基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择*

吴红霞，吴　悦，刘宗田，雷　州
（上海大学 计算机工程与科学学院，上海 200072）

摘　要：针对 SNP 的全基因组关联分析面临 SNP 数据的高维小样本特性和遗传疾病病理的复杂性两大难点，将特征选择引入 SNP 全基因组关联分析中，提出基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择方法。该方法包括两个阶段：Filter 阶段，使用 Relief 算法剔除无关 SNPs；Wrapper 阶段，使用基于支持向量机的特征递归消减方法（SVM-RFE）筛选出与遗传疾病相关的关键 SNPs。实验表明，该方法具有明显优于单独使用 SVM-RFE 算法的性能，优于单独使用 Relief-SVM 算法的分类准确率，为 SNP 全基因组关联分析提供了一种有效途径。

关键词：单核苷酸多态性；全基因组关联研究；特征选择；过滤式；缠绕式；组合式
中图分类号：TP309　　**文献标志码**：A　　**文章编号**：1001-3695(2012)06-2074-04
doi:10.3969/j.issn.1001-3695.2012.06.018

Combined SNP feature selection based on relief and SVM-RFE

WU Hong-xia, WU Yue, LIU Zong-tian, LEI Zhou
(*School of Computer Engineering & Science, Shanghai University, Shanghai* 200072, *China*)

Abstract: The genome-wide association study (GWAS) on SNPs faces two big issues: high dimensional SNP data with small sample characteristics and complex mechanisms of genetic diseases. This paper proposed a combined SNP feature selection method through bring feature selection methods into GWAS. The method included two stages: filter stage, it used Relief algorithm to eliminate irrelevant SNP features, wrapper stage, it used support vector machine based recursive feature reduction (SVM-RFE) algorithm to select the key SNPs set. Experiments show that the proposed method has an obviously better performance than SVM RFE algorithm, and also gains higher classification accuracy than Relief-SVM algorithm, which provides an effective way for SNP genome-wide association analysis.

Key words: SNP; GWAS; feature selection; filter; wrapper; combined

0　引言

单核苷酸多态性（single nucleotide polymorphisms，SNPs）是在染色体基因组上的单个核苷酸引起的 DNA 序列多态性。SNPs 具有广泛性、代表性、遗传性、稳定性等特性，是人类基因组最丰富的遗传变异[1]。位于编码区的 SNPs 极有可能直接影响并控制着蛋白质的结构和表达水平，从而与疾病的遗传机理存在关联关系。因此，开展 SNP 关联研究对于定位致病基因、发现复杂疾病的遗传机理有着非常重要的意义[2]。

近年来，基因芯片技术的飞速发展，使得高通量检测 SNP 数据变得越来越简单。但实验所能测试的病例样本数量却不能和被检测的 SNPs 数量同比增长，致使 SNP 数据集中呈现样本数量远小于 SNP 数量（$m \ll n$）。面对具有高维小样本特性的 SNP 数据，如何有效进行海量数据关联分析，成为 SNP 关联研究的第一大难点。

事实上，SNP 关联研究还必须考虑另外一个问题，即遗传疾病病理的复杂性。研究表明，复杂疾病的产生和诱因比较复杂，通常是由多个基因共同作用所致。而 SNPs 的遗传却具有连锁不平衡性，即不同位点的等位基因之间存在着连锁关系而不是自由组合关系[3]。本文不深入探讨 SNP 致病机理，但是在关键 SNPs 筛选结果的表现形式上采取组合形式，即考虑 SNPs 组合效果，而非单个 SNP 的作用。

近些年来，SNPs 研究非常迅猛。在生物信息学领域，众多研究者使用现有统计分析工具，如 weka、R 的 lumi 包、plink 等对 SNP 数据进行分析[4]。这些统计学分析工具将单个 SNP 与疾病的相关程度作为筛选 SNP 的依据，忽略了 SNP 间的关联关系，导致筛选的结果可靠性程度不高。

SNP 的全基因组关联分析旨在全基因组范围寻找与疾病关联程度最高的 SNP 集合，即关键 SNPs。有研究者将数据挖掘、机器学习、模式识别等方法应用于 SNP 关联研究中[5~7]，在一定程度上考虑了 SNP 间的关联关系，或者通过大幅度降维降低计算成本。但是这些方法还不够理想，必须要找到一种既能有效降维，又能充分考虑 SNP 间关联关系的 SNP 数据分析方法。

本文将 SNP 关联研究过程看成是一个特征选择的过程，具体思路是：在机器学习过程中，将 SNP 数据的每一个样本看成机器学习的一个实例，通过对训练集样本进行训练，获取分类模型，用来对测试集进行预测分析；并根据预测准确率高低作为 SNPs 关键性程度指标，成为筛选关键 SNPs 集合

收稿日期：2011-10-14；**修回日期**：2011-11-26　　**基金项目**：上海市科委重点基础项目（09JC1411302）；上海市重点学科建设项目（J50103）

作者简介：吴红霞，女，河南信阳人，硕士研究生，主要研究方向为数据挖掘（wuhongxia@shu.edu.cn）；吴悦，女，教授，博导，主要研究方向为并行计算与应用、数据挖掘与生物信息学；刘宗田，男，教授，博导，主要研究方向为数据挖掘、人工智能和软件工程.

的依据。

特征选择算法按照特征集合的评价策略分为两种[8]:Filter 式算法效率较高,但未考虑特征之间的关系;Wrapper 式算法考虑了特征之间的关系,但算法自身却具有极高的时间复杂度。显然,这些算法单独直接应用于 SNP 的全基因组关联研究的效果都不够理想。本文将 filte 式的 Relief 算法与 wrapper 式的 SVM-RFE 算法进行有效组合,提出一种高效的组合式 SNP 特征选择方法。

1　问题描述

假设:有 n 个样本,每个样本由 m 个 SNP 组成,则原始 SNP 数据集可以表示成 $n \times m$ 的矩阵 $\boldsymbol{M}$。$\boldsymbol{M}$ 的行表示 SNP 特征,列表示样本。每个样本可以表示成这样一个序列向量:

$$\boldsymbol{X}_i = (\text{label}:\text{SNP}_1,\text{SNP}_2,\cdots,\text{SNP}_m) \quad i = 1,2,\cdots,N$$

其中:label 是样本的类别标签,label $\in$ {case,control},case 表示有病样本,control 表示正常样本。SNP_j($j = \{1,2,\cdots,m\}$)为第 j 个 SNP 在样本 $\boldsymbol{X}_i$ 中的表型值,可以是 AA、BB、AB、NC 四种类型。其中,AA 表示野生纯合型;AB 表示突变杂合型;BB 表示突变纯合型;NC 为分型失败标记。

对于 SNP 的特征选择,本文给出以下定义:

在一个 $n \times m$ 的 SNP 数据矩阵 $\boldsymbol{M}$ 中,采用机器学习的特征选择方法,以样本分类为目标,从矩阵 $\boldsymbol{M}$ 中挑选出 $k(k \ll n)$ 个 SNP 构成关键 SNPs 集合,使得 k 尽量小、分类准确率尽量高。序列向量 $\boldsymbol{X}_i$ 即为机器学习过程中的一个实例。

2　基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择

2.1　方法介绍

事实上,考虑 SNP 遗传特性与致病机理,在特征选择过程中,本文对关键 SNP 提出如下假设:关键 SNP 的筛选结果以组合形式出现;关键 SNP 的冗余特征是关键 SNP,无关 SNP 特征的无关特征亦是无关 SNP;尽可能多地删除无关特征,而尽可能保留冗余特征。

Relief 系列算法是公认的效果较好的 filter 式评估算法,能高效剔除无关特征,而不删除冗余特征;SVM-RFE 算法则是典型的 wrapper 式特征选择算法,能有效评估特征的关键性程度。本文取两者的优点,并对算法本身结合 SNP 数据特点进行优化,提出基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择方法。

该方法包括两个关键阶段:Filter 阶段,采用 Relief 算法对 SNP 数据集进行无关特征删除;Wrapper 阶段,使用基于支持向量机的特征递归消减(SVM-RFE)方法对 SNP 特征进行排序。最后,使用十字交叉验证筛选关键 SNPs。该方法示意图如图 1 所示。

2.2　Filter 阶段

Filter 阶段是组合式 SNP 特征选择的第一阶段,职责是在保证关键 SNPs 被保留的前提下,尽可能多地删除与疾病无关的 SNP 特征。

2.2.1　SNP 特征的 Relief 评估

Relief 算法最早由 Kira 等人[9]提出,用于解决二分类问题。该算法核心思想是:“好”的特征应该使同类样本间的距离越近,而使非同类样本之间的距离越远。

本文将 Relief 算法重新设计用于 SNP 特征选择的预筛选过程。两个样本之间的距离依据式(1)进行计算:

$$\text{distance}(s_i,s_j) = \sum_{\substack{k \in \{1,\cdots,n\} \\ i,j \in \{1,\cdots,m\}}} \text{diff}(k,s_i,s_j) \tag{1}$$

其中,$\text{diff}(k,s_i,s_j)$ 是 snp_k 特征在样本 s_i 和 s_j 中表达值的差异程度,按式(2)进行计算:

$$\text{diff}(k,s_i,s_j) = \frac{P(M(k,i) \mid \text{snp} = k)}{P(M(k,i))} - \frac{P(M(k,j) \mid \text{snp} = k)}{P(M(k,j))} \tag{2}$$

其中:$M(k,i)$ 表示第 i 个样本在第 k 个 SNP 处的分型值;$P(M(k,i) \mid \text{snp} = k)$ 表示该分型值在第 k 行中所占的比率;$P(M(k,i))$ 表示该分型值在整个样本集合中所占的比率。$P(M(k,i) \mid \text{snp} = k)$ 和 $P(M(k,i))$ 都是通过对数据样本进行一次性统计计算得出,计算复杂度为 $O(n)$。

为了衡量 SNP 特征的关键性程度,设计 SNP 特征权重计算公式如下:

$$w[\text{snp}_k] = \sum_{i=1}^{m} w[\text{snp}_k] - \text{diff}(\text{snp}_k,s_i,H) + \text{diff}(\text{snp}_k,s_i,M) \tag{3}$$

其中:H 表示 s_i 的同类最近邻,指与 s_i 有相同的类别标签,且距离最近的样本;M 为 s_i 的异类最近邻,指与 s_i 有不同的类别标签,且距离最近的样本。

2.2.2　算法流程

Filter 阶段的 Relief 算法流程如下:

输入:待处理数据集 S,过滤阈值 δ_1。

输出:S_{new}。

a)初始化

初始关键 SNPs 集合:S_{new} = null;待处理 SNPs 集合:$S = \{\text{snp}_1,\text{snp}_2,\cdots,\text{snp}_n\}$;

b)找近邻

for　$k = \{1,2,\cdots,n\}$ //循环特征集

　　GetDiff(snp_i,snp_j);//计算特征差异度

for　$i = \{1,2,\cdots,m\}$ //循环样本集

　　getDis(s_i,s_j);//计算样本距离

　　findNeighbor(s_i);//找出每个样本的最近邻

c)计算特征权重

for　$k = \{1,2,\cdots,n\}$

$w[\text{snp}_k] = \sum_{i=1}^{m} w[\text{snp}_k] - \text{diff}(\text{snp}_k,s_i,H) + \text{diff}(\text{snp}_k,s_i,M)$

d)过滤无关特征

if　$(w[\text{snp}_j] > \delta)$　$S_{\text{new}} = S_{\text{new}} \cup \{\text{snp}_j\}$

判断特征的 Relief 权值是否大于给定阈值 δ。若小于 δ,则说明该特征无关度较高,与疾病关联程度低,在 filter 阶段应该被过滤掉;若大于 δ 则保留,更新当前待分析的 SNP 集合 S_{new}。

2.3　Wrapper 阶段

2.3.1　SVM-RFE 算法介绍

基于支持向量机的特征递归消减(SVM-RFE)方法最早由 Guyon 等人[10]提出。该算法核心是:通过支持向量机(SVM)构造特征权重,每次迭代消去权重最小的特征,直至所有特征被删除。SVM-RFE 是典型的后向选择算法,越是重要的特征越晚被删除。

SVM 是根据结构风险最小化理论的学习算法,使分类误差极小化,并尽可能提高分类器的泛化性能[11]。其基本思想

是:用核函数将输入向量映射到高维特征空间 F 中,再在 F 中构造一个最大间隔的最优超平面,将样本无错误地分隔开,使两类的分类间隔最大[12]。

对于两类判别问题,SVM 训练得到分类模型的过程即是求解最小化泛函 J 的过程。

$$J=\frac{1}{2}w^2+C\sum_{i=1}^{N}\xi_i \tag{4}$$

其中:ξ_i 是松弛变量,对应 X_i 允许偏离的量,C 是用于控制目标函数错分样本惩罚程度的常量参数。

将第 i 个特征对目标函数 J 的影响 $\Delta J(i)$ 用泰勒公式展开,并引入最优解 $J=\frac{1}{2}\|w\|$:

$$\Delta J(i)=\Delta w_i^2 \tag{5}$$

由于 $\Delta w_i=w_i$,第 i 个特征对目标函数 J 的影响大小为 $C_i=w_i^2$。C_i 越大,第 i 个特征对分类贡献率越高,因此通常将 C_i 作为特征的重要性度量。

2.3.2 二次划分

原始 SVM-RFE 算法每次迭代只删除一个特征,算法效率较低。为尽可能减少迭代次数,本文提出一种动态限定的二次划分方法,即在大量数据集上,对特征的分类准则分数进行二次划分,依据划分结果选择每次迭代删除的 SNP 特征集合。

二次划分的具体步骤如下:

a)SVM 训练过程中,计算 SNP 特征的排序准则分数为 $C_i=W^2$。

b)第一次划分。将 SVM 当前训练特征的分类准则系数 C_i 按升序排列,并计算其一阶差分 Δc:

$$\Delta c_i=\begin{cases}0 & \text{if} \quad i=1\\ c'_i-c'_{i-1} & \text{if} \quad i=2,3,\cdots,n\end{cases} \tag{6}$$

在 Δc_i 中选择 $k-1$ 个峰值,将每两个峰值之间的特征作为一组,这样就将特征集合分成了 k 组,记为 $G_1,G_2,\cdots,G_k$。显然 G_1 组中单个特征的分类权值都是最小,而 G_k 中所有特征的分类权值最大。

c)第二次划分。考察 G_1 组特征,引入 Relief 权重计算新的特征排序准则分数,并用上述方法对新的分类准则分数进行第二次划分:

$$W_i=0.5C_i+0.5W_{\text{relief}(i)} \tag{7}$$

假设此次分为的 K 组为 $M_1,M_2,\cdots,M_k$,则待删除的特征集合为 M_1 组全部特征。对每次迭代过程删除的特征,分配相同的特征删除排序系数。

动态划分方法在第二次划分时将衡量特征的无关性程度的 Relief 权重考虑进来,实现了第二次剔除无关性程度较高的 SNP 特征。二次划分参数 K 是每次划分的分组数。通过两次划分,使得每次迭代过程删除特征数量为当前待分析 SNP 集规模的 $1/k^2$。迭代过程中,每次删除特征的数目随着特征集合规模变小而减小,在特征集合规模较小时,筛选的精度不断提高。

2.3.3 算法流程

采用二次划分方法改进后的 SVM-RFE 算法流程如下:

输入:待处理数据集 S_{new}、二次划分参数 K、SVM 参数。

输出:特征排序特征集 R。

a)初始化

训练样本矩阵:$X=\{x_1,x_2,\cdots,x_n\}$

类别标签:$Y=[y_1,y_2]$

初始特征集合 $S=\{\text{snp}_1,\text{snp}_2,\cdots,\text{snp}_s\}$,

特征排序集合 $R=[\,]$;

b)递归循环 SVM 直至 $S=[\,]$

(a)$X_0=X(:,S)$;//当前样本训练集

(b)训练分类器:$\alpha=\text{SVMtrain}(X_0,Y,C,\lambda)$,其中 C 为惩罚因子,λ 为 RBF 核函数参数。

(c)计算排序系数 $C_i=w_i^2$;

if(S.length > 1000)二次划分动态限定

else 删除排序系数最小的特征

(d)更新特征排序集合:

$R=[s(f),R]$,$s(f)$ 为此次迭代预备删除的特征或特征集合。

(e)更新待删除特征集合:

$$S=S(1:f[0]-1,f[\text{length}(f)]:\text{length}(S))$$

3 实验与分析

3.1 实验数据

采用一组高血压数据进行实验,包括 96 个病例样本和 96 个对照样本,每个样本共有 590 960 个 SNP 位点。

在 SNP 检测过程中,人为或者机器缘故会导致 SNP 位点分型不成功,被标记为 NC。本文对原始实验数据进行分型成功率检查,并对原始数据中分型不成功的 SNP 位点,通过数据清洗和数据填充进行校正。同时,数据分析还需要将非数值的 SNP 分析值进行数据编码,方便数据分析的进行。

1)分型成功率检查

本文设置分型成功率阈值为 0.9。统计某一个 SNP 位点的分型成功率,低于阈值,则直接将 SNP 位点从原始数据中清洗掉。

对于分型成功率未达到 100% 但又高于阈值的 SNP,将其值为 NC 的位置进行缺失处理。本文采取的方法是:当遇到 NC 时,使用同类型样本该位点统计比例最高的分型值填充到该位置。这种方法尽可能保持了原数据趋向性,同时完善了数据的可用性。

2)数据编码

为使数据分析更加方便,按照下述编码方式将 SNP 数据进行编码,将分型值 AA 转换为 0,AB 转换为 1,BB 转换为 2。样本类别标示为:正常样本标记为 +1,非正常样本标记为 -1。具体编码模式如表 1 所示。

表 1 SNP 数据编码表

分型值	AA	AB	BB	NC	正常样本	非正常样本
编码	1	2	3	0	+1	-1

3.2 SNP 特征选择方法参数设置及实验比对

3.2.1 参数设置

本文使用 Relief-SVM 和 SVM-RFE 分析以及 Relief-SVM-RFE 方法对 SNP 数据进行特征选择,并对结果进行比对和分析。

实验中,Relief 无关阈值设置为 0.5;SVM-RFE 算法在 LIBSVM 2.91 源代码的基础上编写完成[13],SVM 分类机采用 C-SVC,核函数选用 RBF,惩罚参数 C 为 1,核参数 gamma 为 1/feature,二次划分的分组个数 K 设置为 10,即采用二次划分时,

每次迭代删除特征个数为总特征的 1%。

3.2.2　关键 SNPs 评估

本文采用十字交叉验证方法选择关键 SNPs,通过比较三种方法获得关键 SNPs 的集合规模以及最高分类准确率,判断哪种方法更适合作 SNP 数据分析。

十字交叉验证的办法是将 Relief-SVM、SVM-RFE、Combined 方法获取的前一个最优 SNP 进行十字交叉验证,然后选取前两个最优 SNP 进行十字交叉验证,依此类推至 n 个(n = 1,2,3…),分别获取其分类准确率,直到三种方法获得的分类准确率基本上达到稳定为止。

3.2.3　算法结果分析

由于 filter 式算法和 wrapper 式算法在时间复杂度上的差异非常明显,因此实验将三种方法在获取关键 SNPs 集合的耗时进行了对比。三种方法获取的关键 SNPs 集规模、耗时和分类准确率的对比结果如表 2 所示。

表 2　三种方法关键 SNP 集合规模与分类准确率及耗时对比

方法	比较指标		
	关键 SNPs 集合规模	分类准确率/%	耗时/h
Relief-SVM	38	93	4.2
SVM-RFE	33	98	8.5
Relief-SVM-RFE	31	96	5.2

由表 2 可以看出,组合式 SNP 特征选择方法在特征规模为 31 时达到了分类准确率的稳定,并获得了与 Relief-SVM 相同的分类准确率 95%。由此可见,组合式特征选择能获得比 Relief-SVM 更加精确的关键 SNPs 范围。SVM-RFE 虽然以 33 个特征集规模获得了高出组合式 SNP 特征选择 3% 的分类准确率,但是耗时却比组合式 SNP 特征选择高出 65%。

此外,使用十字验证确定 SNPs 范围时,随着关键 SNPs 规模的扩大,对关键 SNPs 的分类准确率变化趋势进行了统计和比较。三种方法的关键 SNPs 集合规模与分类准确率变化规律如图 2 所示。

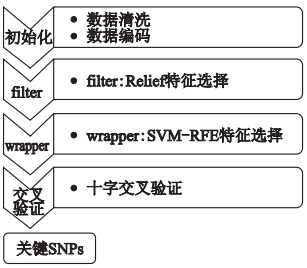

图1　基于Relief和SVM-RFE的组合式SNP特征选择方法

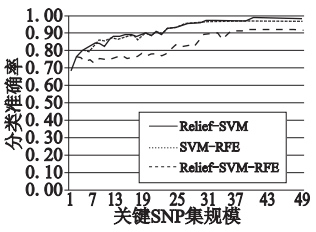

图2　关键SNPs集合规模与分类准确率变化图

4　结束语

SNP 特征选择旨在找到与疾病相关联的 SNPs 集合。该过程既要解决高维小样本 SNP 数据带来的维数灾难问题,又要充分考虑遗传疾病病理的复杂性。对此,本文提出了将 Relief 和 SVM-RFE 算法有效结合的 SNP 组合式特征选择方法。首先使用 Relief 算法对原始 SNP 集进行预筛选,得到去除无关特征的次高维 SNP 集合;然后基于 SVM-RFE 算法找到具有最高分类权重的关键 SNPs。实验证明,该方法在 SNP 数据分析上具有明显优于单独使用 SVM-RFE 算法的性能,优于 Relief-SVM 的分类准确率。

虽然该方法比现有的 SNP 数据分析方法在算法效率和分类效果上有明显的提高,但仍有一些可以提高的地方。目前已有研究者将遗传算法、蚁群算法、决策树、并行算法等算法融入 wrapper 算法中进行组合优化,从而提高筛选性能和精确度[14,15]。此外,结合 filter 和 wrapper 方法的组合式特征选择是串行设计。Filter 和 wrapper 的特征筛选比例如何控制,才能保证 filter 方法不会过滤掉关键特征,且提供最小规模数据集进行 wrapper 筛选,也值得研究探索。最后,实验中所获取的关键 SNPs 在医学上是否具有确切的生物学意义,与疾病之间是否存在关联关系,还有待进一步研究验证。

参考文献:

[1] 王威. 人类基因组单核苷酸多态性研究与疾病三级防护体系[J]. 国外医学卫生学分册,2006,33(6):321-325.

[2] 逄锦钟,钦伦秀,汤钊猷. 人类基因组单核苷酸多态性及其医学应用[J]. 肿瘤,2005,25(4):401-403.

[3] 邹喻萍,葛颂. 新一代分子标记——SNPs 及其应用[J]. Biodiversity Science,2003,11(5):370-382.

[4] 严卫丽. 复杂疾病全基因组关联研究进展[J]. 遗传,2008,30(5):543-549.

[5] SHAH S C, KUSIAK A. Data mining and genetic algorithm based on gene/SNP selection[J]. Artificial Intelligence in Medicine,2004,31:183-196.

[6] PARDI F, LEWIS C M, WHITTAKER J C. SNP selection for association studies: maximizing power across SNP choice and study size [J]. Annals of Human Genetics,2005,69(6):733-746.

[7] PHUONG T M, LIN Z, ALTMAN R B. Choosing SNPs using feature selection[J]. Journal of Bioinformatics and Computational Biology,2006,4(2):241-257.

[8] 毛勇,周晓波,夏铮,等. 特征选择算法研究综述[J]. 模式识别与人工智能,2007,20(2):211-218.

[9] KIRA K, RENDELL L A. The feature selection problem : traditonal methods and a new algorithm[C]//Proc of the 10th National Conference on Artificial Intelligence. Michigan: AAAI Press,1992:129-134.

[10] GUYON I, WESTON J, BARNHILL S, *et al*. Gene selection for cancer classification using support vector machines[J]. Machine Learning,2002,46(1-3): 389-422.

[11] VAPNIK V N. Statistical learning theory[M]. New York: Wiley, 1998.

[12] 汪伟,刘红. 基于遗传算法与支持向量机的基因微阵列分析[J]. 中国组织工程研究与临床康复,2010,14(17):3099-3013.

[13] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[EB/OL]. (2001). http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[14] DING Yuan-yuan, VILKINS D. Improving the performance of SVM-RFE to select genes in microarray[J]. BMC Bioinformatics,2006,7(2):S12.

[15] TAN Feng, FU Xue-zheng, ZHANG Yan-qing, *et al*. Improving feature subset selection using a genetic algorithm for microarray gene expression data[C]//Proc of IEEE Congress on Evolutionary Computation. Vancouver: IEEE Congress,2006:2529-2534.