用支持向量机算法(SVM)对试验图像进行分类预测. 试验结果显示, 若选择适当的小波分解尺度、分类器核函数和相关参数组合, 则 MRI 图像的分类灵敏度、特异度和准确率分别可达 96.9%, 97.3% 和 97.1%. 该处理方法有望为临床提供客观有效的辅助诊断手段, 可作为 DMD 罕见病无创检测的尝试探索.

关键词: 杜兴氏肌营养不良; 无创检测; 磁共振图像; 纹理特征; 小波变换; 支持向量机

中图分类号: Q 334

文献标志码: A

Classification Prediction of Duchenne Muscular Dystrophy with a Machine Learning Method

ZHANG MingHuan^{1,2}, CHEN Ying², SHEN Ying³, MA JunShan¹

(1. School of Optical Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. College of Information Science and Technology, Sanda University, Shanghai 201209, China; 3. Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200092, China)

Abstract: To diagnose and test the orphan neuromuscular disease——Duchenne muscular dystrophy (DMD) in early stage, an experimental plan was designed. First, the magnetic resonance images (MRI) for DMD patients and healthy persons were decomposed into wavelets by using wavelet transform technique. Then, the dimension reduction was conducted with respect to the texture feature parameters extracted from the decomposed images. In the end, on the basis of the texture feature parameters, the classification and prediction of these images were carried out by using support vector machines (SVM). The results show that if the suitable combination of wavelet function, decomposed scale, kernel function and related parameters were selected, the classification sensitivity, specificity and overall correct classification rate of the MRI images can reach 96.9%, 97.3% and 97.1% respectively. This plan might provide an objective and effective auxiliary method for clinical diagnoses.

收稿日期: 2015-05-07

基金项目: 上海市民办高校重点科研项目(2016-SHNGE-01ZD); IBM 大学合作部联合研究项目(D-2111-15-001)

第一作者: 章鸣媛(1980-), 女, 博士研究生. 研究方向: 医学图像处理. E-mail: zmh_cd@163.com

通信作者: 马军山(1967-), 男, 教授. 研究方向: 光学工程、生物医学工程. E-mail: junshanma@163.com

Keywords: *Duchenne muscular dystrophy; noninvasive detection; magnetic resonance image; texture feature; wavelet transform; support vector machine*

杜兴氏肌营养不良(Duchenne muscular dystrophy, DMD)是一种严重的神经肌肉罕见疾病^[1],由 Duchenne 于 1868 年首先描述.该疾病以骨骼肌进行性坏死为主要病理特征,是常见的致死性骨骼肌遗传疾病之一.DMD 患者几乎均在幼年发病,发病率为活产男婴的 1/3 500.

目前针对 DMD 尚无有效的治疗方法,故高效准确地早期诊断和检测该疾病显得非常重要.DMD 的传统诊断流程一般为:临床疑似→血清肌酸激酶增高→肌电图检查呈肌源性损害→肌肉病理检查或基因检查证实.但肌电图是一项创伤性检查,会带给病人极大不适.由于 DMD 患者多在幼年起病,该项检查必然导致患儿巨大的恐惧与不合作.因此,临床上迫切需要一种无创、敏感的诊断手段,以获取对 DMD 的早期诊断.

作为一种相对安全有效的影像学检查手段,磁共振图像(magnetic resonance image, MRI)可以清晰准确地显示出肌肉的组织结构和病变的位置,因此是神经肌肉疾病诊断中一项非常有价值的辅助检查手段^[2-3].然而,当前对 MRI 图像的信息利用一般仍停留在医生直接读片来获得诊断结果的阶段,而这种方式会受到种种主观因素的干扰,且人眼对图像的分辨率远低于图像本身,图像中蕴含的大量特征信息尚有待发掘.如果能够利用计算机进行辅助分析,则可很好地降低这些因素的影响,提取更多有效的信息,提高诊断准确率.有学者通过分析 MRI 图像所表现的病变组织的纹理特征^[4]来进行辅助诊断,例如利用小波变换来提取图像的纹理特征信息并进一步分析,力求为诊断提供更为客观的依据.

支持向量机(support vector machines, SVM)算法是建立在统计学习的 VC(Vapnik - Chervonenkis)维理论和结构风险最小化理论基础上的机器学习方法,在解决小样本、非线性和高维度模式识别的问题中表现出极大优势.目前 SVM 算法在模式识别、生物学工程、医学影像的分类预测等方面都有广泛的应用.谢红梅等^[5]提出了基于 SVM 的脑部 CT 图像的分类算法.Sohail 等^[6]提出了基于 SVM 的卵巢超声图像的分类算法.Sela 等^[7]采集了肝部 MRI 图像,并提出了基于 SVM 的肝部纤维化等级分类的算法.Fan 等^[8]提出了基于 SVM 的脑部 MRI 图像

的分类算法.以上文献均显示,利用 SVM 算法对医学影像进行分类和预测,可获得较高的分类准确率,达到较好的分类效果.

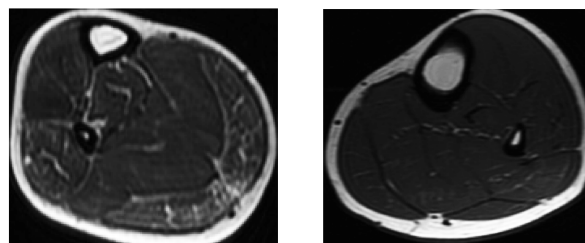
为此,本文通过计算机辅助分析,先对试验 MRI 图像进行小波分解,从每幅图像中均提取出 12 个纹理特征参数并降维,遴选得到 8 个参数;再基于这些纹理特征参数,利用 SVM 算法对试验图像进行有效的分类和预测.力求以良好的分类预测结果为临床医师提供一种 DMD 罕见病的辅助诊断方法,以此作为 DMD 无创检测的技术探索.

1 数据采集与试验方法

1.1 数据采集

本文涉及的所有样本数据均来自上海交通大学附属新华医院小儿神经科门诊,以受检者的 MRI 图像为试验数据对象,利用 GE SIGNA HORIZON 1.5T 磁共振成像仪进行 MRI 成像.根据 DMD 病变位置选定分析图像,共获取有效的 MRI 图像 244 幅,包括 127 幅患者图像和 117 幅健康儿童图像.DMD 罕见病患者组(8.4 ± 1.7 岁)与健康儿童对照组(8.8 ± 1.9 岁)的人数各 15 例,均为男性.DMD 患者组均为经肌肉活检后,病理结果确诊为 DMD 罕见病患者;健康儿童对照组均为运动发育正常、无慢性疾病史、无遗传性疾病家族史的志愿者.本研究已获得受检者家属的知情同意.

图像样例如图 1 所示.



(a) DMD 患者的 MRI 图像

(b) 健康儿童的 MRI 图像

图 1 DMD 患者和健康儿童的 MRI 图像样例

Fig.1 MRI samples of healthy children and patients

1.2 试验方法

1.2.1 利用小波变换提取 MRI 图像的纹理特征参数

DMD 罕见病患者骨骼肌的病理变化主要表现

为肌纤维退变、坏死和再生,间质中有明显的纤维组织和脂肪组织增生,故导致其 MRI 图像上的高亮度信号发生改变.目前已有很多方法来提取纹理特征,如结构分析方法、模型化方法、统计分析方法和频谱分析方法等^[9].频谱分析方法中的小波变换法是近年来广泛应用的数学工具,它可进行时间变换和频率变换,具有多分辨率分析的特点,且在时频两域均具备表征信号局部特征的能力^[10],故本文采用小波变换法来提取图像的纹理特征信息.利用尺度函数和小波函数的正交性,通过二维小波变换,从所选定的每幅 MRI 图像(包括 DMD 罕见病患者及健康对照组)中分别提取出 12 个纹理特征参数.二维小波变换的分解算法过程如下:

$$c_{k,m}^{j,1} = \sum_{l,n} h_{l-2k} h_{n-2m} c_{l,n}^{j+1} \quad (1)$$

$$d_{k,m}^{j,1} = \sum_{l,n} h_{l-2k} g_{n-2m} c_{l,n}^{j+1} \quad (2)$$

$$d_{k,m}^{j,2} = \sum_{l,n} g_{l-2k} h_{n-2m} c_{l,n}^{j+1} \quad (3)$$

$$d_{k,m}^{j,3} = \sum_{l,n} g_{l-2k} g_{n-2m} c_{l,n}^{j+1} \quad (4)$$

从滤波器的角度看,二维小波变换是对图像数据在两个维度上作用两次滤波,这样可得到 4 组系数: $[cA_j, cD_j^H, cD_j^V, cD_j^D]$. 其中, cA_j 是对应图像矩阵 $\mathbf{X}$ 的近似系数, cD_j^H 是细节系数矩阵的水平分量, cD_j^V 是细节系数矩阵的垂直分量, cD_j^D 是细节系数矩阵的对角分量.

本研究选取 DB 作为小波基函数,其消失矩阵的阶数选定为 4,分别对 MRI 图像进行尺度 $L=2, 3, 4$ 的分解.由小波变换过程可知,对一幅图像每做一次(层)小波变换,即可获得水平(H)、垂直(V)和对角(D)共 3 组小波系数,这是最基本的小波变换系数.进一步地,分别对水平、垂直、对角以及全部小波系数求取均值、方差和能量,即可获得每幅图像在第 $i(i=2, 3, 4)$ 层上的 12 个纹理特征参数,分别是:水平均值、水平方差和水平能量;垂直均值、垂直方差和垂直能量;对角均值、对角方差和对角能量;全部均值、全部方差和全部能量.

1.2.2 利用主成分分析进行图像纹理特征降维

主成分分析(PCA)又称主分量分析,由 Pearson 于 1901 年最先提出,又由 Hotelling 于 1933 年做了进一步发展.PCA 是一种优秀的多元统计方法,从数学角度看,这是一种数据的降维处理技术,能将多个变量合并成少数几个不相关的综合指标^[11].PCA 是对一系列相关的变量进行正交变换,得到一些彼此互不相关的变量,即主成分.实际应用中,一般要从所有

主成分中提取累计贡献率尽可能高的综合变量^[12].

为提升预测模型性能,故考虑利用 PCA 进行参数降维,在保证模型精度的前提下,力求以较少的参数个数达到试验目的.根据特征根 λ_c 的数值决定提取主成分的数目.设置 $\lambda_c=1$,即方差大于等于 1 的所有主成分将被保留,其余均被舍弃.以分解尺度 $L=2$ 为例,经 PCA 分析后可得到 3 个贡献率最大的主成分,这 3 个主成分对整体的累积贡献率达到 83.67%.图 2 给出了主成分碎石图,即按特征根从大到小排列的主成分散点图.可见,前 3 个主成分特征根的变化非常明显,后 9 个特征根的变化趋于平稳,因此提取 3 个主成分较为合适.根据纹理特征参数对主成分的贡献情况,遴选得到 8 个纹理特征参数,分别是全部方差、全部能量、对角能量、对角方差、垂直能量、垂直方差、水平能量、水平方差.

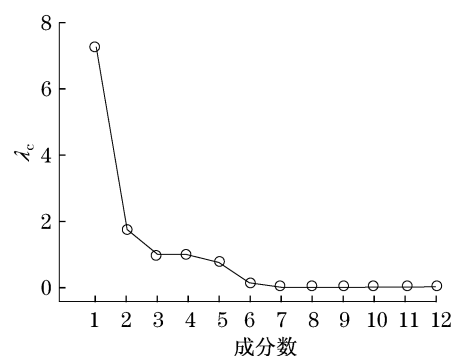


图 2 主成分碎石图

Fig.2 Main components of scree plot

1.2.3 支持向量机分类器设计

支持向量机(SVM)是建立在统计学习的 VC 维理论和结构风险最小化理论基础上的机器学习方法,在解决小样本、非线性和高维度模式识别的问题中表现出极大优势^[13].该理论的关键在于引入了核函数概念.由于低维空间内的样本集常常难以被准确分类,可将低维空间内的样本集映射到高维空间,但会因此增加计算复杂度.核函数的引入完美地解决了该难题,大大提高了机器学习的非线性处理能力.常用的非线性核函数有径向基核函数(RBF)、多项式核函数和 sigmoid 核函数,其中 RBF 应用最广泛^[14-15].有研究证明,RBF 是一种普适核函数,其收敛域较宽,是比较理想的分类函数^[16-17],其表达式为

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (5)$$

式中: x 为空间任一点; y 为核函数的中心; γ 为内核参数,改变 γ 值会令映射函数随之改变,从而决定了线性分类所能达到的最小误差值.另外, RBF

还需确定惩罚因子 C , 该参数用来控制对于错分样本的惩罚程度。

试验图像被随机分成 10 组, 并保证每组中病人图像和正常人图像的分布情况接近总体分布。采用十折交叉验证法组织样本数据, 训练模型并实现对测试样本的分类预测。研究选用 RBF 作为 SVM 的核函数, 同时设置多种 (C, γ) 组合, 根据以往研究经验, 依次选择 $C = 1, 10, 100, 1\ 000$; $\gamma = 0.083, 5, 80, 1\ 000$ 。比较预测结果随 16 种 (C, γ) 组合的变化情况, 确定最优 (C, γ) 组合, 以获得推广能力最优的

SVM 模型。

2 试验结果

从每幅图像中分别提取出 12 个纹理特征参数, 分别是水平均值、水平方差和水平能量; 垂直均值、垂直方差和垂直能量; 对角均值、对角方差和对角能量; 全部均值、全部方差和全部能量。以某位 DMD 患者的某幅试验图像为例进行数据展示, 归一化后的参数示例如表 1 所示。

表 1 某 DMD 患者的某幅试验图像的纹理特征参数(归一化后)

Tab.1 Texture feature parameters of a DMD patient's experimental image (normalized)

L	水平			垂直			对角			全部		
	均值	方差	能量	均值	方差	能量	均值	方差	能量	均值	方差	能量
2	0.107 8	-1	-1	0.039 3	0.032 1	-0.268 9	0.481 0	-0.715 4	-0.870 9	0.042 5	-0.702 8	-0.827 5
3	0.011 6	-0.492 1	-0.648 3	0.250 4	-0.444 3	-0.632 2	0.938 5	-0.531 2	-0.729 1	0.326 7	-0.613 5	-0.735 8
4	-0.259 3	-0.379 6	-0.557 9	0.434 4	-0.256 3	-0.483 3	0.829 0	-0.140 6	-0.335 3	0.432 4	-0.332 5	-0.486 9

在临床医学上, 通常可用灵敏度来衡量试验方法正确地判定患病者的能力, 可用特异度来衡量试验方法正确地判定无病者的能力。

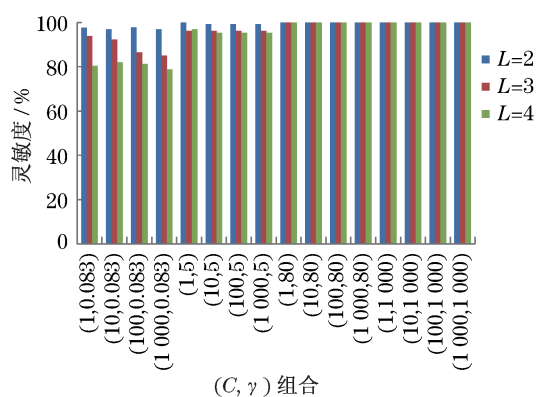


图 3 不同 (C, γ) 参数组合下的灵敏度

Fig.3 Sensitivities for the combinations of different parameters (C, γ)

图 3 和图 4 分别为对 MRI 图像进行尺度 $L = 2, 3, 4$ 的小波分解后, 再对所得统计数据进行分类预测的灵敏度和特异度。其中, 横坐标为 16 种 (C, γ) 参数组合。

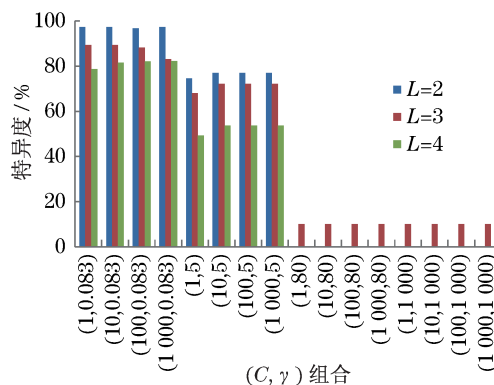


图 4 不同 (C, γ) 参数组合下的特异度

Fig.4 Specificities for the combinations of different parameters (C, γ)

3 分析与讨论

由图 4 可知, 在后 8 组 (C, γ) 参数组合下, 对图像进行尺度 $L = 2$ 和 $L = 4$ 的小波分解时, 均出现了特异度大幅为零的情况, 且 $L = 3$ 时的分类预测效果也不理想, 故只保留前 8 组组合。

3.1 最优分解尺度

在 3 个分解尺度 $L = 2, 3, 4$ 中, $L = 2$ 是最优的。表 2 分别给出了 MRI 图像在 3 个分解尺度下的

灵敏度和特异度的 8 组统计结果(均值 $\pm$ 方差)。可以看出, 在 3 个分解尺度中, $L = 2$ 的分类结果最好。

表 2 MRI 图像在 $L = 2, 3, 4$ 下的分类结果

Tab.2 Classification results for MRI images at $L = 2, 3, 4$

L	灵敏度	特异度
2	0.983 4 $\pm$ 0.000 1	0.867 3 $\pm$ 0.012 4
3	0.927 6 $\pm$ 0.002 1	0.792 4 $\pm$ 0.008 2
4	0.881 5 $\pm$ 0.006 6	0.667 9 $\pm$ 0.023 7

再利用 t 检验对 3 个尺度分别进行两两差异性

检验,如表 3 所示.可知试验图像的 3 个分解尺度之间均存在显著性差异($p < 0.05$).

表 3 3 个尺度下的差异性检查
Tab.3 Differences in check with three levels

分类结果	p		
	$L = 2, 3$	$L = 2, 4$	$L = 3, 4$
灵敏度	0.004 2	0.004 5	0.036 8
特异度	0.000 3	3.96×10^{-6}	1.60×10^{-3}

因此,在 3 个分解尺度中, $L = 2$ 是最优分解尺度.故只需对 MRI 图像进行 $L = 2$ 的小波分解,即可提取最有效的纹理特征参数.

3.2 (C, γ) 组合对分类结果的影响

在本文所选取的 16 种 (C, γ) 组合下, SVM 模型的分类结果有以下特点:当 C 呈增加趋势时,在本文所选取的参数组合中,模型对 DMD 患者和健康儿童的预测差异均不大,即样本预测的灵敏度和特异度受惩罚因子 C 的影响不大;当 γ 呈增加趋势时,在本文所选取的参数组合中,模型对 DMD 患者的预测差异不大,即样本预测的灵敏度受内核参数 γ 的影响不大.但当 γ 增加时,模型对健康儿童的预测能力急剧下降,即特异度迅速降低,甚至出现了特异度大幅度为零的情况.其变化趋势如图 5 所示.

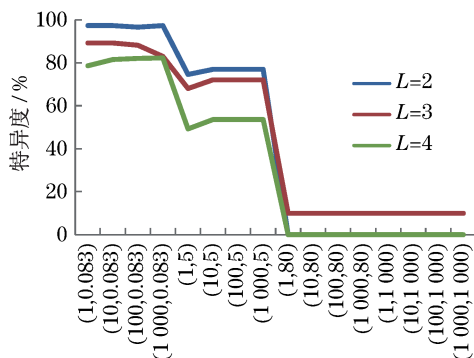


图 5 不同 (C, γ) 参数组合下的特异度变化趋势

Fig.5 Trend of specificities for the combinations of different parameters (C, γ)

综合以上分析,再结合原始试验数据可知:对 MRI 图像进行 $L = 2$ 的小波分解后,可得到最有效的纹理特征参数.针对这些参数的分类预测问题, RBF 的最优 (C, γ) 参数组合共有 4 组,分别是 $(1, 0.083)$ 、 $(10, 0.083)$ 、 $(100, 0.083)$ 及 $(1000, 0.083)$. 其中 $(1, 0.083)$ 组合的分类预测效果最好,其分类预测的灵敏度、特异度和分类准确率分别可达 96.9%, 97.3% 和 97.1%.

为检验所遴选的 8 个参数能否充分表征试验图

像的特征,本研究再利用最初的 12 个参数进行同样的分类预测试验,仍采用十折交叉验证,并保证样本的组织形式不变.结果显示,其分类预测的灵敏度、特异度和分类准确率分别是 97.6%, 97.3% 和 97.5%.可见,利用所遴选的这 8 个纹理特征参数来进行分类预测,其试验精度几乎没有降低.

4 结 论

利用小波变换及主成分分析,从试验图像中优化提取 8 个纹理特征参数,对这些参数进行基于 SVM 的建模,以实现 MRI 图像的分类预测.根据分类预测结果可知,若选择合适的小波基函数、小波分解尺度、SVM 分类器核函数及相关参数组合,则该模型具有极高的分类预测精度,其灵敏度、特异度和分类准确率分别可达 96.9%, 97.3% 和 97.1%.因此,本文所提及的试验方法是判断该受检者是否罹患 DMD 的一种可行的辅助诊断手段,能有效降低因有创检测带来的巨大身心痛苦,可作为 DMD 罕见病无创检测的尝试探索.

参考文献:

- [1] 黄艳军,郭虎,何燕,等. Duchenne 肌营养不良 15 例临床及基因分析并文献复习[J]. 中华临床医师杂志(电子版),2014,8(7):1275-1278.
- [2] 李悦,李小明,卜碧涛. 磁共振在肌肉疾病诊断中的应用[J]. 神经损伤与功能重建,2011,6(4):302-304.
- [3] Del Gaizo A, Banerjee S, Terk M. Adult onset glycogen storage disease type II (adult onset Pompe disease): report and magnetic resonance images of two cases[J]. Skeletal Radiology, 2009, 38(12):1205-1208.
- [4] Liu C, Yang T, Wang J, et al. Primary methods and applications of image texture research in computer-aided diagnosis[J]. Journal of Clinical Rehabilitative Tissue Engineering Research, 2009, 13(39):7721-7727.
- [5] 谢红梅,连宇,彭进业. 基于 Ncut 分割和 SVM 分类器的医学图像分类算法[J]. 数据采集与处理, 2009, 24(6):734-737.
- [6] Sohail A S M, Bhattacharya P, Mudur S P, et al. Classification of ultrasound medical images using distance based feature selection and fuzzy-SVM[C]// Proceeding of the 5th Iberian Conference. Berlin Heidelberg: Springer, 2011:176-183.
- [7] Sela Y, Freiman M, Dery E, et al. fMRI-based hierarchical SVM model for the classification and grading of liver fibrosis[J]. IEEE Transactions on

- Biomedical Engineering, 2011, 58(9): 2574 - 2581.
- [8] Fan Y, Resnick S M, Davatzikos C. Feature selection and classification of multiparametric medical images using bagging and SVM [C] // Proceedings of the SPIE, 2008: 6914.
- [9] 刘丽, 匡纲要. 图像纹理特征提取方法综述[J]. 中国图象图形学报, 2009, 14(4): 622 - 635.
- [10] 张德丰. MATLAB小波分析[M]. 2版. 北京: 机械工业出版社, 2012.
- [11] 洪素珍. 如何有效利用主成分分析中的主成分[D]. 武汉: 华中师范大学, 2008.
- [12] 李小娟. 基于支持向量机的医学图像相关技术研究[D]. 沈阳: 沈阳理工大学, 2013.
- [13] 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.
- [14] 王斐. 基于多核函数 SVM 的非平衡数据分类研究[D]. 镇江: 江苏大学, 2013.
- [15] 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.
- [16] 周绍磊, 廖剑, 史贤俊. RBF-SVM 的核参数选择方法及其在故障诊断中的应用[J]. 电子测量与仪器学报, 2014, 28(3): 240 - 246.
- [17] 王霞, 王占岐, 金贵, 等. 基于核函数支持向量回归机的耕地面积预测[J]. 农业工程学报, 2014, 30(4): 204 - 211.

(编辑: 董 伟)

(上接第 153 页)

- [15] Clementi L A, Vega J R, Gugliotta L M. Particle size distribution of multimodal polymer dispersions by multiangle dynamic light scattering solution of the inverse problem on the basis of a genetic algorithm [J]. Particle & Particle Systems Characterization, 2012, 27(5/6): 146 - 157.
- [16] Xu F, Cai X S, Ren K F, et al. Application of genetic algorithm in particle size analysis by multispectral extinction measurements [J]. China Particuology, 2004, 2(6): 235 - 240.
- [17] Epstein P S, Carhart R R. The absorption of sound in suspensions and emulsions. I. water fog in air[J]. The Journal of the Acoustical Society of America, 1953, 25(3): 553 - 565.
- [18] Allegra J R, Hawley S A. Attenuation of sound in suspensions and emulsions: theory and experiments [J]. The Journal of the Acoustical Society of America, 1972, 51(5): 1545 - 1564.
- [19] Saraswat M, Sharma A K. Genetic algorithm for optimization using MATLAB[J]. International Journal of Advanced Research in Computer Science, 2013, 4(3): 155 - 159.
- [20] Mohamad N, Ariffin M K A, Ali A, et al. Development of genetic algorithm toolbox using MATLAB in cutting tool path optimization [J]. Scientific Research and Essays, 2013, 8(38): 1848 - 1857.
- [21] 章维, 苏明旭, 蔡小舒. 基于超声衰减谱和相速度的颗粒粒径测量[J]. 化工学报, 2014, 65(3): 898 - 904.

(编辑: 丁红艺)