高技术通讯 2013 年 第23 卷 第12 期:1236 ~ 1241

doi:10.3772/j.issn.1002-0470.2013.12.004

Bagging 选择性集成演化硬件 DNA 微阵列数据分类方法①

王 进② 冉仟元 丁 凌 赵 蕊

(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

摘 要 为了提高演化硬件(EHW)分类系统的泛化能力和减少硬件代价,提出了一种用于 DNA 微阵列数据分类的演化硬件多分类器选择性集成学习方法。重点讨论了基于 Bagging 的选择性集成学习策略和基于虚拟可重构结构的演化硬件分类系统构架。通过对原始数据训练集的随机重采样生成训练子集完成对演化硬件基分类器的训练,并选择其中识别率较高的基分类器进行集成以获得更高的分类性能。演化硬件分类系统对 DNA 微阵列数据的学习与分类均在 Xilinx Virtex xcv2000E FPGA 硬件平台上实现。通过对急性白血病和肺癌数据集的对比实验表明:相对于传统演化硬件集成学习方法,这种方法在保证较高识别率的基础上有效降低了硬件代价,且具有更短的学习时间和较强的泛化能力。

关键词 演化硬件(EHW),Bagging,DNA 微阵列,选择性集成

0 引 言

相对于主要依据症状、体征、影像检查、组织细胞病理等形态学信息进行癌症诊断的传统方法,DNA 微阵列数据分析从基因分子水平进行疾病性状的定义,研究肿瘤的生长机理,在致癌基因识别、癌症诊断等方面能够快速、经济,诊断可靠,避免癌症诊断的侵犯性检查。自 1999 年 Golub 等人在白血病 DNA 微阵列数据上成功进行癌症分类以来[1],基于 DNA 微阵列技术的癌症分类研究已经被人们广泛接受并日益成为生物信息学研究的热点之一。基于 DNA 微阵列技术的癌症分类方法有多种,但其性能各不相同。本文引入了 Bagging 集成学习和演化硬件(evolvable hardware,EHW)相结合的分类方法,该方法克服了以往方法的局限,性能有明显提高。

1 相关工作

在 DNA 微阵列数据癌症分类应用中,多种不同的传统模式识别方法如禁忌搜索[2]、遗传规划[3]、贝叶斯网络[4,5]、决策树[6]、神经网络集成[7]等已经被广泛研究和使用,取得了较好的分类效果。然而,传统的模式识别技术存在学习速度慢、学习结果可读性差、不易分析等局限。针对上述局限,Torressen[8]、王进[9]等人提出了基于演化硬件(EHW)的分类方法。EHW 方法通过在 FPGA 上实现基于笛卡尔遗传规划(Cartesian genetic programming,CGP)[10]的硬件演化,在获得与其他传统分类方法相似分类性能的同时,提高了系统的学习速度和学习结果的可读性。近年来,EHW 分类方法已被广泛应用于声纳谱识别[8]、人脸识别[11]、字符识别[12]、道路限速标志识别[13]等领域。

在 DNA 微阵列数据分类研究方面,王进等提出了一种基于虚拟可重构结构的 EHW 分类技术[14],并在急性白血病 DNA 微阵列数据分类应用中取得了较好的效果。但是,该方法也存在以下不足:(1)单一 EHW 基分类器识别率不高;(2)通过集成大量 EHW 基分类器,可以提高对 DNA 微阵列数据的识别率,但带来了过大的硬件实现代价;(3)通过多次学习获得的 EHW 分类系统分类性能差异较

① 国家自然科学基金(61203308,61075019),教育部留学回国人员科研启动基金(教外司留[2010]1174 号)和重庆市大学生创新创业训练计划(201210617003)资助项目。

② 男,1979 年生,博士,教授;研究方向:演化计算,模式识别,智能信息处理;联系人,E-mail:wangjin_liips@yahoo.com.cn

(收稿日期:2012-12-30)

大,稳定性不足。

本文针对上述 EHW 分类方法在 DNA 微阵列数据分类中的局限,提出了一种基于 Bagging 选择性集成的 EHW 分类方法。主要通过引入 Bagging 集成学习方法增加基分类器间的差异性,同时结合选择性集成策略,在提高 EHW 分类系统泛化能力和稳定性的同时,能有效降低系统硬件实现代价、减少系统学习时间。本文选取麻省理工学院提供的急性白血病数据集[1]和哈佛医学院提供的肺癌数据集[15]进行实验。通过对数据集进行特征选择和规格化等预处理,基于 Bagging 选择性集成策略对演化学习后的 EHW 基分类器进行集成建立 DNA 微阵列数据分类系统,对分类系统的时间开销、硬件代价、平均识别率、识别率方差等性能指标进行对比分析,验证了本文方法的有效性。

2　DNA 微阵列数据预处理

对 DNA 微阵列数据的预处理主要包括特征选择和数据规格化处理两个步骤。特征选择的主要目标是去除微阵列数据中的冗余和含噪音基因,降低分类算法的计算复杂度。目前,常见的信息基因选择方法主要分为两类:基于过滤(filter)和基于打包(wrapper)的特征选择。从机器学习角度,过滤方法独立于分类算法,以数据的内在属性作为特征的评价准则,计算复杂度低。为了便于 EHW 系统的硬件实现,提高系统学习速度,本文采用基于过滤的信噪比(signal-to-noise ratio,SNR)方法[1]进行信息基因选择:

$$P(g,c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \tag{1}$$

其中,$\mu_1(g)$、$\mu_2(g)$ 分别表示基因在类型 1 和类型 2 中的平均值;$\sigma_1(g)$、$\sigma_2(g)$ 分别表示基因在类型 1 和类型 2 中的标准差。式(1)中 $P(g,c)$ 绝对值越大,则基因对分类的相关性越好。根据求得的 $P(g,c)$ 值,我们从正值和负值中分别选择绝对值较大的 $n/2$ 个基因(本文中 $n=32$)作为信息基因。采用式(2)对 n 个信息基因进行规格化处理:

$$nor_n = \frac{e_m - g_{ave_i}}{g_{SD_i}} \tag{2}$$

其中 e_m 是所选中的基因在样本 m 中的表达水平,g_{ave_i} 表示选中的基因 g_i 在训练集所有样本中的平均表达水平,g_{SD_i} 表示选中的基因 g_i 在训练集所有样本中的标准差。求出归一值之后,进行二值化处理,即如果 $nor_n \geqslant 0$,把该值定义为 1;否则定义为 0。

3　基于 Bagging 选择性集成的演化硬件分类系统

与传统结构和功能固定、设计完成后不可更改的硬件电路相比,EHW 是一种基于可重构结构,通过引入学习算法能够自动且动态地改变自身结构和功能的新型电子器件[16-18]。EHW 分类系统基于可重构逻辑器件的高效、快速等特性,具有可在线适应、实时性强,学习结果可分析性好,识别速度快等特点[8,14]。

3.1　演化硬件分类系统

EHW 分类系统的总体架构如图 1 所示,自上而下主要分为 4 个层次:EHW 分类系统、种类识别器、功能单元阵列(EHW 基分类器)、功能单元(function element,FE)。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)[12]实现。VRA 是一种基于商业 FPGA 的快速可重构平台,具有重构速度快、通用性强、易于实现等特点,为 EHW 分类系统提供了一个更为简单有效的内进化技术途径。

图 1(a)为分类系统的顶层结构图。考虑到急性白血病和肺癌数据识别均为两类分类问题,因此 EHW 分类系统中仅使用了 2 个种类识别器。经过预处理的 32 位信息基因作为系统输入,通过数据总线并行送到 2 个种类识别器;经种类识别器处理后的结果输入到最大值检测器进行比较,并将输入样本识别为种类识别器输出值较大者所对应的类别。

图 1(b)中的种类识别器由多个 FE 阵列和一个计数器构成,通过数据总线输入到种类识别器的信息基因并行到达 k 个 FE 阵列的输入端口。通过 FE 阵列执行相应的功能后,每个 FE 阵列的输出分别送到该类的计数器,以计算该种类识别器中所有 FE 阵列输出值之和。

本文采用的 FE 阵列结构如图 1(c),是一个包含 8 行 4 列的 FE 二维网络。如图 1(d)所示,每个 FE 由 2 个多路选择器、1 个功能模块和 1 个 D 型触发器组成。对于第 1 列 FE,其输入端口连接到经过预处理的 32 位信息基因,其输出连接到下一列 FE 的输入。其他列 FE 的输入端口连接到前一列的输出上,而输出端口通过 FE 功能模块执行相应的逻辑组合后提供给下一列 FE 作为输入。在演化过程中,通过遗传算法(evolutionary algorithm,EA)产生

的染色体配置系统的功能电路。染色体的编码决定每个 FE 单元的执行功能和整个 FE 阵列的连接方式。例如:对于图 1(d)中的功能单元 FE_{12},其染色体编码为(101,111,100),那么它的第 1 个输入端口连接第 1 列 FE_{61} 的输出 a,第 2 个输入端口连接 FE_{81} 的输出 b,输出端口为 a、b 执行功能 4(a and b)逻辑后的输出。

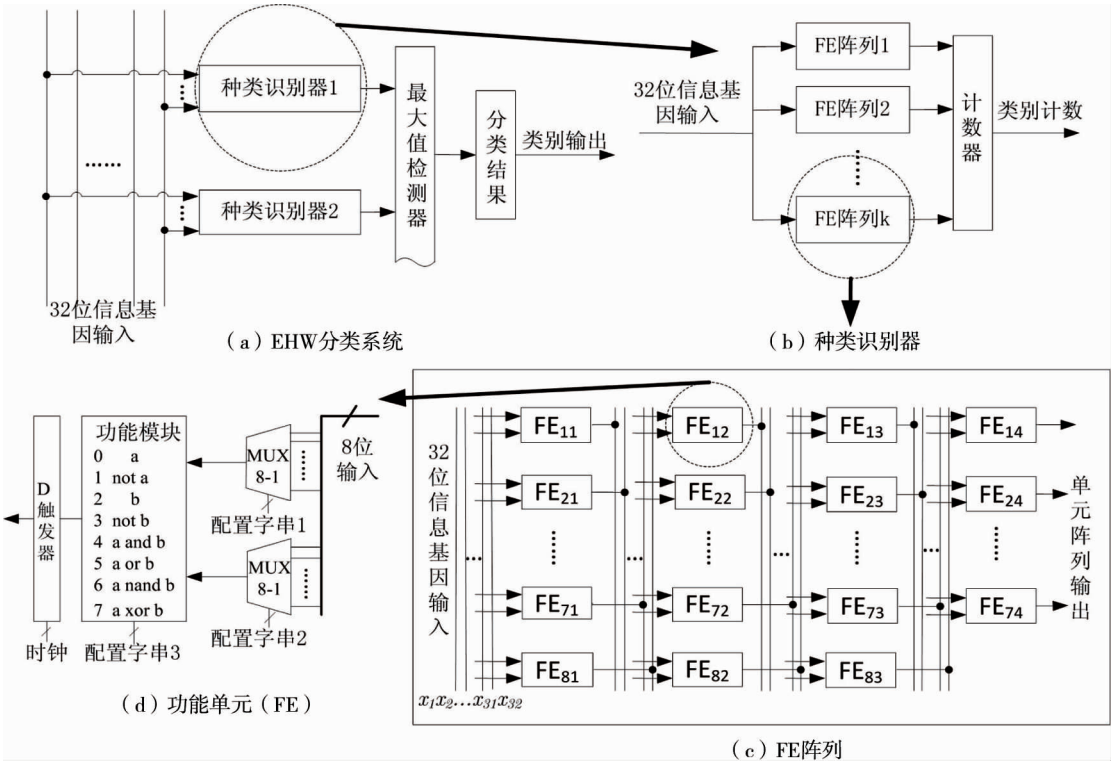

图 1 EHW 分类系统总体框图

3.2 演化算法

每个 FE 阵列的演化均采用 1 + λ 演化策略,(λ =4)。演化算法具体的执行过程为:

步骤(1):随机产生一个初始种群,种群中含有 λ 个个体。

步骤(2):根据适应值评价函数[14]计算种群中每个个体的适应值。

步骤(3):根据适应值选出最优个体。

步骤(4):对最优个体按 0.8% 的突变率对选中的位进行变异操作,生成最优个体的 λ 个突变体。

步骤(5):将最优个体与 λ 个变异体组成新的 1 + λ 个个体的种群。转至步骤(2),重复执行步骤(2)—(5)直至满足终止条件。

在实验中,EA 算法的终止条件定义为达到预定最大迭代次数:16777216 代,或者达到最大适应值。

3.3 Bagging 选择性集成

集成学习是一种能够有效提高分类系统泛化能力的方法,其基本思想是:在原始数据集上构建多个尽可能相互独立的基分类器,通过对多个分类器结果进行某种组合来决定最终的分类[19]。对集成学习效果的影响主要包括三方面:(1)基分类器之间的差异度;(2)基分类器的识别率;(3)系统的集成方法。

本文设计了一种基于 Bagging[20] 的 EHW(Bagging-EHW)选择性集成学习方法。在基分类器学习阶段,为保证每个样本子集的差异性和最大相关性,采用 Bagging 算法的重采样技术对经预处理后的训练集进行重复随机采样,然后用 EHW 分类方法对子训练集 S_i 进行训练得到基分类器。由于输入样本序列的改变,EHW 的演化结果会产生很大的变化,所以通过控制不同的子训练集样本的输入,将有效增大基分类器的差异,这有助于获得更好的集成效果。另一方面,由于重复随机抽样只选择原样本空间中一个局部进行基分类器训练,也降低了 EHW 算法的复杂度和时间开销。在分类器集成阶段,本文基于选择性集成的思想[21],仅从训练完毕的 EHW 基分类器序列中选择部分性能较好的基分类器进行集成。为了评价演化硬件基分类器(即 FE 阵列)的性能,本文把 FE 阵列的误差值作为评价函数 M。对一个训

练集样本序列而言,FE 阵列的实际输出值和理想输出值之差绝对值的和,即为该 FE 阵列的误差值:

$$error = \sum_{i=1}^{N} | ideal - practical | \quad (3)$$

式(3)中,*ideal* 是每个样本在 FE 阵列的理想输出值(其中样本数据在自己对应的种类识别器中 FE 理想输出值为“1111111”,在其它的种类识别器中 FE 理想输出值为“0000000”),*practical* 是该样本在 FE 阵列实际的输出值,*N* 是训练集样本的个数,*error* 即为该 FE 阵列对所有训练集样本的误差值,其值越小表示该基分类器的分类性能越好。

完成选择后的基分类器序列,采用大多数投票策略[22]来决定集成系统的分类结果。基于 Bagging 的 EHW 选择性集成学习算法的主要步骤如下:

输入:训练集 L_{tr},基分类器算法 C,基分类器个数 $t(t=20)$,选择的基分类器个数 $s(s=5)$,评测方法 M.

输出:选择的基分类器集合 $S = \{C_1^*, C_2^*, \cdots C_s^*\}$

训练过程:

初始化:令基分类器集合 $T=\Phi$.

For n = 1,2,$\cdots t$

基于训练集 L_{tr},采用 bootstrap 方法获取新的训练集 $L_{tr}^{(t)}$.

应用基分类器算法 C 于 $L_{tr}^{(t)}$ 训练得到基分类器,将其加入集合 T.

EndFor(得到初始的基分类器集合 $T = \{C_1, C_2, \cdots C_t, \}$)

选择过程:

在训练集 L_{tr} 上对每个基分类器 $C_i(i = 1,2,\cdots t)$ 进行测试,得到其输出 O_i.

集成过程:

利用评测方法 M 基于 $O_i(i = 1,2,\cdots t)$ 对 T 中每个基分类器进行评测,从中选择 s 个基分类器 $C_1^*, C_2^*, \cdots C_s^*$. 采用大多数投票方法集成最优基分类器.

4　结果与讨论

选用麻省理工学院的急性白血病数据集[1]和哈佛医学院的肺癌数据集[15]来验证本文方法的有效性。急性白血病数据集来源于 72 位不同病人的急性白血病样本,急性白血病分为急性淋巴细胞白血病(acute lymphoblastic leukemia, ALL)和急性骨髓细胞白血病(acute myeloid leukemia, AML)两种类型。选取数据集样本中 38 个样本(27 个 ALL,11 个 AML)作为训练集,另外独立的 34 个样本(20 个 ALL,14 个 AML)作为测试集。肺癌数据集包含恶性胸膜间皮瘤(malignant pleural mesothelioma, MPM)和肺腺瘤(lung adenocarcinoma, ADCA)两种类型,来源于 181 位不同病人的肺癌样本,其中 32 个样本(16 个 MPM,16 个 ADCA)作为训练集,另外 149 个样本(15 个 MPM,134 个 ADCA)作为测试集。

硬件平台采用 Celoxica 公司的 RC1000 板卡,在 Xilinx ISE6.3 开发环境下使用硬件描述语言 VHDL 设计硬件演化系统,用 ModelSim 进行功能仿真,综合实现后通过 PCI 总线接口下载程序到板卡的 Virtex xcv2000E FPGA 芯片中执行在线演化,得到最终的分类系统。

在白血病和肺癌数据集下,进行了本文方法(Bagging-EHW)与传统 EHW 方法[14]的全集成和选择性集成对比实验,对比结果见表 1 和表 2。相对于 EHW 采用完整的训练集训练各基分类器,Bagging-EHW 对训练集进行重采样构造了较小的训练子集。从实验结果可以看出:(1)相对于单一分类器,集成学习可以有效地提高系统的识别率(见①VS②,④VS⑤);(2)就全集成而言,在集成同样数目基分类器的前提下,Bagging-EHW 方法的识别率高于 EHW 方法(见②VS⑤,③VS⑥);(3)在同样集成 5 个基分类器的情况下,选择性集成的效果要明显好于全集成(见⑤VS⑦)。

表 1　Bagging-EHW 和 EHW 在白血病数据集中的平均识别率、硬件代价、平均演化时间、识别率方差对比

实验方法	集成方法	白血病			
		识别率(%)	硬件代价(slice)	演化时间(ms)	识别率方差(%)
EHW	①单个基分类器	85.19	3106	26.4	32.100
	②全集成 5 个基分类器	93.68	5986	132.2	5.340
	③全集成 10 个基分类器	95.00	11204	263.1	1.210
Bagging-EHW	④单个基分类器	86.76	3106	1.8	29.600
	⑤全集成 5 个基分类器	93.82	5986	9.6	5.020
	⑥全集成 10 个基分类器	95.88	11204	18.3	1.240
	⑦20 个基分类器选择性集成 5 个	95.88	5986	35.7	1.370

表 2 Bagging-EHW 和 EHW 在肺癌数据集中的平均识别率、硬件代价、平均演化时间和识别率方差对比

实验方法	集成方法	肺 癌			
		识别率(%)	硬件代价(slice)	演化时间(ms)	识别率方差(%)
EHW	①单个基分类器	93.96	3106	1.3	2.590
	②全集成 5 个基分类器	94.43	5986	6.4	0.130
	③全集成 10 个基分类器	95.64	11204	12.8	0.005
Bagging-EHW	④单个基分类器	94.36	3106	0.8	2.140
	⑤全集成 5 个基分类器	96.72	5986	4.1	0.110
	⑥全集成 10 个基分类器	97.98	11204	7.9	0.002
	⑦20 个基分类器选择性集成 5 个	97.54	5986	15.5	0.002

表 1、表 2 同时也给出了不同实验设定下的硬件实现代价(FPGA CLB Slice 占用)对比。从表中可以看出,在识别率相当甚至更优的情况下,Bagging-EHW 选择性集成方法的硬件代价低于其他 EHW 方法。对于白血病和肺癌数据集,采用 Bagging-EHW 选择性集成 5 个基分类器的识别率均优于采用 EHW 全集成 10 个基分类器时的识别率(见③VS⑦)。而此时 Bagging-EHW 选择性集成的硬件代价却仅 EHW 全集成 10 个基分类器的一半。而对于系统演化时间问题,由于 Bagging-EHW 重复随机抽样只选择原样本空间中的一部分进行训练,降低了演化硬件的算法和时间开销(见①VS④)。根据表 1、表 2 中的识别率方差对比可以看出,全集成和选择性集成 EHW 分类系统都比单分类器系统的稳定性强,并且在集成同样基分类器数目的条件下,选择性集成比全集成方法具有更好的稳定性。

表 3 给出了本文方法和其他传统分类方法在白血病和肺癌数据集下的识别率对比。可以看出,本文方法的识别率与其他方法具有一定的可比性。

表 3 本文方法与其他模式分类方法的识别率对比

实验方法	白血病(%)	肺癌	文献
Bagging-EHW	95.88	97.54%	本文
加权投票	85.29	–	[1]
神经网络集成	95.90	–	[7]
Bagging-C4.5	91.18	93.29%	[6]
BN	94.12	–	[5]
CBN	97.60	–	[4]

5 结 论

本文针对传统演化硬件集成学习方法识别稳定性差、硬件代价过高的问题,提出了一种基于 Bagging选择性集成的演化硬件 DNA 微阵列数据分类方法。该方法通过引入 Bagging 策略增加演化硬件基分类器间的差异性,同时通过选择性集成学习方法选择性能较好的部分基分类器集成到最终的分类系统。实验结果表明,与传统演化硬件分类方法相比,本文方法在保证系统识别率的同时有效降低了硬件资源消耗和系统学习时间,为实现更大规模的系统应用提供了可能。而与其他 DNA 微阵列数据分类方法相比,本文方法在具有较好分类性能的同时,具有可在线适应、实时性强、学习结果可读性好、识别速度快等优点。

参考文献

[1] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286(5439): 531-537

[2] Shen Q, Shi W M, Kong W, et al. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Computational Biology and Chemistry*, 2008, 32(1): 53-60

[3] Sandin I, Andrade G, Viegas F, et al. Aggressive and effective feature selection using genetic programming. In: Proceedings of the IEEE Congress on Evolutionary Computation, Brisbane, Australia, 2012. 1-8

[4] Piao H Y. A correlational Bayesian network for DNA microarray data analysis. In: Proceedings of the International Conference on Biomedical Engineering and Informatics, Shanghai, China, 2011. 1702-1705

[5] Helman P, Veroff R, Atlas S R, et al. A Bayesian network classification methodology for gene expression data. *Journal of Computational Biology*, 2004, 11(4): 581-615

[6] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2003, 2(2 suppl): 75-83

[7] Chao S B, Won H. Cancer classification using ensemble of neural networks with multiple significant gene subsets. *Applied Intelligence*, 2007, 26(3): 243-250

[8] Glette K, Torresen J, Yasunaga M. An online EHW pattern recognition system applied to sonar spectrum classification. In: Proceedings of the International Conference on Evolvable Systems: From Biology to Hardware, Wuhan, China, 2007. 1-12

[9] Wang J, Jung J K, Lee Y M, et al. Using reconfigurable architecture-based intrinsic incremental evolution to evolve a character system. In: Proceedings of the International Conference on Computational Intelligence and Security, Xi'an, China, 2005. 216-223

[10] Sekanina L, Frienl S. An evolvable combinational unit for FPGAs. *Computing and Informatics*, 2004, 23(5): 461-486

[11] Glette K, Torresen J, Yasunaga M. An online EHW pattern recognition system applied to face image recognition. In: Proceedings of the EvoWorkshops, Valencia, Spain, 2007. 271-280

[12] Wang J, Chen Q S, Lee C H. Design and implementation of a virtual reconfigurable architecture for different applications of intrinsic evolvable hardware. *IET Computers & Digital Techniques*, 2008, 2(5): 386-400

[13] 王进,康雄. 基于演化硬件的道路限速标志识别方法. 江苏大学学报,2011,32(6):689-694

[14] 王进,陈文,李钟浩. 用于癌症分子分型的虚拟可重构结构演化硬件. 华中科技大学学报,2012,40(4):23-28

[15] Gordon G J, Jensen R V, Hsiao L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 2002, 62: 4963-4967

[16] Haddow P C, Tyrrell A M. Challenges of evolvable hardware: Past, present and the path to a promising future. *Genetic Programming and Evolvable Machines*, 2011, 12(3): 183-215

[17] 张开锋,肖山竹,陶华敏等. 基于 GAL 软核的 EHW 平台设计技术研究. 武汉大学学报,2012,45(2):126-130

[18] 杨华秋,段欣,来金梅. 一种新型的单芯片级可进化硬件系统. 复旦大学学报,2012,51(1):47-53

[19] 张春霞,张讲社. 选择性集成学习算法综述. 计算机学报,2011,34(8):1399-1410

[20] Breiman L. Bagging predictors. *Machine Learning*, 1996, 24(2): 123-140

[21] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 2002, 137(1-2): 239-263

[22] Kuncheva L I. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(2): 281-286

Bagging-based selective ensemble of EHW for classification of DNA microarray data

Wang Jin, Ran Qianyuan, Ding Ling, Zhao Rui

(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065)

Abstract

In order to improve the generalization ability and reduce the hardware cost of evolvable hardware (EHW) classification systems, a bagging-based selective ensemble learning method using EMW multiple classifiers was proposed for the classification of DNA microarray data. A bagging-based selective ensemble learning strategy and a virtual reconfigurable architecture-based EHW classification system were studied. In the system learning process, several training subsets were generated by using random sampling from the original training set. The final EHW classifier was built by using the evolved base classifiers with the high classification rate. Both the system learning and the system classification of the EHW for the classification of microarray data were implemented on a Xilinx Virtex xcv2000E FPGA. The comparison of the experimental results of acute leukemia and lung dataset showed the proposed method's advantages of much lower hardware cost, higher recognition rate, shorter learning time and generalization ability compared with traditional EHW ensemble learning schemes.

Key words: evolvable hardware (EHW); Bagging; DNA microarray; selective ensemble