江苏大学学报(自然科学版)
JOURNAL OF JIANGSU UNIVERSITY(Natural Science Edition)
2017 年 1 月 Jan. 2017
第 38 卷 第 1 期 Vol. 38 No. 1

doi: 10.3969/j.issn.1671 -7775.2017.01.015

基于 AdaBoost 集成学习的演化硬件 DNA 微阵列数据分类

王 进，黄 超，冉仟元，邓 欣，陈乔松
(重庆邮电大学 计算智能重庆市重点实验室，重庆 400065)

摘要： 为了更好地解决 DNA 微阵列数据的分类问题并进一步提高系统的识别率，提出了一种用于 DNA 微阵列数据分类的演化硬件多分类器 AdaBoost 选择性集成学习方法. 在系统集成阶段，介绍了 2 种改进的 AdaBoost 算法，分别探讨了以样本标记提升抽样有效容量和直接面向组合分类器分类精度提升的选择性集成策略. 对急性白血病、肺癌、结肠癌数据集进行了试验. 结果表明，基于 AdaBoost 集成学习的演化硬件方法对白血病、肺癌、结肠癌的平均识别率为 97.06%，99.32%，和 94.44%. 相对于传统演化硬件集成学习方法，文中方法保证更优识别率的同时有效降低了硬件实现代价.

关键词： 机器学习；演化硬件；DNA 微阵列；AdaBoost；选择性集成
中图分类号： TP391.4 **文献标志码：** A **文章编号：** 1671 -7775(2017)01 -0086 -07

引文格式： 王 进，黄 超，冉仟元，等. 基于 AdaBoost 集成学习的演化硬件 DNA 微阵列数据分类[J]. 江苏大学学报(自然科学版)，2017，38(1)：86 -92.

AdaBoost-based ensemble learning of evolvable hardware for classification of DNA microarray data

WANG Jin, HUANG Chao, RAN Qianyuan, DENG Xin, CHEN Qiaosong
(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: To solve the classification issues of DNA microarray data and enhance the recognition rate, an AdaBoost-based selective ensemble learning method was proposed with evolvable hardware (EHW) multiple classifiers. At the system ensemble stage, two improved AdaBoost algorithms were introduced. A sample labeling method was used to improve the effective capacity of sampling, and a selective ensemble strategy was employed to directly promote the classification precision of combined classifier. The experiments were completed on acute leukemia, lung cancer and colon cancer. The results show that the average accuracies of the proposed AdaBoost-based EHW for acute leukemia, lung cancer and colon cancer dataset are 97.06%, 99.32% and 94.44%, respectively. The proposed scheme achieves higher classification rate and lower hardware cost than the traditional EHW ensemble learning methods.

Key words: machine learning; evolvable hardware; DNA microarray; adaboost; selective ensemble

DNA 微阵列技术通过安装数千甚至上万个核酸探针于一个固体表面，能够在一次试验中同时分析成千上万的基因相关表达，使人们可以从分子水平了解生命本质. 近年来，DNA 微阵列技术已经被广泛地应用于生物医学上的癌症分子诊断、生物标记发现、分子靶向治疗等领域[1-2]. 相对于传统基于

收稿日期：2016 -02 -12
基金项目：国家自然科学基金资助项目(61203308，61309014，61403054)；重庆市基础与前沿研究计划项目(cstc2014jcyjA40001)；重庆教委科学技术研究项目(KJ1400436)
作者简介：王 进(1979—)，男，重庆人，教授(wangjin@cqupt.edu.cn)，主要从事机器学习与模式识别研究.
黄 超(1990—)，男，贵州遵义人，硕士研究生(ChinaHuangChao@hotmail.com)，主要从事机器学习与模式识别研究.

形态学的癌症分类方法，基于 DNA 微阵列数据的癌症分子分型技术能够有效地提高识别精度，辨别新的癌症亚型，促进临床治疗. 传统的模式识别方法如支持向量机[3-4]、人工神经网络[5]、决策树[6]、朴素贝叶斯[7]等已在癌症分子分型应用中表现出良好的分类效果. 然而上述方法大多采用软件实现的方式构建模型，受限于计算机本身的速度瓶颈和算法本身的时间复杂度，因此这些方法普遍存在着演化学习速度缓慢的局限. 同时，传统模式识别方法的学习结果通常是一些权值或回归系数表达，往往难以进行有效的分析，进而发现癌症的致病基因生物标记和探索它们之间的相互作用.

针对上述局限，近年来出现了一种基于 FPGA (field programmable gate array) 实现的内部演化硬件分类方法[8]. 该方法以其学习速度快(达到 ms 级)、数据处理能力强、演化结果电路易于分析等优点，目前已被广泛应用于声纳谱识别[9]、人脸识别[9]、肌电信号识别[10]等领域. 结合演化硬件分类器和选择性集成学习方法，目前文献[11]的研究已在急性白血病、结肠癌等 DNA 微阵列数据分类应用中取得了较好的效果. 但上述方法得到的结果电路结构复杂，需要较多的逻辑门和硬件资源开销.

为了进一步优化结果电路，降低分类器的硬件实现代价，文中拟在硬件演化过程中采用一种多目标适应值函数优化方法，将适应值函数评价准则分为 2 个层次，降低硬件代价和整个系统的时延；并提出 2 种改进的 AdaBoost 算法，直接面向组合分类器分类精度提升的选择性集成策略；通过对急性白血病[11]、肺癌[12]、结肠癌[2] 3 个 DNA 微阵列数据集进行试验，对分类器在采用多目标函数优化后的逻辑门平均使用数量、平均识别率、平均演化时间等性能指标与传统方法进行对比分析，以验证文中方法的有效性.

1　演化硬件分类系统

1.1　数据预处理与演化硬件分类器

DNA 微阵列数据具有特征维度高、样本数量少、非线性、高噪音、高相关冗余、数据分布不平衡等特点. 采用何种特征选择方法从这些数据中提取有用的信息基因，并使分类算法的计算复杂度控制在合理范围内，将直接关系到整个系统的性能. 文中采取基于信噪比(signal-to-noise ratio，SNR)的特征选择方法进行信息基因选择[11]，在处理样本数据时保留 n 位基因作为演化硬件基分类器的输入信息. 针对演化硬件分类器的 DNA 微阵列数据预处理流程详见文献[11].

与结构和功能一次性固定不可逆转的传统硬件电路相比较，演化硬件是一种基于可编程逻辑器件，通过应用演化算法能够自动地、动态地改变其自身结构和功能从而适应其周围环境变化的新型电子器件. 文中采用的 EHW 分类系统中，各基分类器的在线演化都是基于虚拟可重构结构(virtual reconfigurable architecture，VRA)[8,13]来实现的. 图 1 给出了一个演化硬件分类系统的总体框图.

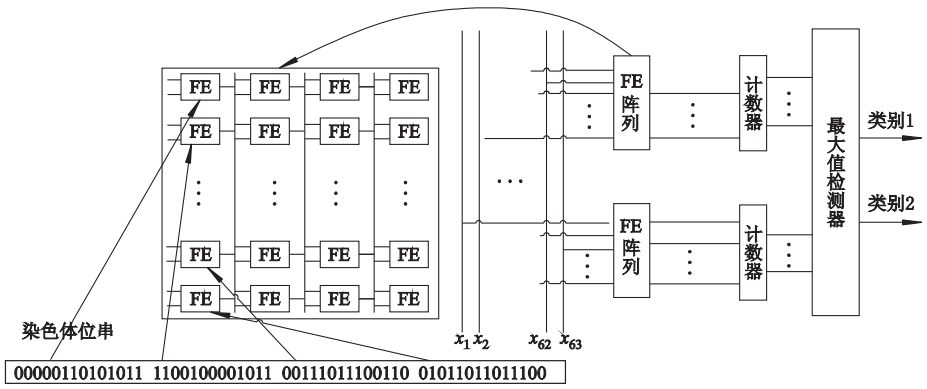

图 1　演化硬件分类系统总体框图

针对文中处理的癌症二分类问题，首先将经过前期预处理的 n 位信息基因(图 1 中 $n=64$)作为系统输入，通过数据总线并行地送到多个功能单元(function element，FE)阵列(每个 FE 阵列等同于一个基分类器)，经过 FE 阵列处理后把结果按类别分别输入 2 个计数器进行计数，最后通过最大值检测器进行比较，将输入样本识别为计数器输出值较大者所对应的类别. FE 阵列采用固定的 8×4 拓扑网

状结构，前一级联的输出作为后一级联的输入. 通过固定长度的染色体位串设定 FE 阵列中每个 FE 的功能和输入连接方式就可以控制整个电路的功能演化，关于演化硬件分类器基因型——表现型的解码过程在文献[8]中有详细的描述.

1.2 多目标适应值函数优化策略

文献[11]中的演化硬件多分类器在 FE 阵列演化过程中采用了一种 $1+\lambda$ 的演化策略，其优化目标为得到一个分类功能正确的组合逻辑电路. 然而这些演化结果电路通常结构比较复杂，在系统输入和输出之间包含较多的级联和逻辑门，一定程度上增加了硬件实现代价，降低了数据处理效率. 为了在实际应用中降低演化硬件分类系统的硬件实现代价，提高数据处理速度，文中在传统 $1+\lambda$ 演化策略（$\lambda=4$）[8]基础上引入了一种多目标适应值函数优化策略. 该策略前期仅用于组合逻辑电路演化设计领域[13]，但是在演化硬件分类器应用中，能否在保证基分类器分类正确的同时进一步优化其逻辑门使用数量，基分类器逻辑门使用数量的减少是否会影响组合分类器的分类性能等问题还缺乏相关文献进行探讨. 文中以减少基分类器逻辑门的使用数量为主要目标，着重优化少量基分类器逻辑门的识别能力以能够达到文献[11]相同的效果. 优化基分类器逻辑门会伴随着对其模型的适应值进行优化调整，同时适应值的优化也是多目标优化策略所考虑的次要目标. 2 个优化目标在整个演化硬件微阵列数据分类问题中相辅相成. 以下将着重介绍文中的优化策略层次.

文中所采用的适应值函数评价准则分为 2 个层次. 在第 1 层次中，选择设定真值表演化得到功能正确的分类器作为适应值评价准则. 评价函数用于评估候选电路的优良程度，通过读取电路的实际输出值与期望输出值进行比较，计算出对应个体的适应值，即为

$$F=\sum_{i=0}^{b}\sum_{j=0}^{a}(w_{ij}\otimes v_{ij}), \tag{1}$$

式中：a，b 分别为每个 FE 阵列的输出个数和输入样本个数；w_{ij}为第 i 个输入样本数据在 FE 阵列的第 j 个输出位信息（样本的分类器预测类别标签值）；v_{ij}为对应样本的期望输出值（样本的实际类别标签值）. 如果 w_{ij}和 v_{ij}的值相等，则适应值 F 加 1，否则 F 不变.

在第 2 层次中，可以选择逻辑门使用数、CMOS 门使用数或者以输入和输出之间级联所需要的染色体位串的长度等作为适应值评价准则. 但逻辑门使用数与 CMOS 门使用数一般成线性相关性（比如：1 个逻辑与门对应 6 个 CMOS 门，1 个逻辑或门对应 4 个 CMOS 门），而在设计 FE 阵列时，为了实现方便，采用了固定的拓扑网状结构，用固定长度的染色体位串来配置，所以在输入和输出之间为固定的长度，因此在第 2 层次中以逻辑门使用数作为适应值评价准则. 具体的执行过程如下：

1）保留在第 1 层次中满足电路功能完全正确条件的个体；

2）对步骤 1）得到的个体按预先设定的突变率（文中采用 0.8% 的固定突变率）进行突变操作；

3）保留突变操作后电路功能完全正确的个体，统计每个个体转化为电路后逻辑门的使用数量；

4）选择逻辑门使用数最少的个体继续执行步骤 2），直至迭代结束.

步骤 2）中的突变率根据数据集不同而调整，对于文中的数据集，如果突变率过快经过几次迭代后特征产生很大差异，因此会直接影响结果. 多次试验结果显示设置为 0.8% 最合理. 试验迭代执行步骤 2）至步骤 4）直至满足终止条件，试验中演化算法的终止条件为完成了预定的演化代数.

2 基于 AdaBoost 的选择性集成

2.1 AdaBoost 算法

AdaBoost 算法是一种具有代表性的集成学习方法[14]. 文中多处涉及到与它的对比分析及推广，故先简单介绍 AdaBoost 算法. 设实例空间为 X，训练样本集为 $D=\{(x_1,c_1),(x_2,c_2),\cdots,(x_m,c_m)\}$，此处 $x_i\in X$，文中只考虑二分类问题，此时 $c_i\in\{-1,+1\}$. 设每轮训练使用的训练集为 $[D_t(t=1,2,\cdots,T)]$，后续试验中选择的基分类器的个数为 5，故此处 $T=5$，在 D_t 上用演化硬件分类方法训练生成分类器：$h_t(x)$：$x\to(-1,+1)$. 函数 $[h_t(x_i)\neq c_i]$ 根据是否满足条件输出 0 或 1. AdaBoost 算法中，通过 D_t 训练得到带不同权值的基分类器 $h_t(x)$，强分类器由 T 个弱分类器加权投票，分类错误率小的弱分类器赋予更大的投票权重.

AdaBoost 算法步骤如下：① 初始化样本权值：$w_i^1=1/m(i=1,2,3,\cdots,m)$；② 使用学习算法，基于样本权值 w_i^t 对样本集抽样，训练得到分类器：$h_t(x)$：$x\to(-1,+1)$；③ 计算 $h_t(x)$ 的错误率：

$\varepsilon_t = \sum_{i=1}^{m} w_i^t [h_t(x_i) \neq c_i] / \sum_{i=1}^{m} w_i^t$; ④ 调整样本权值: $w_i^{t+1} = w_i^t \times \begin{cases} e^{-\alpha_t}, & h_t(x_i) = c_i, \\ e^{\alpha_t}, & h_t(x_i) \neq c_i; \end{cases}$ ⑤ 循环执行 T 次步骤②－④,以训练模型; ⑥ 循环结束后,最后的强分类器为 $H(x) = \mathrm{sgn}(f(x))$,其中 $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$.

2.2　基于样本标记的改进 AdaBoost 算法

传统 Adaboost 算法在每轮迭代之后,易分类样本的权值都会被更新为小于 $1/m$ 的数值. 但在每轮迭代学习过程中,每个易分类样本依然具有一定被选入训练子集的概率. 为了避免易分样本的重复抽取,增加训练子集的有效容量,对 Adaboost 算法做以下改进.

设样本集$\{(x_i, c_i, l_i) \mid i = 1, 2, \cdots, m\}$ 包含 m 个训练样本,c_i 为类别标签,l_i 为选择标签,基于样本标记的改进 AdaBoost 算法(AdaBoost_M1)步骤如下: ① 初始化: $w_i^1 = 1/m$, $l_i = 1 (i = 1, 2, \cdots, m)$; ② 根据样本权值 w_i^t,通过对原始训练集中满足 $l_i = 1$ 的样本进行抽样产生训练集 D_t,使用学习算法,基于训练集 D_t 训练得到分类器 $h_t(x)$: $x \rightarrow (-1, +1)$; ③ 计算 $h_t(x)$ 的错误率: $\varepsilon_t = \sum_{i=1}^{m} w_i^t [h_t(x_i) \neq c_i] / \sum_{i=1}^{m} w_i^t$; ④ 调整样本权值: $w_i^{t+1} = w_i^t \times \begin{cases} e^{-\alpha_t}, & h_t(x_i) = c_i, \\ e^{\alpha_t}, & h_t(x_i) \neq c_i, \end{cases}$ 更新每个样本的权值,对于 $w_i^{t+1} < \frac{X}{m}$ 的样本,置 $l_i = 0$; ⑤ 迭代执行 T 次步骤②－④,以训练模型; ⑥ 循环结束后,最后的强分类器为 $H(x) = \mathrm{sgn}(f(x))$,其中 $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$.

样本可选标签 l_i 标示样本在抽样时是否可被选取的状态. 算法初始化后,原始训练集中所有样本的可选标签 l_i 均被置为 1,即在第 1 次对训练集进行抽样时,所有样本均有同样的概率被选中. 在每轮迭代过程中,一个样本的权值一旦被更新为小于 $1/m$ 的数值,则其对应的可选标志位 l_i 被置为 0,表示在此轮样本抽取过程中该样本不能被抽中; 即在迭代过程中,只有可选标志 l_i 为 1 的样本可被抽样构成训练子集. 相对于传统 AdaBoost 算法,基于样本标记的改进 AdaBoost 算法避免了对于那些权值小于 $1/m$ 的易分样本的重复抽样,提高了算法的收敛性. 同时训练样本数目的减少和迭代过程中不同训练样本子集的构建可降低分类器的学习时间开销和增大基分类器之间的差异性,有助于获得更好的集成效果.

2.3　面向分类精度快速提升的改进 AdaBoost 算法

基于样本标记的改进 AdaBoost 算法主要从增加抽样样本有效容量和训练样本子集之间的差异性着手,控制集成分类系统中基分类器的特性,而在进行基分类器集成的时候都是以基分类器错误率最小为选择依据. 在演化硬件分类器集成学习过程中,根据文献[15]的理论分析结果,文中尝试实现了一种面向集成分类器分类准确率提升最大的改进 AdaBoost 算法(AdaBoost_M2).

集成分类器能解决单个分类器的训练数据量小、假设空间小、局部最优这 3 个问题,所以,集成分类器的预测能力会优于单个分类器的预测能力. 文中采取分类器集成来提升组合分类器的分类正确率,即满足如下条件的样本数量增多[15]:

$$\sum_{t: h_t(x_i) = c_i} \alpha_t - \sum_{t: h_t(x_i) \neq c_i} \alpha_t > 0. \qquad (2)$$

在每一轮学习后选取新的基分类器时,以满足式(2)的样本比率增加最大为依据,而不是传统的选择训练错误率 ε_t 最小为依据.

AdaBoost_M2 中的 θ_t,实际上是前 t 个分类器组合得到的集成分类器对于样本集的分类正确率,根据 $\theta_t - \theta_{t-1}$ 最大化来训练和选取 $h_t(x)$,体现了直接面向目标(最大化提升 $H(x)$ 的分类精度)的思路,并且错误率采用了真正的训练错误率 ε_t'.

AdaBoost_M2 算法步骤如下: ① 初始化: $w_i^1 = 1/m$, $\delta_i^0 = 0 (i = 1, 2, \cdots, m)$, $\theta_0 = 0$; ② 使用学习算法,基于样本权值 w_i^t 对样本集抽样,训练得到分类器 $h_t(x)$: $x \rightarrow (-1, +1)$; ③ 计算 $h_t(x)$ 的错误率: $\varepsilon_t' = \sum_{i=1}^{m} [h_t(x_i) \neq c_i] / m$; 令 $\beta_t = \varepsilon_t' / (1 - \varepsilon_t')$, $\alpha_t = \frac{1}{2} \mathrm{lb}(1/\beta_t)$; ④ 计算样本的组合累计权值: $\delta_i^t = \begin{cases} \delta_i^{t-1} + \alpha_t, & h_t(x_i) = c_i, \\ \delta_i^{t-1} - \alpha_i, & h_t(x_i) \neq c_i; \end{cases}$ ⑤ 统计组合累计权值大于零的样本比率 $\theta_t = \sum_{i=1}^{m} [\delta_i^t > 0] / m$; 根据 $\theta_t - \theta_{t-1}$ 最大化原则选取 $h_t(x)$; ⑥ 迭代执行 T 次步骤②－⑤,以训练模型; ⑦ 最后的强分类器 $H(x) = \mathrm{sgn}(f(x))$,其中:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x).$$

3 试验结果与分析

为验证2种基于AdaBoost改进算法的演化硬件分类器性能，以及多目标适应值函数优化策略使用前后对演化时间、识别率、硬件代价等方面的影响，采用了急性白血病[1]、肺癌[12]、和结肠癌[2]这3个公开的DNA微阵列数据集进行对比试验，表1给出了3个数据集的相关信息.

硬件环境采用Celoxica公司的RC1000板卡，演化硬件分类系统在Xilinx ISE6.3开发环境下使用VHDL语言设计，仿真实现后通过PCI总线接口下载程序到Virtex. 在xcv2000E FPGA芯片中执行在线演化，得到最终的组合分类器. 因演化硬件分类器的FPGA实现过程已在文献[8]有了详细说明，此处不再重点阐述.

表1 3个DNA微阵列数据集信息表

数据集	训练集数目	测试集数目	基因个数
急性白血病	38	34	7 129
肺癌	32	149	12 533
结肠癌	38	24	2 000

在选择性集成5个基分类器的固定条件下，选择的信息基因位数不同会对演化硬件分类系统识别率造成影响. 图2给出了试验结果.

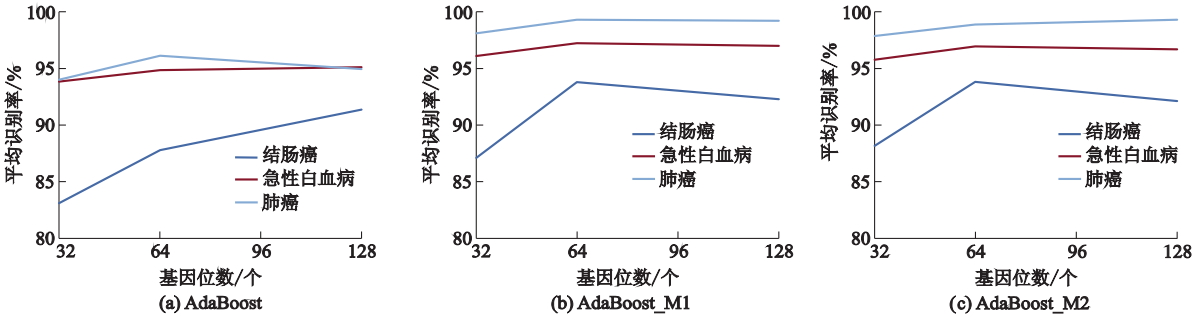

图2 不同信息基因位数下3种AdaBoost集成方法对3个癌症数据集的测试集平均识别率

如图2所示，AdaBoost_M1和AdaBoost_M2方法对急性白血病、肺癌、结肠癌数据集的识别率会随着信息基因数的增加而增加；但达到一定程度后（$n=64$），对急性白血病和结肠癌的识别率反而会有所下降，这有可能是由于过多的信息基因将导致冗余和噪声基因的增加，反而降低了演化硬件分类器识别率. 试验过程中，信息基因位数被限制到128位及以下，主要是因为多路选择器的硬件实现代价过大，无法在选定的Virtex xcv2000E FPGA芯片上实现有效处理128位以上的演化硬件分类器.

由图2可见，相比传统的AdaBoost算法，AdaBoost_M1和AdaBoost_M2算法在3种DNA微阵列数据集表现更加优秀，在相同特征维度的数据集上的识别率都有明显的增加，这说明AdaBoost_M1和AdaBoost_M2模型更加适应于高噪声，数据样本维度高，样本数量较少的数据集. 文中提出的优化策略提高了传统AdaBoost算法冗余数据和噪声数据的泛化能力. AdaBoost_M1相较于AdaBoost_M2性能有一些差异，主要体现在对Colon数据集训练过程中产生的误差. 经过对Colon数据集的分析，可发现该数据集样本间的特征取值区间比较大. 由于AdaBoost_M1算法在更新样本特征权值时太过于绝对（对小于阈值的权值都置为0，样本权重被视为均等所以阈值设为$1/m$），使其在学习数据集样本间取值区间较大的数据时就会出现不应该被置0的权值. 此外，试验结果显示AdaBoost_M2算法就避免了类似的问题，且表现良好.

表2-4给出了在50次试验中，使用多目标适应值函数优化（post_MO）和未使用多目标适应值函数优化（pre_MO）策略时3种AdaBoost集成学习算法在3个癌症数据集上的系统平均演化代数、系统独立测试集平均识别率、基分类器逻辑门平均使用数、和系统平均演化时间对比.

由表2-4可见，多目标适应值函数优化策略的引入在时间开销可以接受的条件下（所有引入多目标适应值函数优化策略的分类器演化时间都在0.4 s以内），很明显地减少了演化硬件基分类器的逻辑门使用数量，在控制分类电路结构复杂度和降低硬件代价方面具有很明显的效果. 表2-4的结果证明使用多目标适应值函数优化策略减少基分类器逻辑门使用数量并不显著影响集成后组合分类器的分类效果.

表 2　pre_MO 和 post_MO 策略下 3 种集成学习方法对急性白血病数据集的学习结果对比

试验方法	pre_MO				post_MO			
	演化代数	识别率/%	逻辑门/个	演化时间/ms	演化代数	识别率/%	逻辑门/个	演化时间/ms
AdaBoost	294	96.64	17.9	5.7	213(8 192)	95.98	1.8	159.0
AdaBoost_M1	196	99.32	21.4	1.2	213(8 192)	99.32	1.2	49.6
AdaBoost_M2	301	98.66	20.1	5.8	213(8 192)	99.32	1.6	159.0

表 3　pre_MO 和 post_MO 策略下 3 种集成学习方法对肺癌数据集的学习结果对比

试验方法	pre_MO				post_MO			
	演化代数	识别率/%	逻辑门/个	演化时间/ms	演化代数	识别率/%	逻辑门/个	演化时间/ms
AdaBoost	402	95.56	18.4	9.3	214(16 384)	94.70	1.7	377.5
AdaBoost_M1	259	97.06	21.5	1.9	214(16 384)	97.06	1.5	119.2
AdaBoost_M2	386	97.06	19.9	8.9	214(16 384)	96.76	1.9	377.5

表 4　pre_MO 和 post_MO 策略下 3 种集成学习方法对结肠癌数据集的学习结果对比

试验方法	pre_MO				post_MO			
	演化代数	识别率/%	逻辑门/个	演化时间/ms	演化代数	识别率/%	逻辑门/个	演化时间/ms
AdaBoost	4 288	90.54	22.5	98.8	214(16 384)	89.13	2.3	377.5
AdaBoost_M1	2 147	93.38	22.0	15.6	214(16 384)	94.12	1.4	137.6
AdaBoost_M2	3 899	94.44	21.9	89.8	214(16 384)	94.44	1.8	377.5

对于 3 种 AdaBoost 集成学习算法，在集成 5 个基分类器的条件下，改进的 AdaBoost_M1 和 AdaBoost_M2 这 2 种算法在 3 种数据集上都取得了比原始 AdaBoost 算法更高的分类准确度. 对学习同样的癌症数据集，AdaBoost_M1 通过添加样本选择标签，在每轮训练中避免对识别正确的样本进行重复采样学习，在降低抽样样本整体容量的同时也显著减少了系统的学习时间和演化代数. 而对于 AdaBoost_M2 算法，文中用试验验证了文献[15]提出理论的有效性.

表 5-7 列出了文中方法与其他传统模式识别方法在急性白血病、肺癌、结肠癌数据集下的独立测试集平均识别率对比. 在系统演化时间不超过0.4 s，硬件资源消耗很低的情况下，文中方法仍然具有一定的可比性，同时验证了基于 AdaBoost 选择性集成的演化硬件分类方法在对 DNA 微阵列数据分类中的有效性.

如表 5-7 所示，AdaBoost 集成学习演化硬件的识别率在肺癌数据集中与最优识别率(FG-HN)只相差 0.08%，在急性白血病数据集中与最优识别率(SVM-RFE + MRMR)相差 1.29%，在结肠癌数据集中与最优识别率(FH-HN)相差 0.56%. 可见，文中提出的 AdaBoost 集成学习的演化硬件能够在 3 组 DNA 微阵列数据试验中保持分类性能. 同时结合表 2-4 分析得出，采用了多目标适应值函数优化策略优化的 AdaBoost 演化硬件方法减少了其基分类器逻辑门的数量. 由此验证了文中的假设，即采用多目标适应值函数优化策略进行优化的 AdaBoost 演化硬件方法，能够减少基分类器逻辑门的数量而不影响其分类性能; 同时也验证了伴随基分类器逻辑门数量的减少，适应值优化也起到了很大的作用，但相比传统 AdaBoost 性能提升明显.

表 5　不同分类方法对肺癌数据集的测试集平均识别率对比

试验方法	识别率/%	参考文献
AdaBoost + EHW	96.64	本文
AdaBoost_M1 + EHW	99.32	本文
AdaBoost_M2 + EHW	99.32	本文
FG-HN	99.40	[16]
Bagging-C4.5	93.29	[17]

表 6　不同分类方法对结肠癌数据集的测试集平均识别率对比

试验方法	识别率/%	参考文献
AdaBoost + EHW	90.54	本文
AdaBoost_M1 + EHW	94.12	本文
AdaBoost_M2 + EHW	94.44	本文
EHW 选择性集成	88.33	[11]
FG-HN	95.00	[16]
SVM-RFE + MRMR	91.68	[3]
KNN	83.90	[5]

表7 不同分类方法对急性白血病数据集的测试集平均识别率对比

试验方法	识别率/%	参考文献
AdaBoost + EHW	95.56	本文
AdaBoost_M1 + EHW	97.06	本文
AdaBoost_M2 + EHW	97.06	本文
EHW 选择性集成	95.42	[11]
FG-HN	95.60	[16]
SVM-RFE + MRMR	98.35	[3]
Bagging-C4.5	91.18	[17]
GCM + CCM	97.10	[18]
NN	97.10	[5]

4 结 论

1) 基于改进 AdaBoost 集成学习方法通过在传统演化硬件分类器中引入多目标适应值函数优化策略,以得到功能正确的结果电路和逻辑门使用数量最少作为适应值评价准则,在保证组合分类器识别率的同时,有效减少了硬件资源消耗.

2) 据试验 AdaBoost_M1 和 AdaBoost_M21 方法显示,适应值的更新方式不同对性能影响也有所不同. 提升演化硬件分类系统学习效率的同时,提高了整个系统的分类性能. 通过与其他传统模式识别方法在 3 个 DNA 微阵列数据集下的对比试验,验证了文中方法的有效性.

参考文献(References)

[1] MARISA L, DE REYNIÈS A, DUVAL A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value [J]. Plos Medicine, 2013, 10(5), article number: e1001453.

[2] WIŚNIEWSKI J R, OSTASIEWICZ P, DUŚ K, et al. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma [J]. Molecular Systems Biology, 2012, doi: 10.1038/msb.2012.44.

[3] MUNDRA P A, RAJAPAKSE J C. SVM-RFE with MRMR filter for gene selection [J]. IEEE Transactions on Nanobioscience, 2010, 9(1): 31 - 37.

[4] MAULIK U, CHAKRABORTY D. Fuzzy preference based feature selection and semisupervised SVM for cancer classification [J]. IEEE Transactions on Nanobioscience, 2014, 13(2): 152 - 160.

[5] FERNÁNDEZ-NAVARRO F, HERVÁS-MARTÍNEZ C, RUIZ R, et al. Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection [J]. Applied Soft Computing, 2012, 12(6): 1787 - 1800.

[6] CHEN K H, WANG K J, TSAI M L, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm [J]. BMC Bioinformatics, 2014, doi: 10.1186/1471-2105-15-49.

[7] WIN S L, HTIKE Z Z, YUSOF F, et al. Cancer recognition from DNA microarray gene expression data using averaged one-dependence estimators [J]. International Journal on Cybernetics & Informatics, 2014, 3(2): 1 - 10.

[8] SALVADOR R, OTERO A, MORA J, et el. Self-reconfigurable evolvable hardware system for adaptive image processing [J]. IEEE Transactions on Computers, 2013, 62(8): 1481 - 1493.

[9] GLETTE K, KAUFMANN P. Lookup table partial reconfiguration for an evolvable hardware classifier system [C] // Proceedings of the 2014 IEEE Congress on Evolutionary Computation. Piscataway: IEEE, 2014: 1706 - 1713.

[10] KAUFMANN P, GLETTE K, GRUBER T, et al. Classification of electromyographic signals: comparing evolvable hardware to conventional classifiers [J]. IEEE Transactions on Evolutionary Computation, 2013, 17(1): 46 - 63.

[11] 王进, 陈文, 冉仟元, 等. 用于微阵列数据癌症分类的演化硬件多分类器 [J]. 江苏大学学报(自然科学版), 2013, 34(4): 410 - 415.
WANG J, CHEN W, RAN Q Y, et al. Multiple classifiers based on evolvable hardware for cancer classification with microarray data [J]. Journal of Jiangsu University(Natural Science Edition), 2013, 34(4): 410 - 415. (in Chinese)

[12] FUKUTA K, OKADA Y. Informative gene discovery in DNA microarray data using statistical approach [C] // Proceedings of the 2011 International Conference on Advances in Intelligent Control and Innovative Computing. Heidelberg: Springer Verlag, 2012: 377 - 394.

[13] WANG J, LEE C H. Virtual reconfigurable architecture for evolving combinational logic circuits [J]. Journal of Central South University, 2014, 21(5): 1862 - 1870.

[14] DUDA D, KRETOWSKI M, BÉZY-WENDLING J. A computer-aided diagnosis of liver tumors based on multi-image texture analysis of contrast-enhanced CT. selec-

(下转第 102 页)

and Building Materials, 2002, 16(8): 473 – 487.

[8] LEE S J, HU J, KIM H, et al. Aging analysis of rubberized asphalt binders and mixes using gel permeation chromatography [J]. Construction and Building Materials, 2011, 25(3): 1485 – 1490.

[9] 原健安. 改性剂对沥青粘结性的影响 [J]. 西安石油学院学报, 1999, 14(3): 51 – 54.
YUAN J A. Effect of modifier on the cohesiveness of asphalt [J]. Journal of Xi'an Petroleum Institute, 1999, 14(3): 51 – 54. (in Chinese)

[10] 延西利, 梁春雨. 沥青与石料间的剪切粘附性研究 [J]. 中国公路学报, 2001, 14(4): 25 – 27.
YAN X L, LIANG C Y. Study of the shear adhesiveness between bitumen and rock [J]. China Journal of Highway and Transport, 2001, 14(4): 25 – 27. (in Chinese)

(责任编辑 赵 鸥)

(上接第92页)

tion of the most appropriate texture features [J]. Studies in Logic, Grammar and Rhetoric, 2013, 35(1): 49 – 70.

[15] 冷强奎. 组合凸线器框架下分片线性分类器的构造方法研究 [D]. 北京: 北京工业大学, 2015.

[16] 王进, 张军, 胡白帆. 结合最优类别信息离散的细粒度超网络微阵列数据分类 [J]. 上海交通大学学报, 2013, 47(12): 1856 – 1862.
WANG J, ZHANG J, HU B F. Optimal class-dependent discretization-based fine-grain hypernetworks for classification of microarray data [J]. Journal of Shanghai Jiao Tong University, 2013, 47(12): 1856 – 1862. (in Chinese)

[17] PIAO Y J, PIAO M H, PARK K, et al. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data [J]. Bioinformatics, 2012, 28(24): 3306 – 3315.

[18] 卢新国, 林亚平, 骆嘉伟, 等. 癌症识别中一种基于组合GCM和CCM的分类算法 [J]. 软件学报, 2010, 21(11): 2838 – 2851.
LU X G, LIN Y P, LUO J W, et al. Classification algorithm combined GCM with CCM in cancer recognition [J]. Journal of Software, 2010, 21(11): 2838 – 2851. (in Chinese)

(责任编辑 梁家峰)