DOI: 10.13382/j.jemi.2017.01.002

R-AdaBoost 带钢表面缺陷特征选择算法*

刘 坤 赵帅帅 屈尔庆 周 颖
(河北工业大学 控制科学与工程学院 天津 300130)

摘 要:带钢表面缺陷形式的复杂多变给特征的选择带来了困难,为此,提出一种融合特征筛选和样本权值更新的 R-AdaBoost 特征选择算法。该算法在 AdaBoost 算法的每个循环中通过 Relief 算法进行特征的筛选与降维,通过筛选后的特征利用样本的类内类间差去除噪声样本,然后根据 AdaBoost 的动态权值更新样本库,再利用每个循环优化选择得到的最优特征与弱分类器级联成最终的 AdaBoost 强分类器,进行带钢表面缺陷的检测与定位。实验结果表明,针对带钢实际生产线上的划痕、褶皱、山脉、污点等多种缺陷,该算法可以有效提取出具有高区分性和独立性的特征,同时提高了缺陷检测算法的准确率。

关键词: AdaBoost 算法;Relief 特征筛选;特征选择;缺陷检测
中图分类号: TP391.4 文献标识码: A **国家标准学科分类代码**: 510.40

R-AdaBoost strip surface defect feature selection algorithm

Liu Kun Zhao Shuaishuai Qu Erqing Zhou Ying
(School of Control Science and Engineering, Hebei University of Technology, Tianjin 300130, China)

Abstract: The complex and various defects of the steel surface bring great difficulty to the feature extraction and selection. Therefore, this paper proposes a new R-AdaBoost future selection method with a fusion of feature selection and sample weights updated. The proposed algorithm selects features and reduces the dimension of features via Relief feature selection according to updated samples in each cyle of AdaBoost algorithm, and uses reduced features to remove noise samples by intra class difference among samples, and then update sample library according to dynamic weight of AdaBoost. The weak classifiers are trained by the resulting optimal features, and combined to generate the final AdaBoost strong classifier, and detect and locate strip surface defects by AdaBoost two classifiers. Aiming at a variety of defects such as scratch, wrinkle, mountain, stain, etc. in the actual strip production line, the experimental results show that the proposed R-AdaBoost algorithm can effectively extract features with high distinction and independence and reduce the feature dimension, and simultaneously improve the accuracy of defect detection.

Keywords: AdaBoost algorithm; relief feature selection; feature selection; defect detection

1 引 言

由于带钢表面缺陷形成原因的不同,导致缺陷呈现形式存在着多样性、随机性和复杂性等特点[1],现有的缺陷检测方法通过提取缺陷的空域特征和变换域特征[2-3]并将多种类型的特征进行组合以满足多样和复杂缺陷形式的要求,但过高的特征维度难以满足算法实时性的要求。因此,提取和选择有效的图像特征是提高带钢表面缺陷检测精度和实现实时性的重要手段。

AdaBoost 算法[4]作为一种经典的分类器集成算法,还具有特征选择的功能,其通过加大权值来加强对难分样本的学习,提高弱分类器的精度,针对正确的训练样本,可以有效提取出具有最大区分能力的特征。其在实际应用中主要存在以下两个问题:1)训练分类器的特征往往含有冗余特征和无关特征,这导致分类器集成的实

收稿日期:2016-06 Received Date: 2016-06
*基金项目:国家自然科学基金(61403119)、河北省自然科学基金(F2014202166)资助项目

际结果与理论值相差甚远;2)训练集样本中存在的噪声样本会影响最终分类器的分类效果,会导致分类器的泛化能力较差,达不到理想的分类效果。

针对这些问题,许多研究者提出解决方案。文献[5]使用 AdaBoost 算法对特征进行有效地降维,但其并没有考虑训练样本对特征筛选和分类器效果的影响;文献[6]通过动态地选择弱分类器并集成强分类器,有效降低了分类系统的复杂度,但忽略了训练集中无关样本和冗余特征的影响。

为此,本文提出了一种 R-AdaBoost 特征选择算法,本文算法在 AdaBoost 分类器集成框架下融入样本选择和特征选择,去除噪声样本和动态更新样本库,并利用 Relief 算法[7]有效剔除了训练样本中的冗余特征,提高了分类器的学习效率和学习精度。通过在实际生产线上采集的带钢表面缺陷库上验证表明,本文算法有效降低了特征的维度,并可以去除训练样本中的噪声样本,使分类器具有更好的泛化能力。

2 带钢表面缺陷特征提取

目前,带钢表面缺陷特征的提取包括灰度特征、几何特征、纹理特征、投影特征和频域特征[8-9]等不同类型的特征。常见带钢表面缺陷特征[10]如图 1 所示。其中灰度特征和几何形状特征分别从缺陷的自身结构和像素级别上对图像提取有效特征,但灰度特征和几何特征对噪声和光照比较敏感。

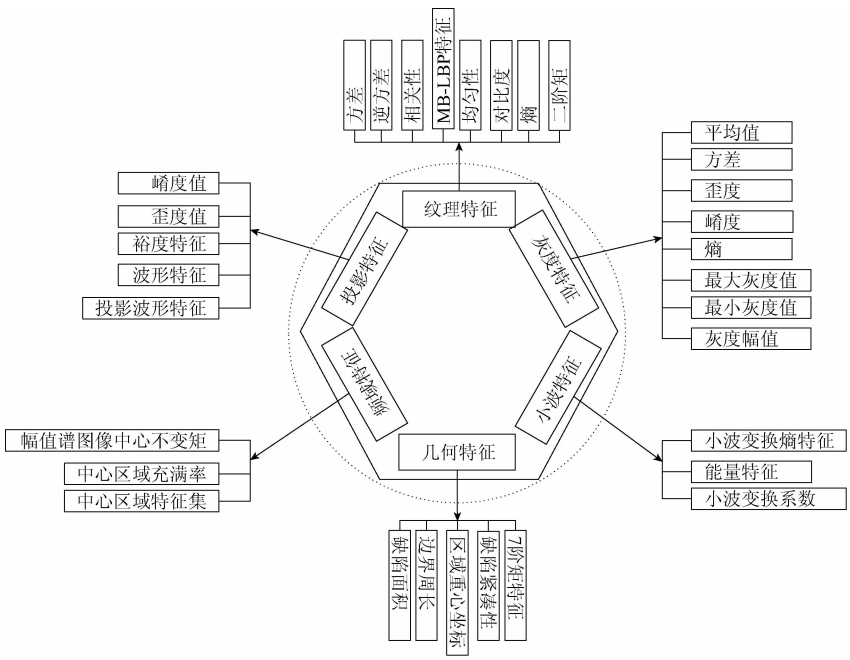

图 1 常见的带钢表面缺陷特征

Fig. 1 Common features of the strip surface defects

局部二进制模式(LBP)最早被作为一种有效的纹理描述算法提出,并且被用于纹理分析、动作检测等领域[11],并且成功应用在不同光照条件下的人脸识别上[12],具有旋转不变性和灰度不变性的优点;近年来,小波变换[13]不断应用于图像处理领域,它具有很好的时域和空域的局部化特性,通过伸缩平移运算对信号逐步进行多尺度细化,最终达到高频处时间细分,低频处频率细分,这样可以去除一些冗余信息,挖掘图像的有效信息。

针对小波特征的优点,本文先对图像进行预处理,然后对图像进行二维小波离散变换并对图像进行三层分解,对每层分解的细节系数提取均值和方差二次特征作为带钢表面缺陷特征。

3 R-AdaBoost 算法原理与流程

3.1 R-AdaBoost 算法原理

R-AdaBoost 算法采用了 AdaBoost 算法集成框架,首先采用 Relief 特征选择算法进行特征选择,然后通过类内类间差判断噪声样本并进行去除,再根据更新的样本权值更新训练样本库,以保证在每个循环中尽量选择较难区分的样本和去除噪声样本及保留最优的特征。这样的优化特征子集训练的分类器往往具有更高的分类精度,且随着训练集中噪声样本的去除和样本库的不断更新及冗余特征、无关特征的去除,大大降低了特征的维数和分类器的复杂度,并使最终强分类器的分类精度也有

所提升。R-AdaBoost 算法原理如图 2 所示。

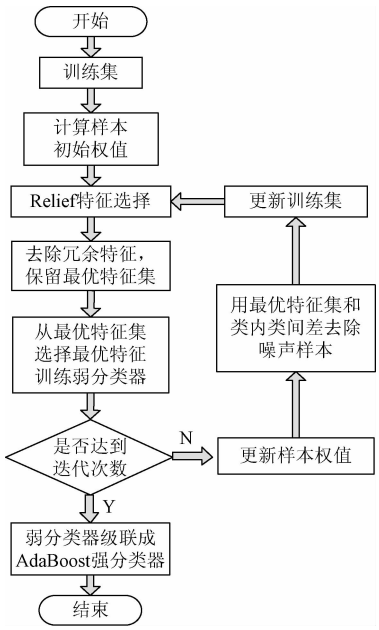

图 2　R-AdaBoost 算法原理

Fig. 2　Schematic diagram of R-AdaBoost algorithm

在 R-AdaBoost 算法中，本文利用 Relief 算法在每次循环根据更新的样本库对特征进行选择，对于权值为负的特征即冗余特征和无关特征去除，保留对样本有效的特征，并且根据每次循环更新的样本库重新选择特征。Relief 特征筛选方法具体步骤如下。

1）初始化：假设样本集为 S，样本权重为 W，W 初始值为 0，抽样次数为 M，特征数为 N，最近邻个数为 K。

2）对于 $i=1,\cdots,M$

（1）从样本集 S 中随机挑选一个样本 D；

（2）寻找 K 近邻同类样本 H，K 近邻不同类样本 M；

（3）对于 $j=1,\cdots,N$，计算 $W(j)$，计算公式如下：

$$W(j)=W(j)-\mathrm{dif}(D,H,j)+\mathrm{dif}(D,M,j) \tag{1}$$

其中差异度衡量方法根据下式计算：

$$\mathrm{dif}(D,H,j)=\frac{D(j)-H(j)}{\max(j)-\min(j)}$$

$$j=1,2,\cdots,N \tag{2}$$

式中：$\max(j)$ 和 $\min(j)$ 分别为该特征在所有样本特征值中的最大值和最小值，$D(j)$ 和 $H(j)$ 分别为样本 D 和 H 的第 j 个特征值。W 为最终的特征排序结果。

3.2　样本选择

本文借鉴 Relief 算法的思想，通过在不同特征维度上样本的类内类间差判断样本是否为噪声样本。对于某个样本 S，从样本集中找到与样本 S 最近的同类样本 H 和不同类样本 M，依次对 N 个特征计算每个特征与样本之间的差异度，若不同类样本之间的差异度大于同类样本之间的差异度，说明该特征判断该样本 S 为正常样本，对所有的特征计算与样本 S 之间的差异度，若能够区分该样本的特征数目大于不能够区分该样本的特征数目，则该样本为正常样本，否则为噪声样本，噪声样本去除方法具体步骤如下。

1）初始化：假设样本集为 D，特征个数为 N，记每个特征的判断结果为 1 或 −1，能够区分样本为 1，反之为 −1，W 为样本的最终判断结果。

2）对于 $i=1,\cdots,N$

（1）从样本集中随机选择一个样本 S；

（2）寻找样本 S 近邻同类样本 H，近邻不同类样本 M；

（3）通过差异度衡量公式来计算 W：

$$W(i)=W(i)+\mathrm{sign}(\mathrm{dif}(S,M,i)-\mathrm{dif}(S,H,i)) \tag{3}$$

3）判断结果：若 W 为正值，则该样本为正常样本，若为负值，则该样本为噪声样本。差异度衡量方法公式按式（2）计算。

R-AdaBoost 算法通过在每次循环中去除冗余特征和样本的选择进行特征的筛选。传统 AdaBoost 算法仅仅根据特征值训练弱分类器，计算每个特征的平均值作为阈值来计算该特征的分类错误率，本文算法在计算错误率时将权值和类别标签结合[14]。在计算弱分类器错误率时，对于每个特征 j，将该特征从小到大排序，L 为实际的分类位置，θ_L 为该特征阈值即分类位置两侧特征的平均值，L 从该特征最小值开始搜索，L 的左侧为 λ（$\lambda\in\{-1,1\}$），右侧为 $-\lambda$，根据判断结果与实际类别标签计算分类错误率，得到最小错误率的位置即为分类位置，两侧特征的平均值即为该特征的阈值 θ_L。

R-AdaBoost 算法通过去除噪声样本、无关特征和更新样本库进行特征提取，具体步骤如下。

1）输入：给定训练样本 $s1=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中 $y_i\in\{1,-1\}$ 表示 x_i 的正确的类别标签，$i=1,\cdots,n$，n 为样本个数，T 为迭代次数，$s2$ 为扩充样本集。

2）初始化：训练集上样本的初始权重分布为：

$$W_1(i)=\frac{1}{n} \tag{4}$$

3）对于 $t=1,\cdots,T$，有：

（1）归一化权重 $w_t(i)=\dfrac{w_t(i)}{\sum_{i=1}^{n}w_t(i)}$ (5)

（2）利用 Relief 算法对特征排序，保留具有较大权值的特征集，并进行噪声样本的去除；

（3）在当前权值下，对于每个特征 j，训练弱分类器 h_j；

（4）将特征与类别标签相结合，挑选出具有最小错误率的弱分类器，计算弱分类器 h_t 的错误率如下：

$$\varepsilon_t = \sum_{i=1}^{n} w_i \left| h_t(x_i) - y_i \right| \tag{6}$$

(5)计算此弱分类器的权重系数为:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{7}$$

(6)根据权重系数更新权重分布如下:

$$\omega_{t+1}(i) = \frac{\omega_t(i)\exp\{-\alpha_t y_i h_i(x_i)\}}{C_t} \tag{8}$$

式中:C_t 为归一化常数,且 $\sum_{i=1}^{n}\omega_t(i) = 1$;

(7)对权重系数进行降序排序,将权重系数和大于 0.9 的样本保留,其他样本和噪声样本由 *S*2 扩充集随机挑选的样本替代,权值保持不变。

4)输出:T 次循环得到的 T 个弱分类器为筛选的特征,将弱分类器级联成 AdaBoost 强分类器。

4 实验结果及分析

为了验证本文特征选择算法的有效性,本文采用邯钢实际生产线上的图像库,并从中选取了十几种不同类型和复杂程度的缺陷图像,图 3 所示为 6 种典型的缺陷样本,包括划痕、麻点、山脉、污点、圆点、褶皱。实验所用的训练集中,缺陷样本共 2 000 张,非缺陷样本共 2 000 张,均从图像库中手工截取 32 ×32 大小样本,测试样本包含 120 张缺陷图像,每类缺陷各 20 张,大约共扫描 3 000个 32 ×32 窗口。

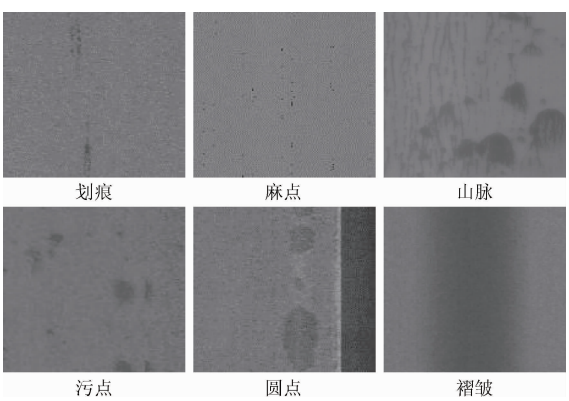

图 3 不同类型的带钢表面缺陷

Fig. 3 Different types of strip surface defects

对于训练集,先用二维离散小波变换函数 *dwt*2 对预处理后的图像在不同尺度上分解,得到 4 个分量:近似分量 *cA*、水平细节分量 *cH*、竖直细节分量 *cV*、对角细节分量 *cD*。对近似分量 *cA* 用 *dwt*2 再次分解,对 3 个细节分量系数直接提取均值和方差特征,即每层对近似系数继续分解,从细节系数中提取特征,本文选用 Haar 小波、Daubechies 小波、Bior 小波等多种小波基函数,对图像进行三层分解,提取均值和方差共 300 多维二次特征作为带钢缺陷特征。

为消除特征属性因大小不一而影响特征的比较,首先对数据进行归一化处理,这里采用最大最小规格化方法,取消由于特征之间量纲和数据变化大小等标准不同造成的额外偏差影响。数据规格化方法如下:

$$V = \frac{\nu - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \tag{9}$$

式中:$\min_A$ 和 $\max_A$ 分别为特征属性 A 的最小值和最大值,最大最小规格方法将属性 A 的一个值 ν 映射为 V 且 $V \in [new\_\min_A, new\_\max_A]$

为了验证噪声样本去除的有效性,对训练样本用类内类间差进行噪声样本的判断和去除。先用 Relief 对具有不同区分度的特征排序,然后选择不同维度的特征利用类内类间差对训练样本进行噪声的判别和去除,用去躁后的训练集训练 AdaBoost 强分类器,然后得到测试样本的测试结果对比图如图 4 所示。

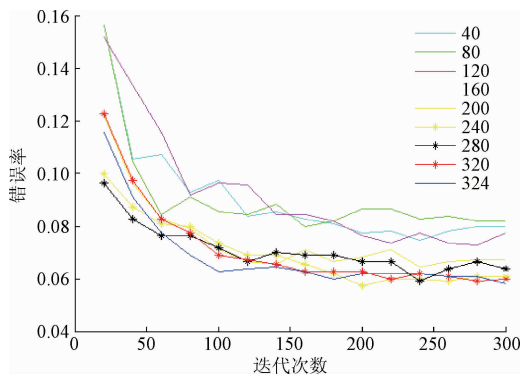

图 4 不同维度特征去除噪声样本的效果

Fig. 4 Effect of removing noise samples at different dimension feature

从图 4 中可以看出,当特征维度较低时,随着迭代次数的增加,测试误差波动幅度比较大,且测试精度比较低,过拟合现象依然存在;当选用的特征维度达到 240 维时过拟合现象基本消除,曲线几乎接近平稳,测试误差波动大大减小,且整体测试精度明显提高。综合考虑算法的复杂度与测试精度,R-AdaBoost 算法在每次循环中选用 240 维度的特征来去除噪声样本。

本文用检测率作为测试性能指标来验证算法的有效性,对提取的特征用本文算法和其他两种方法进行筛选,统计分类器对所有扫描窗口判断错误的个数并计算错误率,本文算法和其他两种特征选择算法对比如图 5 所示,其中横坐标为分类器训练迭代次数,纵坐标为前 n 次迭代集成分类器后的测试误差。第 1 种用 AdaBoost 算法对特征进行筛选;第 2 种算法首先利用 Relief 算法去除冗

余特征,然后再用 AdaBoost 算法对其他特征进行筛选;第 3 种为本文算法筛选特征生成分类器得到的测试结果。

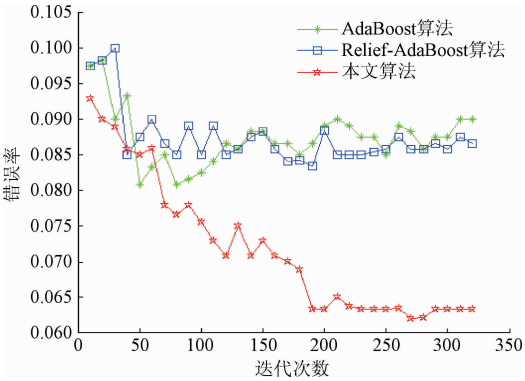

图 5　本文算法与其他算法的对比

Fig. 5　Comparison of the proposed algorithm with other algorithms

从图 5 中可以看出,由于训练集中存在噪声样本和冗余样本,使用单纯的 AdaBoost 分类器随着迭代次数的增加出现了过拟合,测试错误率较高;第二种算法先通过 Relief 算法去除冗余特征,然后再用其它特征训练分类器,过拟合程度有所下降,但测试错误率依然较高;本文算法通过在循环中不断的去除噪声样本,有效改善了噪声样本导致的过拟合现象,且算法的错误率较低。本文的 R-AdaBoost 算法不仅降低了特征的维度,同时还提高了缺陷检测率。

为了进一步验证本文算法的有效性,本文采用 $G_{均值}$ 和 $F_{测度}$ 作为两个性能评估指标,首先通过计算检测结果的真正率(TP),真负率(TN),假正率(FP)和假负率(FN)得到 TP_r 、TN_r 、$P_{recision}$ 和 R_{ecall} 四个性能指标,具体计算公式如下:

$$TP_r = \frac{TP}{TP + FN} \tag{10}$$

$$TN_r = \frac{TN}{TN + FP} \tag{11}$$

$$P_{recision} = \frac{TP}{TP + FP} \tag{12}$$

$$R_{ecall} = \frac{TP}{TP + FN} \tag{13}$$

通过这 4 个性能指标定义 $G_{均值}$ 和 $F_{测度}$ 如下:

$$G_{均值} = \sqrt{TP_r \times TN_r} \tag{14}$$

$$F_{测度} = \frac{2 \times P_{recision} \times R_{ecall}}{P_{recision} + R_{ecall}} \tag{15}$$

式中:$G_{均值}$ 为所有类的综合准确率,其值越大,则表示真正率和真负率同时都大,说明缺陷检测效果好。另一方面,对于给定的真正率,$F_{测度}$ 越大,则假正率和假负率越小。

3 种算法的实验结果如表 1 所示,其中 G 和 F 分别代表 $G_{均值}$ 和 $F_{测度}$。从表中可以看出,选用相同数量的特征,用本文算法筛选的特征比其他两种算法得到的特征可以更有效的表达带钢表面缺陷信息,筛选出来的特征不仅得到较高的检测率,同时也得到了较低的误检率和漏检率,且在特征数目达到 200 时 $G_{均值}$ 和 $F_{测度}$ 值最高,这也说明选取部分合适特征可以达到较高的 $G_{均值}$ 和 $F_{测度}$,这表明本文算法筛选出的特征不仅可以取得较高检测率,同时可以降低误检率和漏检率,与图 5 得到的结果相一致。

表 1　不同特征数目下本文算法与其他两种算法的 $G_{均值}$ 和 $F_{测度}$

Table 1　G_{mean} and $F_{measure}$ of the proposed algorithm with other two algorithms

分类器	AdaBoost 算法		Relief-AdaBoost		本文算法	
特征数目	G	F	G	F	G	F
100	92.3	92.1	91.3	91.2	92.6	92.4
150	91.7	91.5	91.5	91.6	92.8	92.9
200	91.2	91.1	91.9	91.8	93.8	94.0

5　结　　论

本文提出了一种融合样本选择和特征选择的 R-AdaBoost 特征选择算法,利用 AdaBoost 集成算法框架,通过样本选择和特征选择挑选最优特征集,利用筛选出的最优特征训练弱分类器集,并采用加权投票的方式集成 AdaBoost 强分类器。采用本文算法对实际带钢现场缺陷图像库进行测试,结果表明,本文算法可以利用更少的训练集对特征进行有效的搜索,大幅度降低特征维数,经过样本选择和特征选择的特征集构造的分类器与两种典型的特征选择算法相比,具有更好的分类性能,有效降低缺陷检测的误检率和漏检率。

参考文献

[1] GHORAI S, MUKHERJEE A, GANGADARAN M, et al. Automatic defect detection on hot-rolled flat steel products[J]. IEEE Transactions on Instrumentation and Measurement, 2013, 62(3): 612-621.

[2] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. The Journal of Machine Learning Research, 2003, 3(6): 1157-1182.

[3] HUA J, TEMBE W D, DOUGHERTY E R. Performance of feature-selection methods in the classification of high-dimension data[J]. Pattern Recognition, 2009, 42(3): 409-424.

[4] 杨宏晖,王芸,孙进才,等. 融合样本选择与特征选择

的 AdaBoost 支持向量机集成算法[J]. 西安交通大学学报,2014,48(12):63-68.
YANG H H, WANG Y, SUN J C, et al. An AdaBoost support vector machine ensemble method with integration of instance selection and feature selection [J]. Journal of Xi'an Jiao Tong University,2014, 48(12):63-68.

[5] 刘天键. 基于熵的特征选择的 AdaBoost 改进算法[J]. 闽江学院学报,2009, 30(2):60-64.
LIU T J. The AdaBoost algrithm of feature selection based on entropy [J]. Journal of Min Jiang University, 2009, 30(2):60-64.

[6] 郝红卫,王志彬,殷绪成,等. 分类器的动态选择与循环集成方法[J]. 自动化学报, 2011, 37(11): 1290-1295.
HAO H W, WANG ZH B, YIN X CH, et al. Dynamic selection and circulating combination for multiple classifier systems [J]. Acta Automatica Sinca, 2011, 37(11):1290-1295.

[7] KIRA K, RENDELL L A. The feature selection problem: Traditional methods and a new algorithm[C]. AAAI, 1992, 2: 129-134.

[8] MWANGI B, EBMEIER K, MATTHEWS K, et al. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder [J]. Brain, 2012, 135 (PT5): 1508-1521.

[9] SAEYS Y, INZA I, LARRANAGA P. A review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23(19):2507-2517.

[10] 邢芝涛. 基于并行分类器集成的板带钢表面缺陷图像识别[D]. 沈阳:东北大学,2011.
XING ZH T. Recognition of strip steel surface defect images based on parallel classifiers integration [D]. Shenyang: Northeastern University,2011.

[11] RASHID R D, JASSIM S A, SELLAHEWA H. LBP based on multi wavelet sub-bands feature extraction used for face recognition [C]. International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2013: 1-6.

[12] AHONEN T, HADID A, PIETIKAINEN M. Face description with local binary patterns; application to face recognition[J]. IEEE Transacations on Pattern Analysis and Machine Intelligence,2006,28(12):2037-2041.

[13] 郑勋烨. 经典与新型小波理论及其在图像处理中的应用[D]. 北京:中国地质大学,2014.
ZHENG X Y. The classical and novel wavelet theory with application in the field of image processing [D]. Beijing: China University of Geosciences,2014.

[14] WEN X, SHAO L, FANG W, et al. Efficient feature selection and classification for vehicle detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(3): 508-517.

作者简介

刘坤,1980 年出生,2009 年于清华大学获得博士学位,现为河北工业大学控制科学与工程学院副教授,目前主要研究方向为图像处理、计算机视觉与模式识别。
E-mail:liukun03@ mails. thu. edu. cn

Liu Kun was born in 1980, received Ph. D. from Tsinghua University in 2009. She is Currently an associate professor in School of Control Science and Engineering, Hebei University of Technology. Her research interests include image processing, computer vision and pattern recognition.