基于量子粒子群和随机森林的特征选择方法

杨明旭[1]，洪文财[1]，米 红[2]
（1、厦门大学自动化系 福建 厦门 361005 2、浙江大学非传统安全与和平发展研究中心 浙江 杭州 310028）

[摘 要]：提出一种基于量子粒子群和随机森林封装的特征选择方法。将量子粒子群算法用于特征选择，优化特征子集，采用随机森林分类器评价特征子集的性能，指导特征子集更新，以寻求最优的特征子集。

[关键词]：量子粒子群；特征选择；随机森林

0、引言

基因表达数据分析的主要任务就是对样本进行分类，希望用较少的基因数目获得较好的分类效果，而基因表达谱数据集的一个显著特点是样本少、维数高，大量的样本属性中仅有少量基因包含了样本的分类信息。因此如何找出影响样本信息的特征基因，就成为基因表达谱分析的关键。

目前常用的特征选择方法可分为过滤法（Filter）、封装法（Wrapper）[1]。过滤法计算简便、速度快；封装法较复杂但分类效果优于过滤方法。在基因表达谱的特征选取中，结合 Filter、Wrapper 的优点，采用启发式搜索算法和分类器进行封装逐渐成为当前的热点。

本文结合了 Filter、Wrapper 的优点，提出一种基于量子粒子群和随机森林封装的特征选择方法。将量子粒子群算法用于特征选择，优化特征子集，采用随机森林分类器评价特征子集的性能，指导特征子集更新，以寻求一组最优的特征子集。实验结果表明，基于量子粒子群和随机森林的特征选择方法能够找到最少的特征子集达到比较高的分类效果，同随机森林自身的重要性排序具有可比性。

1、量子粒子群和特征选择

1.1 量子粒子群简介

粒子群优化算法（Particle Swarm Optimization，PSO）[2]，是由J. Kennedy和R. C. Eberhart等于1995年开发的一种演化计算技术，来源于对鸟类和鱼群捕食等行为的模拟。在鸟类捕食的群体行为中，每只鸟被看作一个粒子，而每个粒子代表一个被优化问题的解。在D维搜索空间中，设微粒x_i本身所找到的最佳位置为$p_i=(p_{i1},p_{i2},\ldots,p_{iD})$，称为粒子个体最优点。整个粒子群迄今为止搜索到的最佳位置为$p_g=(p_{g1},p_{g2},\ldots,p_{gD})$，称为粒子群全局最优点。粒子当前速度为$v_i=(v_{i1},v_{i2},\ldots,v_{iD})$，每个粒子将根据一定的公式来调整自己下一步位置。

基本的PSO粒子群系统，粒子的收敛索空间是一个有限的区域，不能保证以概率1搜索到全局最优解。Sun等人从量子力学的角度出发提出了一种改进的PSO算法——具有量子行为的粒子群算法(Quantum-behaved Particle Swarm Optimization)。其粒子的速度和位置信息都将归结为一个参数β，如下所示：

$$p=\frac{r_1 pBest+r_2 gBest}{r_1+r_2} \tag{1}$$

$$mBest=\frac{\sum_{i=1}^{M} p_i}{M} \tag{2}$$

$$X(t+1)=p\pm\beta\,|mbest-X(t)|\ln(\frac{1}{u}) \tag{3}$$

$$\beta=\frac{(1-0.5)*(Iter\_\max-iter)}{Iter\_\max}+0.5 \tag{4}$$

式中，$pBest$为个体极值，$gBest$为全局极值，$mBest$为中值最优位置，M为群体中所含粒子数，r_1，r_2，u是$(0,1)$之间的随机数，β为系数创造力，调节它的值控制算法的收敛速度。通常情况下，β从1.0线性减小到0.5时，算法可以达到比较好的效果。在迭代过程中，式(3)中±是由随机数u决定的，当u大于0.5时取+，否则取−。

1.2用于特征选择的量子粒子群

量子粒子群工作在连续空间，对其进行离散二进制处理，使其可应用于特征选择。

将各特征作为位置点，若有D个特征，则粒子为D 维向量。产生初始粒子群，随机产生n个粒子，每个粒子x(t)为D 维二进制向量，值为(0,1)之间的随机数。对x(t)取整，x_B(t)=round(x(t))，每个粒子表示为由0、1构成的二进制向量，将值为0的特征选出，值为1的特征不选择，由此得到每个粒子的初始特征子集。

每个初始特征子集，用随机森林分类器评价其性能，适应度函数定义如下：

$$fitness=Accuaray-k*ones/All \tag{5}$$

式中，$Accuaray$是每个粒子分类的错误率，$ones$是每个粒子选取的特征数，All是全体特征数，k是准确率和特征数的平衡系数，k值越大，表示特征数量越受重视。计算所有粒子适应度的大小，每个粒子的初始位置作为个体极值$pBest$，全局极值$gBest$，为适应度值最大的粒子。

根据式(1)-(4)，更新每一个粒子$x(t+1)$，限制更新后的粒子$x(t+1)$的每一维是在$[0,1]$之间的数，将大于1的设为1，小于0的设为0。对$x(t+1)$取整，$x\_B(t+1)=round(x(t+1))$。每个粒子又表示为由0、1构成的二进制向量，将值为0的特征选出，值为1的特征不选择，由此得到每个粒子的初始特征子集。如此反复，得到一组组新的特征子集。

2、量子粒子群和随机森林的特征选择

基于量子粒子群和随机森林的特征选择结合了Filter、Wrapper的优点，利用信噪比[4]方法去除掉了多数不相关的特征，减少算法计算复杂度后，将量子粒子群算法用于特征选择，采用随机森林分类器评价特征子集的性能，指导特征子集的计算和更新，使搜索快速收敛。具体过程如下：

2.1信噪比过滤不相关的基因

以信噪比方法衡量基因的重要性

$$sn(i)=\left|\frac{u_+(i)-u_-(i)}{\sigma_+(i)+\sigma_-(i)}\right| \tag{6}$$

$sn(i)$是第i个特征的表达差异值，$u_+(i)$是第i个特征类标识为正类的样本的平均值。$\sigma_+(i)$是其标准差。$u_-(i)$是第i个特征类标识为负类的样本的平均值，$\sigma_-(i)$是其标准差。选取前300个$sn(i)$值较大的基因，滤去了多数不相关特征，大大减少了特征选择的计算复杂度。

2.2分类器的选择

随机森林是Leo Breiman于2001年提出的一个组合分类器算法，是由许多单棵分类回归树(CART)组合而成的，最后由投票法决定分类结果。整体的泛化误差取决于森林中单棵树的分类效能和各分类树之间的相关程度。Breiman采用Bagging和Randomization相结合的方法，在保证单棵分类树效能的同时，减少

各分类树之间的相关度，提高了组合分类器的性能。能较好地解决小样本、高维数数据的分类问题，且分类速度快，因此随机森林作为搜索过程的分类器。

随机森林同时是一种重要的特征选择方法，可与提出的方法进行比较。

2.3量子粒子群和随机森林的特征选择

信噪比得到的300个基因，用量子粒子群算法进行特征选择，用随机森林分类器评价特征子集的性能，其适应度函数如式(5)所示，k取值为0.02。具体算法实现如下：

(1)产生初始粒子群

按上文提到的方法产生初始粒子群，转换成二进制向量，得到初始特征子集；

(2)根据式(5)计算所有粒子的适应值，每个粒子的初始位置作为个体极值pBest，全局极值gBest为适应度值最小的粒子；

(3)更新粒子的速度和位置

根据式(1)-(4)，更新每一个粒子。由更新后的二进制粒子向量x_B(t+1)，将值为0的特征选出，值为1的特征不选择，得到新的特征子集，根据式(5)计算所有粒子的适应值。更新个体极值和全局极值。若更新后的二进制粒子向量x_B(t+1)全为1，无特征可选。则x_B(t+1)更新为一组随机产生的二进制向量；

(4)判断循环是否终止，产生全局最优解和最优特征子集。否则，返回至(3)。

3、实验及分析

为了验证量子粒子群特征选择的性能，采用4个基因数据集进行实验，数据集如表1所示：

数据集	样本数	特征数	类别
Colon	62	2000	2
DLBCL	77	5469	2
Leukemia	72	3571	2
Prostate_Tumor	102	10509	2

表1 基因数据集

用信噪比得到的300个特征利用随机森林分类器对所有样本进行分类。树大小ntree=500，其余参数为默认，得到300个特征的重要性排序。根据特征重要性排序，取出最重要的特征做为初始特征子集，样本按类别分为5份，每次取1份做为测试集，剩余4份为训练集。5份测试集的准确率平均得到该特征的适应度值，做20次得其平均值及标准差。再取出剩余特征中最重要的特征，与初始特征组成新的特征子集。重复，直到特征数达到50个。记录下最少特征子集达到最比较好的分类效果。

量子粒子群特征选择，利用信噪比获得的300个特征基因进行迭代，粒子群大小设定是30，迭代次数为100，对每一个特征子集，样本按类别分为5份，每次随机取1份做为测试集，剩余4份为训练集。做20次得其平均值及标准差。实验结果如表2所示：

数据集	随机森林		量子粒子群特征选择	
	特征数	准确率	特征数	准确率
Colon	3	0.884±0.014	3	0.893±0.025
DLBCL	12	0.942±0.007	4	0.946±0.015
Leukemia	15	0.979±0.007	12	0.982±0.015
Prostate_Tumor	19	0.938±0.008	8	0.942±0.016

表2 实验结果

基因表达数据具有维数高、样本少等特点，采用一定的特征基因选择方法以减少特征数是非常必要的。本文结合Filter、Wrapper的优点，提出了基于量子粒子群算法和随机森林分类器相结合的特征基因选择方法，从实验结果可以看出，特征选择算法去掉大部分不相关基因，减少特征数，提高分类准确率，具有较高的有效性和可行性。

提出的方法能取得较好结果主要原因：

(1)使用信噪比挑选出300个基因，滤去多数不相关特征，大大减少了特征选择的计算复杂度；

(2)量子粒子群不断更新粒子，使得粒子具有多样性，避免局部最优；

(3)采用特征组合的方式，避免随机森林等方法每次对单个特征计算忽略特征之间的相关性；

(4)量子粒子群算法在搜索过程中只有一个参数 ，当 从1减少到0.5时，粒子收敛，避免陷入局部最优，提高了算法的性能。

参考文献：

[1] 段艳华. 基于基因表达谱的肿瘤分类特征基因选择研究[D].北京：北京工业大学，2008

[2] KENNEDY J, EBERHART R.C. Particle swarm optimization. In: Proc IEEE Conference on Neural Networks. Piscataway, NJ, 1995, (4): 1942-1948

[3] SUN J, FENG B. Particle swarm optimization with particles having quantum behavior[C]. China: Congress on Evolution Computation, 2004

[4] GOLUB R R,SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J].Science, 1999,289(5439):531-537.

[5] BREIMAN L. Random forests. Machine Learning,2004,45:5-32

(上接第 82 页)

2.攻击阻止 NIDS 不能去阻止攻击，而采用报警方式。

现在一些产品扩展了 IDS 的功能，提供具有中断入侵会话的过程和非法修改访问控制列表来对抗攻击。

4.结束语

HIDS 和 NIDS 都有各自的优点，两者相互补充。两种方式都能发现对方无法检测到的一些入侵行为。例如，如果本地服务器发起的攻击可能不通过网络，无法通过 NIDS 发现，只能使用 HIDS 来判断。NIDS 通过检查所有的数据包头的标志位来进行发现，而 HIDS 并不查看包头的首部。NIDS 可以研究负载的内容，查找特定攻击中使用的命令或语法，而 HIDS 无法看到负载，也无法识别嵌入式的攻击。例如，网络型的入侵检测检查所有的数据包头的标志位，而主机型的入侵检测并不查看包头的首部；如本地服务器发起的攻击可能不通过网络，无法通过网络型的入侵检测来发现，只能使用主机型的入侵检测来判断；网络型的入侵检测可以研究负载的内容，查找特定攻击中使用的命令或语法，而主机型的无法看到负载，也无法识别嵌入式的攻击。因此，网络型和主机型的入侵检测各有优势，两者相互补充才能使网络系统预警通报的实现更加可靠、准确。

参考文献：

[1]张仕斌.网络安全技术.北京:清华大学出版社,2008

[2]王 达.网络管理员必读.北京:电子工业出版社,2007

[3]程柏良.基于异常与误用的入侵检测系统.《计算机工程与设计》,2007 年 14 期

[4]王高平.网络与应用教程.北京:清华大学出版社,2007

[5]黄淑华.计算机网络技术教程.北京:机械工业出版社,2004