改进的BPSO的特征基因选择方法及其在结肠癌检测中的应用研究

柴 欣[1] 孙劲耀[1] 郭 磊[2] 武优西[1]
(河北工业大学计算机科学与软件学院 天津 300401)[1]
(河北工业大学河北省电磁场与电器可靠性重点实验室 天津 300130)[2]

摘 要 为了避免二进制粒子群算法(BPSO)容易陷入局部极值的缺陷,提出了一种改进的二进制粒子群算法(IBPSO)。该算法在运行过程中引入遗传算法的交叉和变异策略,以便增加种群的多样性,避免粒子的早熟收敛;同时采用免疫算法的疫苗机制,通过合理的疫苗提取、疫苗接种、疫苗选择有效地抑制种群退化的可能。首先采用Wilcoxon秩和检验指标来获得对分类起较大作用的预选特征子集,然后利用IBPSO算法对基因的特征子集和支持向量机(SVM)的参数进行寻优,最后采用IBPSO算法对结肠癌检测问题进行了研究。实验结果表明,该方法可以在较少的特征基因下取得较高精度,且所选的特征基因与结肠癌密切相关,进一步验证了方法的可行性和有效性。

关键词 特征选择,粒子群算法优化,支持向量机,秩和检验
中图法分类号 TP181,TP391.4 **文献标识码** A

Feature Gene Selection Based on Improved Binary Particle Swarm Optimization Algorithm and its Application in Detection of Colon Cancer

CHAI Xin[1] SUN Jing-yao[1] GUO Lei[2] WU You-xi[1]
(School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China)[1]
(Province-Ministry Joint Key Laboratory of Electromagnetic Field and Electrical Apparatus Reliability, Hebei University of Technology, Tianjin 300130, China)[2]

Abstract In order to avoid local optimal solution of Binary Particle Swarm Optimization algorithm, an Improved Binary Particle Swarm Optimization (IBPSO) algorithm was presented. In this approach, the crossover and mutational strategies are introduced to increase the diversity of populations and avoid the premature-convergence of particles. Vaccine extraction, vaccination and immune selection are used to realize the vaccine mechanism to control the population degradation. In order to reduce the features of the tumor, Wilcoxon is used to remove the useless genes. IBPSO algorithm is used to optimize the subset of features and the parameters of Support Vector Machine (SVM). Finally, this method mentioned above is applied to detect the key genes of colon cancer dataset. The experimental results show that our approach can get higher classification accuracy with smaller size of feature subset than that of some other approaches and the selected genes are proven to be disease-causing. The experimental results also verify the correctness and effectiveness of our approach.

Keywords Feature selection, Particle swarm optimization algorithm, Support vector machine, Wilcoxon

1 引言

DNA芯片可以在一次试验中同时检测成千上万个基因的表达量,为从分子水平上研究疾病的发病机理和临床疾病诊断提供了强有力的手段。然而,在基因表达谱数据的获取过程中,样本的数目一般为几十或上百例,检测基因的数目却往往高达几千甚至几万,其中含有大量类无关基因。同时,由于功能相似基因表达高度相关,因此存在大量在分类学意义上的冗余基因[1,2]。如何利用这种具有高维、高噪、高相关特点的有限样本基因表达谱数据,识别对疾病有鉴别意义的特征基因或疾病相关基因是机器学习课题研究中的热点之一。

基因选择方法大致可以分为过滤式方法和缠绕式方法。前者依据某个评价指标得分高低评价基因分类能力,计算复杂度低,但忽略了基因之间的相关性;而后者以基因子集上分类器的分类精度作为衡量,获得的子集预测能力高,但速度慢[3,4]。有学者尝试两种方法相结合的特征选择模型:先用

到稿日期:2012-09-12 返修日期:2012-12-11 本文受河北省自然科学基金(H2012202035),河北省教育厅重点项目(ZH2012038),河北省高等学校青年基金项目(SQ121006)资助。

柴 欣(1962—),男,硕士,教授,主要研究领域为信息处理与软计算、生物信息处理技术,E-mail:ch2121@126.com;孙劲耀(1986—),男,硕士生,主要研究领域为信息处理与软计算;郭 磊(1968—),男,博士,副教授,主要研究领域为机器学习、图像处理;武优西(1974—),男,博士,教授,CCF会员,主要研究领域为数据挖掘与智能计算,E-mail:wuc567@163.com(通信作者)。

过滤方法降低数据集维度，剔除部分不相关特征；再用缠绕方法在预选特征集上做进一步特征精选，取得了较好的效果。例如，Shen 等人[5]采用 T-test 指标剔除类无关基因，将粒子群算法结合禁忌搜索用于基因选择，适应值函数采用具有线性判别函数的 Fisher 分类器，在结肠癌数据集上用 8 个特征子集取得了 93.55%的分类正确率。Li 等人[6]采用 Wilcoxon 统计量预选择基因子集，然后采用遗传算法进行基因精选。适应值函数选取训练集上支持向量机分类训练样本的识别率，在结肠癌数据集上用 15 个特征基因子集取得 93.5%的分类正确率。目前组合式特征选择方法虽比单独使用有一定改善，但存在的问题依然很明显，如第二阶段的缠绕过程如何在特征子集规模、预测能力和其他约束条件等多个目标下求得折中解以及分类器参数的选择等[7]。

本文基于粒子群智能算法和统计学习理论的知识，在分析肿瘤基因表达谱特征的基础上，研究了结肠癌分类特征基因的选取问题。首先，本文选择 Wilcoxon 获得对分类起较大作用的预选特征子集，是因为 Wilcoxon 克服了 T-test 对于噪音的敏感和误判阈值会丢失表达数据中信息的缺点。通过 Wilcoxon 来对标准化处理后的基因打分，可以极大地去除噪声基因，获得对分类有帮助的特征子集。然后，选择 SVM 分类器进行分类，虽然 SVM 建立在结构风险最小化原则基础之上，抗噪声干扰能力强，有效地解决了过学习和欠学习问题[8,9]，但是 SVM 受其参数影响，学习效果差异较大；此外，适当的特征子集，仍然可以提高 SVM 学习效果[10]。为了进一步合理选择适当的特征子集和 SVM 的参数，我们提出了改进的二进制粒子群(Improved Binary Particle Swarm Optimization，IBPSO)算法，该算法在二进制粒子群算法基础上引入遗传算法的交叉和变异策略，以便增加种群的多样性，避免粒子的早熟收敛；同时采用免疫算法的疫苗机制，通过合理的疫苗提取、疫苗接种、疫苗选择有效地抑制种群退化的可能。实验结果表明，本方法在平均 16 个关键特征的子集上就取得了 95.16%的交叉正确分类率，并且子集中出现次数高的基因已被证实与结肠癌患病相关。

2 特征基因选择方法

2.1 基于 Wilcoxon 的特征基因初选

过滤方法采用基于信息统计的启发式准则来评价特征的预测能力，并选取较优的特征组成特征子集。该方法可以快速剔除部分噪声特征，缩小算法的搜索范围。

本文采用 Wilcoxon 秩和检验的筛选方法来获得对分类起较大作用的预选特征子集。

$$s(g)=\sum_{i\in N_0}\sum_{j\in N_1} I((x_j^{(g)}-x_i^g)\leqslant 0) \tag{1}$$

式中，N_0、N_1 分别表示基因表达数据的两类样本的个数，$x_i^{(g)}$ 表示第 i 号样本第 g 个基因的表达数据，I 函数为判别函数，取值如下：

$$I(x)=\begin{cases}1, & x\text{ 逻辑表达式为真}\\ 0, & x\text{ 逻辑表达式为假}\end{cases} \tag{2}$$

式中，$S(g)$表示基因在两类样本的表达差异，该分值可作为同一基因在两类别中表达差异的测度，值越接近 0 或越接近最大值 $N_0 * N_1$，则表示相应基因对于分类越重要。因此基因对分类的重要性可以根据如下公式来表示：

$$w(g)=\max(s(g),N_1\times N_0-s(g)) \tag{3}$$

即根据式(3)的值排序来选择重要的特征基因。

2.2 支持向量机

支持向量机(Support Vector Machine，SVM)建立在统计学习理论的 VC 维理论和结构风险最小原理基础上，根据有限的样本信息能够在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中，以期望得到最好的推广能力[11]。对于非线性问题，SVM 需要把输入样本映射到高维特征空间，常用到的内积函数形式主要有：

1)线性核(Linear)函数

$$K(x,y)=x^{\mathrm{T}}y \tag{4}$$

2)多项式(Polynomial)内核函数

$$K(x,y)=(x^{\mathrm{T}}y+1)^p \tag{5}$$

式中，p 为常数。

3)高斯径向基核(Gaussian Radial Basis Function)函数

$$K(x,y)=\exp(-\frac{1}{\sigma^2}\|x-y\|)^2 \tag{6}$$

式中，σ^2 为常数。

2.3 基于改进 BPSO 的特征基因精选

粒子群算法(Particle Swarm Optimization，PSO)是由 Kennedy 和 Eberhart[12]于 1995 年提出的一种新型的群智能进化计算技术，其基本概念源于对鸟类觅食行为的研究。通过研究鸟类飞行觅食行为，模拟鸟群的集体协作使群体以最快的速度找到最优解。在粒子群算法中，每个优化问题的解都是搜索空间中的一只鸟，我们称之为“粒子”。粒子追随当前最优的粒子(全局最优值)和自身经历的个体最优位置(个体最优值)来调节速度，在解空间中搜索最优解。粒子通过式(7)和式(8)来更新自己的位置与速度。

$$v_{id}^{k+1}=\omega v_{id}^k+c_1\gamma_1(p_{id}^k-x_{id}^k)+c_2\gamma_2(p_{gd}^k-x_{id}^k) \tag{7}$$

$$x_{id}^{k+1}=v_{id}^k+x_{id}^k \tag{8}$$

式中，i 表示粒子的编号，$i=1,2,\cdots,n$，d 表示搜索空间的维向量，v_{id}^k 与 x_{id}^k 分别表示第 i 个粒子在 k 代的速度与位置，p_{id} 表示粒子 i 的个体极值，p_{gd} 表示全局最优值。ω 为惯性权重，γ_1 与 γ_2 为 0 到 1 之间的随机数，c_1 与 c_2 为学习因子。

为了将粒子群算法应用于二进制编码来解决实际的组合优化问题，Kennedy 和 Eberhart[13]提出离散二进制粒子群算法。在他们提出的模型中，粒子在飞行过程中的位置以二进制编码形式实现，每一维的粒子位置 x_{id} 限制为 1 或 0，而对速度 v_{id}^k 不做限制，选用速度向量每一维的 $Sigmoid$ 函数值为粒子飞行位置改变的概率。$Sigmoid$ 函数是一类模糊函数，用这类函数的取值表示位置向量的第 d 维变化的概率，其表达式如下：

$$Sig(v_{id})=\frac{1}{(1+\exp(-v_{id}))} \tag{9}$$

$$x_{id}^{k+1}=\begin{cases}1, & Sig(V_{id}^{k+1})>\mathrm{rand}(0,1)\\ 0, & \text{else}\end{cases} \tag{10}$$

式中，当 $x_{id}^{k+1}=1$ 时，第 d 维特征被选中。速度更新变化过程与基本 PSO 算法一致。

分析基本的离散二进制粒子群算法的实现，发现其搜索最好解的过程中存在一些不足：速度更新变化和飞行位置的更改过程中，粒子到达下一个迭代时的状态主要由粒子自身的经验和粒子群中的共享经验决定。在这种单一正反馈实现步骤中，若粒子的自身经验占据了绝对优势，粒子就很容易陷入局部极值的状态。为了避免 BPSO 陷入早熟状态，需要对基本的 BPSO 进行改进。在粒子群算法中增加遗传算子操作

可以扩大种群的多样性，提高全局搜索能力，但会引来种群退化风险[14,15]。参考生物免疫系统中的自调节理念，我们同时引入免疫算法的疫苗机制，通过合理的疫苗提取、疫苗接种、疫苗选择，寻求局部搜索和全局搜索的动态平衡[16]。

本文采用的改进的 BPSO（Improved Binary Particle Swarm Optimization，IBPSO）算法的流程图如图 1 所示，其详细流程说明如下。

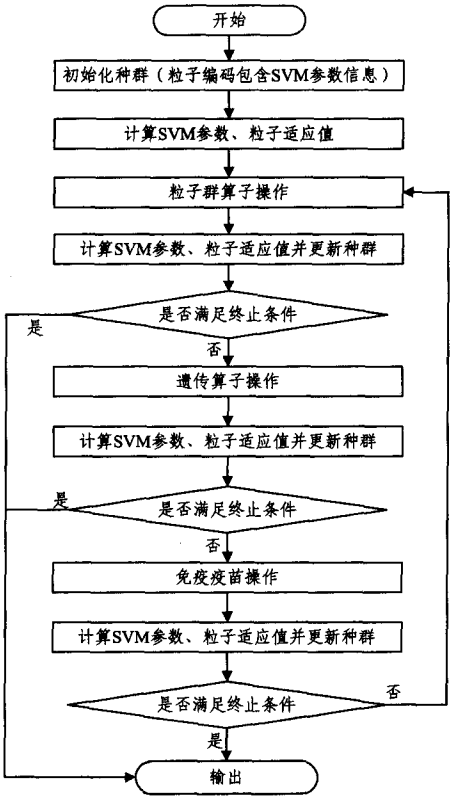

图 1　IBPSO 算法流程图

Step1　随机产生初始化种群，粒子的位置信息采用二进制编码，其中包含 SVM 参数和基因表达信息，并随机产生粒子的初始速度。如图 2 所示，编码前 10 位为参数 r，10－20 位为惩罚参数 C，后 100 位表示寻优子集（1 代表该处基因被选中表达，0 表示该处基因没有被选中）。

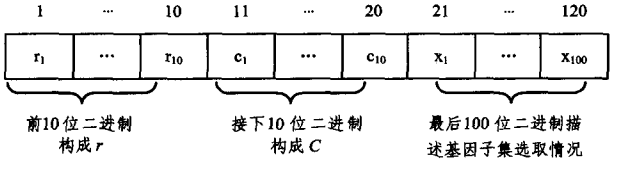

图 2　粒子编码示意图

Step2　计算 SVM 参数值、粒子的适应值，保存初始化时粒子个体与群体的极值信息。

Step3　粒子群算子操作，包括更新粒子各自的速度与位置。

Step4　计算 SVM 参数值、新种群粒子个体的适应值，更新个体极值和全局极值。

Step5　判断是否满足终止条件，如满足，则终止程序，否则进行遗传算子操作。

Step6　遗传算子操作包括选择、交叉、变异操作，增加粒子多样性以防止单纯的粒子群算法陷入局部最优。其中选择采用轮盘赌策略，交叉操作采用单点交叉。

Step7　计算经过遗传操作后粒子位串中 SVM 参数值、粒子适应值，并更新极值信息。

Step8　判断是否满足终止条件，如满足，则终止程序，否则进行免疫疫苗操作。

Step9　计算经过免疫操作后粒子位串中 SVM 参数值、粒子适应值，并更新极值信息。

Step10　判断是否满足终止条件，如满足，则终止程序，否则转 Step3。

2.4　评估函数

我们采用基因子集上留一交叉验证分类精度和子集数量大小来作为其适应问题环境能力的考量，采用以下适应度函数[7]：

$$f(I)=\beta * ACC_{svm}(I)+(1-\beta) * (1-|I|/M) \tag{11}$$

式中，$ACC_{svm}(I)$ 是子集 I 在训练集上利用支持向量机分类训练样本的分类正确率，$|I|$ 是 I 中被选中表达的基因个数，M 为初选基因子集的大小，β 为权重。

2.5　SVM 参数优化

支持向量机采用径向基核函数，其中参数 C 和 γ 在编码粒子时分别用 10 位二进制代码表示[17]，根据式（12）和式（13）将其转成实数：

$$C=3\times\sum_{n=1}^{10}(C_n\times 2^{n-1}) \tag{12}$$

$$\gamma=\frac{\sum_{i=1}^{10}\gamma_i\times 2^{i-1}}{2^{10}-1} \tag{13}$$

将 IBPSO 算法用于基因特征选择时，粒子的长度对应 2.1节中预选择的特征数量与 SVM 高斯核参数编码长度之和。选定的基因子集按照相应的参数去训练分类器获得分类精度。通过式（11）评价各粒子的适应值，粒子跟踪局部和全局两个极值在搜索空间搜寻最优解。终止条件设定为粒子群算法迭代到某一预设代数或达到评价函数的标准。

3　仿真实验与结果

3.1　实验数据

为了验证所提算法的有效性，本文在公开的肿瘤数据集——结肠癌数据集（Colon Cancer Dataset）[18] 上进行了实验仿真。该数据集包括 62 个样本，每个样本含有 2000 条基因，其中 40 个样本为肿瘤，22 个为正常结肠组织。算法采用 Matlab 编程实现，分类器采用 Chang 等开发的支持向量机软件 LIBSVM（http://www.csie.ntu.edu.tw/～cjlin/libsvm）。在开始时对数据进行归一化处理，均值为 0，标准方差为 1。

3.2　实验结果分析

本文首先采用 Wilcoxon 评分准则预选择 100 个基因作为初选特征子集，然后采用 IBPSO 算法进行信息基因的精选。实验中，加速度系数 $c_1=c_2=2$，惯性权重 $\omega=0.9$，交叉概率 $pc=0.9$，变异概率 $pm=0.05$，接种疫苗数量 $r=20$，接种基因位突变概率 $pr=0.8$，权重 $\beta=0.9$，粒子规模设置为 40，最大迭代次数为 50。表 1 给出了在结肠癌数据集分别采用 Wilcoxon 过滤、Wilcoxon 过滤结合 BPSO 与本文方法获得的留一交叉分类正确率（即类别预测正确的样本数占预测中用到的总样本数的比例）及对应获得的子集中有效特征数量，其中 BPSO 参数设置与 IBPSO 相同。

实验结果显示，当只使用 Wilcoxon 过滤方法选择的基因建立分类预测模型时，由于存在大量冗余基因，分类结果较差；而使用 Wilcoxon 过滤与 BPSO 结合做基因选择时，在一定程度上消除了冗余基因，提高了分类精度；而本文提出的算法能有效地防止粒子陷入局部最优解，因此能够在提高分类精度的同时减少特征基因的数目。

表 1　不同基因选择方法的实验对比结果

基因选择方法	特征基因数量	识别精度
Wilcoxon 过滤	100	83.87%
Wilcoxon+BPSO	27	90.32%
Wilcoxon+IBPSO	16	95.16%

3.3　相关工作比较

一些相关文献[6,17,19-21]报道了它们在 Alon 结肠癌数据集[18]上获得的分类结果，具体如表 2 所列。

表 2　不同基因选择方法的实验对比结果

参考文献	基因选择方法	特征基因数量	识别精度
文献[6]	Wilcoxon+GA(SVM)	15	93.50%
文献[17]	Wilcoxon+ DPSO(SVM)	26	93.60%
文献[19]	SNR+DPSO(SVM)	34	89.67%
文献[20]	SVM-RFE(SVM)	50	89.37%
文献[21]	DF+SVM	30	<95.00%
文献[22]	CC+C4.5	50	<90.00%
文献[23]	Fisher-RG-SVMRFE	12	94.70%
本文	Wilcoxon+ IBPSO(SVM)	16	95.16%

注：表中缩写说明如下：SNR=(Signal to Noise Ratio)；RFE=(Recursive Feature Elimination)；DF=(Decision Forests)；GA=(Genetic Algorithm)；CC=(Correlation Coefficient)；Fisher-RG-SVMRFE = Fisher criteria and the Redundancy reduction greedy approach with GO information and SVMRFE algorithm)。

通过表 2 可以看出，本文所提方法与文献[17,19-22]相比，在分类精度和特征数量上均占优势；文献[6,23]获得的子集数目虽然较少，但是子集中可能缺失了某些关键特征，所以获得的分类准确度并没有本文的高。

由于每次交叉验证中训练样本不同，因此选择出的特征子集也不尽相同。我们统计了 62 个子集中基因出现的次数，如图 3 所示。横轴表示预选择的 100 个特征的索引，纵轴为出现次数。

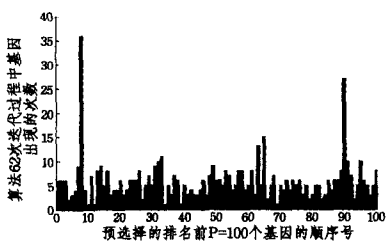

图 3　结肠癌上特征基因出现的次数

为进一步验证算法选择出的基因的有效性，根据出现次数，我们选择了前 $k=25$ 个基因来验证其上的分类推广能力。基因被选择的次数越多，则它对分类越重要，如图 4 所示。横轴上基因按出现次数由大到小排列，纵轴为分类率。从图中可以看出，选择前 6 个出现频率最高的基因取得了 98.4% 的推广分类性能，即 62 个样本 1 个被分错，选择前 25 个基因只取得了 83.9% 的分类率，有所下降。这说明影响疾病分类能力的基因数目是一定的，多余的特征反而会降低模型的判断能力。

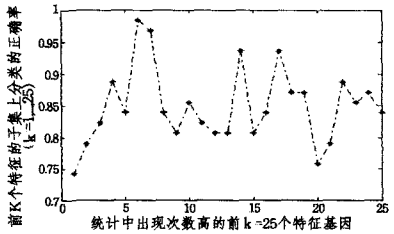

图 4　前 k 个基因的留一交叉验证分类率

同时我们列出了试验中统计出现次数最高的前 8 个基因{H64807，M76378，H64489，U19969，T61661，M63391，H08393，T47377}，并给出简单分析。这些特征除 H64807 外，其余都至少在 T-test、Information gain、Sum of variances、Towing rule、Gini index、Sum minority、Max minority 和 1D SVM 等 8 种排序(http://genomics10.bu.edu/yangsu/rank-gene/compare-alon-colon-cancer-top100.html)中的 3 种以上出现过。其中 M63391 在 7 种排序方法(除 T-test)中被选择排在前 2 的位置，H08393 排在 T-test 方法中的第一位。H64807 与 H08393 出现在文献[24]选择的排名前 10 个特征中。Gordon 在文献[25]中指出 H64807 作为叶酸运输车(folate transporter)与叶酸紧密联系，而临床研究发现叶酸状态的降低会增强患结肠癌的风险。Karakiulakis 等人[26]认为 H08393 会影响胶原蛋白的产生进而影响细胞之间的粘附与分离特性，与癌细胞的转移活动有关。综上，实验中统计出现次数高的基因与结肠癌发病密切相关，进一步地验证了算法选择出的基因的正确性和有效性。

结束语　本文应用粒子群智能算法和统计学习理论进行特征基因选择，首先采用 Wilcoxon 去掉大量噪声基因，减少算法的搜索空间，然后 IBPSO 算法被用来精选特征子集。通过引入遗传算子、疫苗机制来平衡粒子寻优过程中的局部与全局搜索能力，以优化的 SVM 分类器、加权系数的适应值函数指导搜索方向，最终获得小的子集数量与较高的分类推广能力，实验结果表明了这一点；并且实验结果还表明，子集中统计次数出现高的基因确实与结肠癌患病相关，这说明本文提出的方法具有实际应用价值。

参 考 文 献

[1]　李霞，张田文，郭政. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报，2004，27(5)：675-682

[2]　徐菲菲，苗夺谦，魏莱. 基于模糊粗糙集的肿瘤分类特征基因选取[J]. 计算机科学，2009，36(3)：196-200

[3]　Salem D A，Seoud R A，Ali H A. DMCA：A combined data mining technique for improving the microarray data classification accuracy[A]// 2011 International Conference on Environment and Bioscience，2011[C]. Singapore：IACSIT Press，2011：36-41

[4]　周昉，何洁月. 生物信息学中的基因芯片的特征选择技术综述[J]. 计算机科学，2007，34(12)：143-150

[5]　Shen Q，Shi W，Wei K. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data[J]. Computational Biology and Chemistry，2008，32(1)：53-60

[6]　Li S，Wu X，Hu X. Gene selection using genetic algorithm and support vectors machines[J]. Soft Computing，2008，12(7)：693-698

[7] Paul T K, Iba H. Extraction of informative genes from microarray data[A]// Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, 2005 [C]. Washington, DC, USA: ACM, 2005: 453-460

[8] Zhang C, Tian Y, Deng N. The new interpretation of support vector machines on statistical learning theory[J]. Science China Mathematics, 2010, 53(1): 151-164

[9] Damaševicius R. Optimization of SVM parameters for recognition of regulatory DNA sequences[J]. Top., 2010, 18(2): 339-353

[10] Guo L, Wu Y, Zhao L, et al. Classification of mental task from EEG signals using immune feature weighted support vector machines[J]. IEEE Transactions on Magnetics, 2011, 47(5): 866-869

[11] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995

[12] Kennedy J, Eberhart R C. Particle swarm optimization [A]// Proceedings of the IEEE International Conference on Neural Networks, 1995[C]. Perth, Australia, 1995: 1942-1948

[13] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[A]// Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1997[C]. Orlando, USA, 1997: 4104-4108

[14] Xu Y, Liu G. A method of emotion recognition based on ECG signal[A]// Proceedings of International Conference on Computational Intelligence and Natural Computing, 2009[C]. Wu Han, China: CINC, 2009: 202-205

[15] Mohamad M S, Omatu S, Deris S, et al. Particle swarm optimization with a modified sigmoid function for gene selection from gene expression data[J]. Artificial Life and Robotics, 2010, 15(1): 21-24

[16] 吴光华，刘光远，龙正吉. 免疫机制对皮肤电信号情感特征选择的影响[J]. 计算机应用研究，2010，27(12)：4558-4564

[17] 吴希贤. 基于优化算法的基因选择与癌症分类[D]. 长沙：湖南大学，2008

[18] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Science, 1999, 96(12): 6745-6750

[19] 张焕萍，宋晓峰，王惠南. 基于离散粒子群和支持向量机的特征基因选择算法[J]. 计算机应用与化学，2007，9(24)：1159-1162

[20] 王思漫. 基于基因表达谱的肿瘤分类方法研究[D]. 南京：南京理工大学，2012

[21] 李欣. 基于决策森林法的肿瘤基因表达谱数据分析[D]. 北京：北京工业大学，2011

[22] Zhang Z, Li J, Hu H, et al. On the effectiveness of gene selection for microarray classification methods[J]. Intelligent Information and Database Systems Lecture Notes in Computer Science, 2010, 5991(1): 300-309

[23] Mohammadi A, Saraee M, Salehi M. Identification of disease-causing genes using microarray data mining and gene ontology [J]. BMC Medical Genomics, 2011, 4(1): 12

[24] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(3): 389-422

[25] Gordon D. Epidemiologic evidence underscores role for folate as foiler of colon cancer[J]. Gastroenterology, 1999, 116(1): 3-4

[26] Karakiulakis G, Papanikolaou C, Jankovic S M, et al. Increased type iv collagen-degrading activity in metastases originating from primary tumors of the human colon[J]. Invasion and Metastasis, 1997, 17(3): 158-168

(上接第 231 页)

是相同的。不同于在线计算的方法，诊断测试的生成是离线完成的。因此，基于诊断测试的方法可用于在线诊断。

但对于复杂的系统，通过本文提出的方法生成完备和可靠的诊断系统是困难的。利用系统的结构特点来分解是可以尝试的方法。另外，如何在不显著提高成本的情况下，尽量多地生成对应的判定测试，尽量减少不可判定的行为假设集，也是值得研究的。

参 考 文 献

[1] Reiter. A theory of diagnosis from first principles[J]. Artificial Intelligence, 1987, 32: 57-96

[2] Chittaro L. Hierarchical model-based diagnosis based on structural abstraction[J]. Artificial Intelligence, 2004, 155: 147-182

[3] Baroni P. Diagnosis of large active systems[J]. Artificial Intelligence, 1999, 110: 135-183

[4] Console L, Picardi C, Ribaudo M. Process algebras for systems diagnosis[J]. Artificial Intelligence, 2002, 142: 19-51

[5] 张学农，姜云飞，陈蔼祥，等. 值传递诊断过程的的抽象与重用[J]. 计算机学报，2009，32(7)：1264-1279

[6] Zhang Xue-nong. Formal analysis of diagnostic notions[C]// Proceedings of International Conference on Machine Learning and Cybernetics. Xi'an, China, 2012: 1303-1307

[7] 王楠，欧阳丹彤. 基于模型诊断的抽象分层过程[J]. 计算机学报，2011，34(2)

[8] Pencole Y. A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks[J]. Artificial Intelligence, 2005, 164: 121-170

[9] Portinale L, Magro D, Torasso P. Multi-modal diagnosis combining case-based and model-based reasoning: a formal and experimental analysis[J]. Artificial Intelligence, 2004, 158: 109-153

[10] Console L. Temporal decision Trees: Model-based Diagnosis of Dynamic Systems On-board[J]. Journal of Artificial Intelligence Research, 2003, 19: 469-512

[11] Milde H. Integrating model-based diagnosis techniques into current work processes-three case studies from the INDIA project [J]. AI Communications, 2000, 13: 99-123

[12] Gertler J, Singer D. A new structural framework for parity equation-based failure detection and isolation[J]. Automatica, 1990, 26(2): 381-388

[13] 姜云飞，李占山. 基于模型诊断的元件替换与替换测试[J]. 计算机学报，2001，24(6)：666-672

[14] 张学农，姜云飞，陈蔼祥. 一致性诊断的测试[J]. 2008，29(8)：1525-1528