计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(11):168−173 [doi: 10.15888/j.cnki.csa.006637]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于剪枝决策树的人造板表面缺陷识别①

刘传泽[1], 陈龙现[1], 刘大伟[1], 曹正彬[1], 褚　鑫[1], 罗　瑞[1], 王　霄[2], 周玉成[1]

[1](山东建筑大学 信息与电气工程学院, 济南 250101)
[2](中国林业科学研究院, 北京 100091)
通讯作者: 周玉成, E-mail: 1985786210@qq.com

摘　要: 连续压机生产线的发展, 使人造板实现自动化生产, 但缺陷检测环节仍为人工. 缺陷识别是检测中的一个重要环节, 是根据缺陷特征值利用分类器进行识别的过程. 由于人造板连续生产, 实时性要求高, 为实现缺陷的快速、准确识别, 提出了一种基于剪枝的 CART 树对人造板进行缺陷识别. 通过对已有的人造板缺陷图像进行预处理、分割, 获得缺陷的形状、纹理特征作为输入, 通过基于 Gini 指数的 CART 算法生成 CART 树. 针对于自由生长的 CART 树容易出现过拟合的问题, 利用代价复杂度算法对生成的 CART 树进行剪枝, 通过十折交叉验证对剪枝前后的子树进行比较, 获得最优子树. 通过实验证明剪枝后的 CART 树缺陷识别正确率高达 93%, 满足人造板缺陷识别的实时性和正确率的要求, 可以实现人造板在线缺陷检测.

关键词: 人造板; CART 算法; 特征提取; 剪枝; 缺陷识别

引用格式: 刘传泽,陈龙现,刘大伟,曹正彬,褚鑫,罗瑞,王霄,周玉成.基于剪枝决策树的人造板表面缺陷识别.计算机系统应用,2018,27(11):168–173.
http://www.c-s-a.org.cn/1003-3254/6637.html

Defect Recognition of Wood-Based Panel Surface Using Pruning Decision Tree

LIU Chuan-Ze[1], CHEN Long-Xian[1], LIU Da-Wei[1], CAO Zheng-Bin[1], CHU Xin[1], LUO Rui[1], WANG Xiao[2], ZHOU Yu-Cheng[1]

[1](School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China)
[2](Chinese Academy of Forestry, Beijing 100091, China)

Abstract: The automatic production of wood-based panel has been realized with the development of continuous press production line, but the defect detection is still manual. As an important part of detection, defect recognition is a process of using a classifier to identify defects based on feature value. For the reason of the continuous production of wood-based panels, the defects need to be identified quickly and accurately. Therefore, a cart tree is proposed to identify the defects of the wood-based panel in this study. The defect features of shape and texture are firstly obtained using image preprocessing and image segmentation, and then the cart tree is generated by Gini exponent, at last defects are identified by using the cart tree. But it is easy to cause the problem of overfitting using cart tree without pruning, so the study obtains the optimal subtree by using the cost complexity algorithm and 10 cross-validations. The experimental results reflect that the accuracy rate of defect recognition reaches 93% with the proposed cart tree, which can satisfy the requirements of real-time and accuracy on defect identification.

Key words: wood-based panel; CART algorithm; feature extraction; pruning; defect recognition

① 基金项目: 中央级公益性科研院所基本科研业务费专项资金 (CAFYBB2018MB002); 山东省泰山学者优势特色学科人才团队支持计划 (2015162)
Foundation item: Special Foundation for Basic Researchof Central Level Public Welfare Research Institutes (CAFYBB2018MB002); Taishan Scholar Advantagious and Characteristic Talent Team Support Plan of Shandong Province (2015162)
收稿时间: 2018-03-22; 修改时间: 2018-05-08; 采用时间: 2018-05-14; csa 在线出版时间: 2018-10-24

连续压机生产线的普及使人造板生产快速发展, 但是末端的缺陷检测环节仍依靠人工检测, 存在工作强度大、检测正确率低等问题, 因此研发一套缺陷自动检测系统成为人造板行业的迫切需求. 连续压机连续工作, 板材之间的间距为 400 mm, 运动速度高达 1500 mm/s, 缺陷识别的实时性要求较高, 需寻求一种快速、准确的算法完成人造板在线缺陷识别[1]. 现阶段, 主要的缺陷识别算法有贝叶斯[2]、神经网络[3,4]、支持向量机 (SVM)[5,6]、CART 树[7,8]等, 取得了较好的成果. 由于贝叶斯分类方法需要计算出先验概率, 神经网络、支持向量机分类算法存在计算量大、复杂实时性不高, 均不适用人造板缺陷在线识别. CART 树是一种基于二叉树的算法, 通过特征属性值不断对数据进行二分, 最终达到分类的结果[9]. 缺陷特征属性的选择决定了分类的成功与否[10], 根据人造板的特点提取形状、纹理特征用来表征缺陷. 由于 CART 树容易出现过拟合导致缺陷识别率较低的问题, 用代价复杂度的算法对生成的 CART 树进行剪枝, 使其具有更高的准确率. 本研究提出基于剪枝的 CART 树分类算法具有较高的实时性和正确率, 能够将其应用到人造板在线缺陷检测系统中, 促进缺陷检测的自动化发展.

1 最优 CART 树构建

CART 树采用自顶向下的递归方式, 在决策树的内部节点进行属性值的比较, 并根据不同的属性值判断从该节点向下的分支, 在决策树的叶节点得到分类结果[11]. 如图 1 所示, CART 树由根节点、子节点、叶节点三个部分组成, 其中根节点和子节点代表特征属性, 叶节点代表类别.

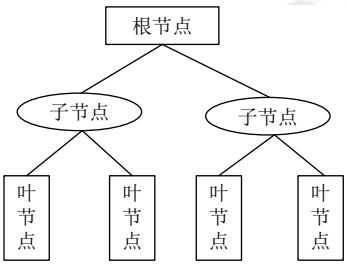


图 1 CART 树示意图

1.1 CART 算法

构建 CART 树的算法流程如下[12]:

1) 建立根节点 N, CART 树开始生长;

2) 如果训练集 L 中只剩下一类样本, 则返回 N 为叶节点;

3) 分别计算当前各个特征属性的 Gini 指数, 将得到最大的 Gini 指数特征属性作为节点, 使 L 划分为 L_1 和 L_2 两个子集, 之后递归构建决策树, 让其充分生长, 不剪枝.

Gini 指数代表当前各个特征属性的分裂的程度, 数值越大说明数据复杂, 不确定性越大, 将 Gini 指数最小的特征属性设置为节点. 其中的 Gini 指数的计算如下:

假设训练集 L 中的样本有 n 个不同的类别 C_i(i=1,…, n). 则概率分布的 Gini 指数定义为:

$$G(p)=1-\sum_{k=1}^{n}p_i^2 \tag{1}$$

式中, p_i 为类别 C_i 在 S 中所占比例. 训练集在某个特征属性 V 下被分为两部分 L_1 和 L_2, 则 L 的 Gini 指数即分裂指数为:

$$G(L)=G(L,V)=\frac{|L_1|}{|L|}G(L_1)+\frac{|L_2|}{|L|}G(L_2) \tag{2}$$

1.2 基于代价复杂度的 CART 树构建

为了提高 CART 树分类器的泛化能力和降低复杂度, 本文采用基于代价复杂度的后剪枝的方法, 使用交叉验证的方式, 确保使人造板缺陷识别的相对误差与构建的决策树中节点数量均保持尽量小, 确定剪枝的阈值 P, 进行剪枝. 因此, 在简化决策树, 可有效保持识别准确率. 具体实现过程如下.

1) 剪枝后各个子树序列 T 构建

T 是一棵是充分生长的 CART 树, 评估决策树的复杂度, 其代价复杂度函数为

$$C_\beta(T)=C(T)+\beta|N_T| \tag{3}$$

式 (3) 中, $C_\beta(T)$值为 T 的代价复杂度; $C(T)$ 值为误分类损失; β 值为剪枝阈值变量; N_T 值为叶节点个数.

以节点为函数 C 为自变量, 则有公式 (4):

$$C_\beta(t)=C(t)+\beta \tag{4}$$

令 β 从 0 开始增加, 直到出现 $C_\beta(t)$=$C_\beta(T)$ 成立的子节点, 得到剪之后的子树 T_2(T_1 为 T), 不断增加 β 的取值, 重复上述过程, 最后只剩下一个根节点, 则得到一系列子树 Tg(g=1, 2,…, n). 当式 (3) 与式 (4) 相等, 则 P 的值为:

$$\beta_g=\frac{C(t)-C(T_t)}{|N_T|-1} \tag{5}$$

式中, $C(t)$ 为子树 T_t 剪枝后节点 t 误分类损失, $C(T_t)$ 是未剪枝时子树 T_t 误分类损失.

2) 最优 DT 构建

在子树 T_g 和 β_g 已知的情况下, 采用交叉验证方法评估子树的分类误差[13], 来确定最佳的剪枝阈值, 在确保正确率的同时, 降低 CART 树的复杂度. 误分类误差的定义为:

$$E_{cv}(T(\beta)) = \frac{1}{N}\sum_{i,j} d(i|j)N_{ij} \tag{6}$$

式中, $E_{cv}(T(\beta))$ 为树 $T(\beta)$ 交叉验证的误分类误差; $d(i|j)$ 为将 j 类误分为 i 类的样本数; N 为训练样本数; N_{ij} 为误分类的样本数.

其交叉验证的过程为, 首先将训练集 L 划分为 K 个子集 $L_n(n=1, 2,\cdots, K)$, 然后从 $L–L_n$ 中生成 K 个子树 T_n. 令 $\beta_g' = \sqrt{\beta_g\beta_{g+1}}$, $T(\beta_g)$ 真实误差以 $T_n(\beta_g')$ 的平均值进行衡量:

$$E_{cv}(T(\beta_g)) = \frac{1}{N}\sum_{i,j} E_{cv}(T_m(\beta_g')) \tag{7}$$

经过循环交叉实验验证, 确定代价复杂度最小的子树 $T_k(\beta)$, 确定最小的相对误差结果是:

$$E_{cv}(T_k(\beta)) = \min E_{cv}(T(\beta_g)) \tag{8}$$

通过上述理论, 即利用 $L–L_n$ 验证误差率、最后确定最佳剪枝阈值 $P(\beta)$, 最终得到误分类误叉最小的 CART 树分类模型.

2 人造板缺陷自动识别

人造板表面缺陷在线识别的实现流程如图 2 所示.

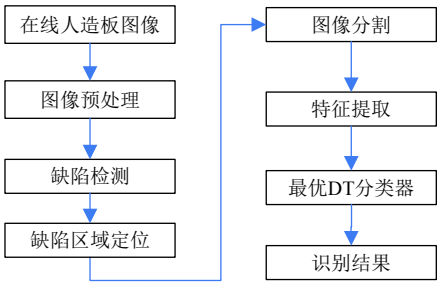


图 2 人造板缺陷自动识别流程

(1) 图像预处理. 对 CMOS 相机获得的在线人造板图像, 用中值滤波去除噪声干扰.

(2) 缺陷检测及区域定位. 将经过预处理的图像进行分块检测, 计算每一各区块的平均灰度值和方差, 将方差大于一定阈值定义为缺陷板, 并将该区块定义为缺陷区域.

(3) 图像分割. 利用 Otsu 阈值分割法对缺陷区域进行分割.

(4) 特征提取. 统计缺陷区域中的形状、纹理特征构成特征向量.

(5) CART 树分类. 通过 CART 树分类器进行分类.

当有新输入的人造板图像时, 首先进行缺陷检测判断是否有缺陷, 如果检测有缺陷, 进行缺陷定位、通过图像分割、提取缺陷的特征值, 作为 CART 树的输入, 通过 CART 树得到缺陷类别. 其中最重要的是特征参数提取和最优决策树 (DT) 的构建.

2.1 特征提取

人造板缺陷特征提取是缺陷识别的关键, 由于人造板图像的背景灰度值均匀, 缺陷区域内的灰度值相似, 各类缺陷的形状差异较大, 基于选取缺陷图像的形状、纹理特征对缺陷进行表征, 其表达式如表 1、表 2 所示.

表 1 形状特征计算公式

序号	形状特征	计算公式
1	面积	$S=N$
2	周长	$L=N_b$
3	圆形度	$O=\frac{4\pi S}{L^2}$
4	长宽比	$OR=\frac{L_{MER}}{W_{MER}}$
5	矩形度	$P=\frac{S}{S_{MER}}$

表 2 纹理特征计算公式

序号	纹理特征	计算公式
1	均值	$u=\frac{1}{N}\sum_{i=0}^{N-1} M(i)$
2	标准差	$\sigma_D=\sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(M(i)-u)^2}$
3	平滑度	$\sigma_P=1-\frac{1}{1+\sigma_D^2}$
4	偏度	$\sigma_S=\frac{1}{(N-1)\sigma_D^3}\sum_{i=0}^{N-1}(M(i)-u)^3$
5	均方根值	$\sigma_R=\sqrt{\frac{1}{N}\sum_{i=0}^{N-1} M(i)^2}$

表 1 中, N 为缺陷区域内像素点的个数; N_b 为缺陷区域边缘像素点的个数; MER 为缺陷区域最小外接矩形; L_{MER} 为 MER 的长; W_{MER} 为 MER 的宽; S_{MER} 为 MER 的面积.

表 2 中, $M(i)$ 为缺陷各点的灰度值, N 为缺陷区域像素点的个数.

2.2 人造板缺陷识别最优 CART 树的构建

采用 CART 算法对大刨花、胶斑、杂物、油污四类缺陷的人造板图像进行训练, 从而获得缺陷自动识

别 CART. CART 分类树的构建实质是建立人造板缺陷特征值与缺陷类别的非线性映射关系. 其输入向量为 $Q(S、L、O、OR、P、u、\sigma_D、\sigma_P、\sigma_S、\sigma_R)$, 输出为 $O(1, 2, 3, 4)$ 分别代表大刨花、胶斑、杂物、油污, 建立映射关系为 $Q\cdots O$.

本文中用 MATLAB 软件进行程序编写, 首先提取 220 张人造板缺陷图像, 提取形状、纹理特征作为实验数据, 抽取其中 200 个数据构成训练集 L_i, 用于训练 CART 树, 剩下的 20 个数据组成训练集 T, 用来验证 CART 树的优劣, 其实现过程如下所示:

1) 将 200 个人造板特征数据 L 分成 10 份即 $L_i(i=1, 2,\cdots, 10)$, 则训练集为 $L–L_i$ 为 CART 树的训练输入, 样本中的缺陷类别标签为训练输出, L_i 为测试数据用来确定剪枝阈值 P;

2) 用 CART 算法训练出自由生长的决策树 T;

3) 通过剪枝获得子树$\{T_1, T_2,\cdots,T_n\}$, 其中 $T=T_1$, 利用 L_i 测每棵子树的真实误差 $E(E_1, E_2, \cdots,E_{10})$ 的平均值, 将 E 中最小的值对应的阈值 P 为最优的剪枝阈值, 从而获得最优的 CART 树.

表 3 测试集混淆矩阵

缺陷类别	大刨花	胶斑	杂物	油污
大刨花	2	0	0	0
胶斑	0	5	0	0
杂物	0	0	6	1
油污	0	0	1	5

2.3 人造板缺陷识别

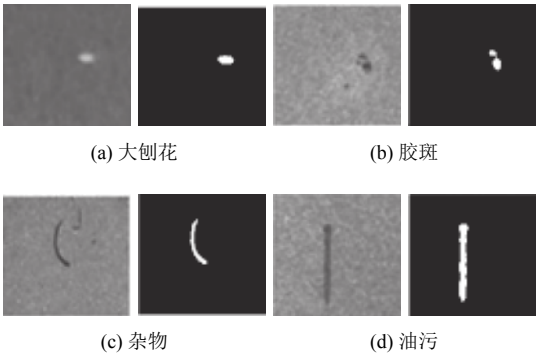
当有新的人造板图像输入时, 首先对图像进行预处理, 通过检测区块的灰度均值和方差判断是否具有缺陷, 如果没有缺陷则不进行分割操作, 如果存在缺陷进行图像分割获得缺陷特征值, 将其作为 CART 树的输入, 通过 CART 树得到输出类别.

本研究的人造板图像是通过人造板生产厂家提供的, 共 220 幅. 通过对人造板缺陷图像其进行预处理、图像分割, 其分割后的图像如图 3 所示, 通过对图 3 中二值图像获得形状特征值, 将分割后的缺陷区域映射到原图中进而获得纹理特征, 从而获得表征缺陷的十个特征数据 ($S, L, O, OR, P, u, \sigma_D, \sigma_P, \sigma_S, \sigma_R$).

3 实验结果及分析

3.1 最优人造板分类决策树构建

在训练集 L 下根据 CART 算法训练出的未经过剪枝的 CRAT 树 T 如图 4 所示, 获得一系列子树及 $P(\beta_1, \beta_2, \cdots, \beta_k)$ 值. 图 4 采用 IF-THEN 的形式表现出数据的分类过程. 当有新数据输入时, 根据根节点的特征属性与相应的数据进行大小判断数据的流向 (此根节点属性为 σ_P, IF $x8>0.05592$, 进入左子树), 进而逐渐的识别出人造板的四类缺陷, 之后确定剪枝阈值 P, 进行剪枝操作.



(a) 大刨花 (b) 胶斑

(c) 杂物 (d) 油污

图 3 四类缺陷分割前后图像

采用 CART 算法对训练数据 $L–L_i$ 得到的子树$\{T_1, T_2,\cdots, T_{10}\}$, 通过由 T 得到的剪枝阈值$\sqrt{\beta_g\beta_{g+1}}$对子树进行剪枝, 通过交叉验证确定最终剪枝阈值 P, P 的确定过程如图 5 所示.

从图 5 可知, 随着 P 值增大 (复杂度增加), 交叉验证相对误差逐渐增大. 图 5 中, P 的值先为无穷大, 叶节点的数量代表树的复杂度, 剪枝过程从 1~12 开展, 共 12 次; 叶节点的数量从 18 变为 8. 通过图 4 可以看出, 交叉验证误差在 P 的值为 0.001 开始增加, 所以最大剪枝次数为 5, 此时交叉验证误差为 0.05, 最终的叶节点数为 8. 通过剪枝后, 决策树的叶节点树量减少, 树的复杂度明显降低, 最终剪枝后的最终结构如图 6 所示.

利用测试集的 20 组特征值数据对如图 6 所示构建的决策树进行检测, 其结果如表 3 所示, 其中行表示实际的缺陷类别, 列代表预测的缺陷类别. 从图中可知, 总的识别率达到 95%, 可以满足人造板的识别正确率. 其中杂物和油污的识别率较低, 主要是这两种缺陷灰度值较为相似, 形状表现形式不一, 容易识别错误.

3.2 算法比较

为了验证本研究提出剪枝的 CART 树分类算法的可行性, 与神经网络、支持向量机 (SVM) 以及未剪枝的 CART 树三种算法在人造板表面缺陷识别上正确率和时间进行比较.

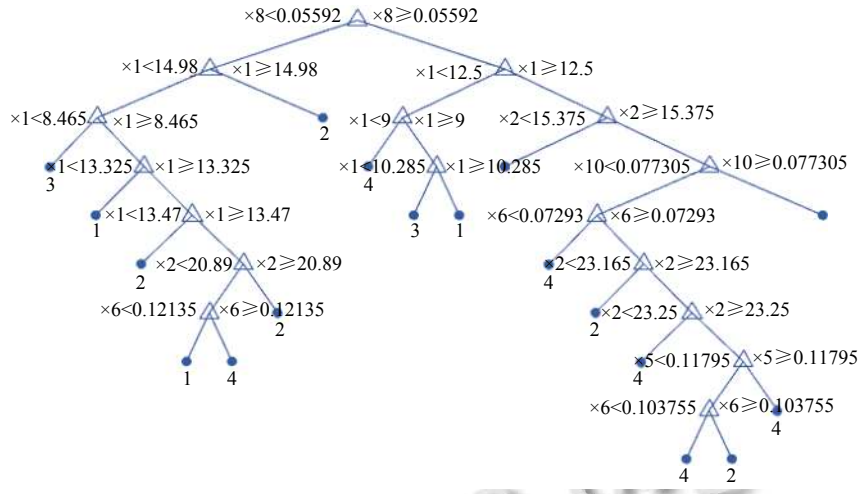


图 4 充分生长的决策树

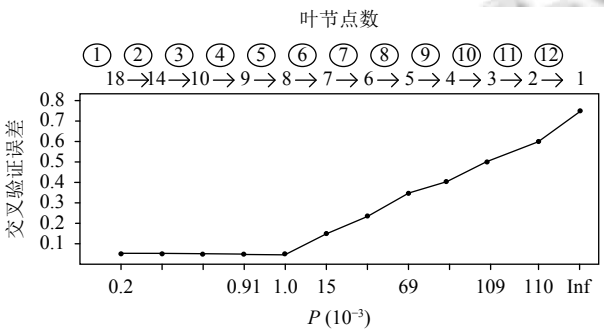


图 5 交叉验证结果

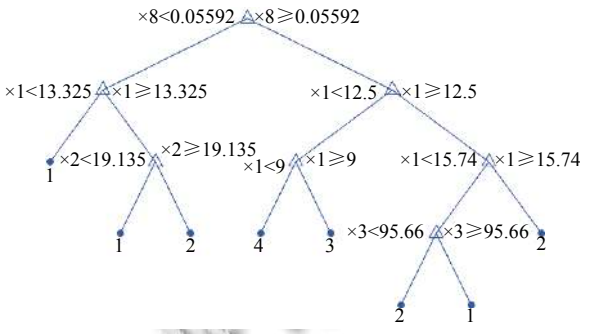


图 6 剪之后的决策树

3.2.1 正确率

利用测试集中的 20 组特征值数据对三种方法进行验证, 其结果如图 7 所示, 其中大刨花、胶斑、杂物、油污缺陷数量为 2、5、7、6. 由图可知, 剪枝后决策树分类器的识别正确率较神经网络、SVM 两种算法有很大的优势, 对于杂物、油污两种难以区分的缺陷仍有较高的识别效率. 剪枝后 CART 树较未剪枝的决策树识别能力强, 这是由于未剪枝决策树在训练的时候对训练样本数据进行充分表达, 泛化能力较弱, 由此可知剪枝后 CART 具有较强的泛化能力.

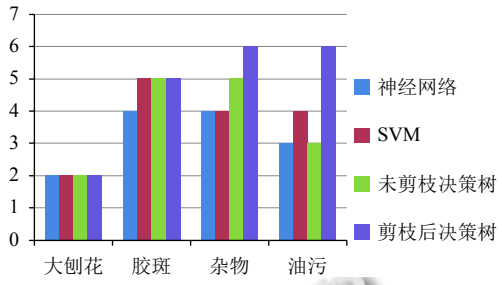


图 7 三种算法缺陷识别比较

3.2.2 算法耗时分析

在 Intel Core I5-7500U 2.4 GHZ, 6 GB 内存的硬件环境下和 Windows10, MATLAB2016b 的软件环境下, 对剪枝后的 CART 树分类算法、神经网络和 SVM 三种方法在获得的人造板缺陷数据上 CPU 运行时间进行分析. 其中图像分割、特征提取为相同程序, 耗时分别为 320 ms、90 ms. 三种算法在训练集、测试集耗时如表 4 所示, 由表可知, 未剪枝的 CART 树在训练和测试耗时最短. 神经网络和 SVM 主要是由于计算复杂, 训练所需时间较长.

表 4 耗时分析

算法	训练耗时 (s)	测试耗时 (ms)
神经网络	60	24
SVM	83	43
未剪枝 CART 树	5	14

从表 4 可以看出, 神经网络、SVM 两种算法时间

很短, 但属于黑盒, 分类过程难以让人理解, 不易应用到实际中. 而决策树在分类过程是根据不同的特征属性的特征值大小对数据不断的进行二分, 容易理解且容易实现. 结合人造板缺陷识别正确率、识别时间和可实行性可知, 基于剪枝 CART 树的分类模型满足人造板在线检测实时性和正确率的要求.

4 应用和结论

将本研究获得的剪枝后 CART 树得出的模糊规则 (IF-THEN) 编写入 C++程序中, 嵌入到基于机器视觉的人造板在线缺陷检测系统中, 缺陷检测率高达 98%, 缺陷识别正确率高达 93%. 本研究证明了设计的基于剪枝 CART 树的分类方法在人造板检测系统中应用的可行性和优势, 填补了我国人造板行业缺陷检测的空白, 推动人造板行业缺陷检测自动化发展, 为以后开展各类相似的缺陷检测提供了一种可行的方式.

参考文献

1 张国梁, 瞿国富, 侯晓鹏, 等. 连续平压机入口角度动态控制方法. 农业机械学报, 2015, 46(4): 372–378, 343.

2 杜超, 王志海, 江晶晶, 等. 基于显露模式的数据流贝叶斯分类算法. 软件学报, 2017, 28(11): 2891–2904. [doi: 10.13328/j.cnki.jos.005350]

3 Mu HH, Zhang MM, Qi DW, *et al*. The application of RBF neural network in the wood defect detection. International Journal of Hybrid Information Technology, 2015, 8(2): 41–50. [doi: 10.14257/ijhit]

4 梁浩, 曹军, 林雪, 等. 基于贝叶斯神经网络的近红外光谱实木地板表面缺陷检测. 光谱学与光谱分析, 2017, 37(7): 2041–2045.

5 Tan M, Hu ZF, Wang BY, *et al*. Robust object recognition via weakly supervised metric and template learning. Neurocomputing, 2016, 181: 96–107. [doi: 10.1016/j.neucom.2015.04.123]

6 Fan MY, Wei L, He ZZ, *et al*. Defect inspection of solder bumps using the scanning acoustic microscopy and fuzzy SVM algorithm. Microelectronics Reliability, 2016, 65: 192–197. [doi: 10.1016/j.microrel.2016.08.010]

7 黄新波, 李文君子, 宋桐, 等. 采用遗传算法优化装袋分类回归树组合算法的变压器故障诊断. 高电压技术, 2016, 42(5): 1617–1623.

8 常辉, 胡修林, 张蕴玉. 基于 CART 算法的卫星星座原始构型选择策略. 华中科技大学学报 (自然科学版), 2011, 39(6): 1–5.

9 张亮, 宁芊. CART 决策树的两种改进及应用. 计算机工程与设计, 2015, 36(5): 1209–1213.

10 Cui WC, Chen S, Yu TS, *et al*. Feature extraction of X-ray chest image based on KPCA. Proceedings of the 2nd International Conference on Computer Science and Network Technology. Changchun, China. 2013. 1263–1266.

11 杜春蕾, 张雪英, 李凤莲. 改进的 CART 算法在煤层底板突水预测中的应用. 工矿自动化, 2014, 40(12): 52–56.

12 Mburu JW, Kingwara L, Ester M, *et al*. Use of classification and regression tree (CART), to identify hemoglobin A1C (HbA1C) cut-off thresholds predictive of poor tuberculosis treatment outcomes and associated risk factors. Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, 2018, 11: 10–16. [doi: 10.1016/j.jctube.2018.01.002]

13 杨柳, 王钰. 泛化误差的各种交叉验证估计方法综述. 计算机应用研究, 2015, 32(5): 1287–1290, 1297. [doi: 10.3969/j.issn.1001-3695.2015.05.002]