◎博士论坛◎

面向局部特征的支持向量机递归特征消除

杨 帆[1],王华珍[2],米 红[1]
YANG Fan[1],WANG Hua-zhen[2],MI Hong[1]

1 厦门大学 信息科学与技术学院,福建 厦门 361005
2.华侨大学 计算机科学与技术学院,福建 厦门 361021
1.School of Information Science and Technology,Xiamen University,Xiamen,Fujian 361005,China
2.College of Computer Science and Technology,Huaqiao University,Xiamen,Fujian 361021,China
E-mail:yang@xmu.edu.cn

YANG Fan,WANG Hua-zhen,MI Hong.Support Vector Machine—Recursive Feature Elimination for localized feature selection.Computer Engineering and Applications,2009,45(28):1-5.

Abstract: Support Vector Machine—Recursive Feature Elimination (SVM-RFE) is one of state-of-the-art method for gene selection.SVM-RFE was originally designed to solve binary feature selection problems and has been extended to solve multiclass problems in several recent studies.This paper illustrates the limitations of the present multi-class gene selection methods from the perspective of Pareto Optimum,describes a new procedure for selecting significant genes for each class,and proposes a new implementation for SVM-RFE. Experiments on 8 cancer and tumor gene expression dataset demonstrate its superiority over two other RFE methods.By considering each class during the gene selection stages,the new method can identify genes leading to more accurate classification.

Key words: gene expression data;multi-classification;gene selection;support vector machine

摘 要:基于支持向量机的递归特征消除(SVM-RFE)是目前最主流的基因选择方法之一,是为二分类问题设计的,对于多分类问题必须要进行扩展。从帕累托最优(Pareto Optimum)的概念出发,阐明了常用的基因选择方法在多分类问题中的局限性,提出了基于类别的基因选择过程,并据此提出一种新的SVM-RFE设计方法。8个癌症和肿瘤基因表达谱数据上的实验结果证明了新方法优于另两种递归特征消除方法,为每一类单独寻找最优基因,能够得到更高的分类准确率。

关键词:基因表达谱;多分类问题;基因选择;支持向量机
DOI:10.3778/j.issn.1002-8331.2009.28.001 **文章编号**:1002-8331(2009)28-0001-05 **文献标识码**:A **中图分类号**:TP301

1 引言

基因芯片技术对医疗诊断和基因水平上的研究具有重要意义,由其产生的基因表达谱数据包含巨大的信息量,设计科学有效的分析方法,从基因表达谱数据中获取更多的知识,是生物信息学研究中的热点和前沿问题。在研究中常使用已知疾病类型的基因表达数据,构建分类模型,使用病人组织样本的基因表达数据,以判断其病理状况,例如,是否患有癌症,是癌症的哪种亚型,等等[1]。基因芯片一次采集成千上万个基因的表达水平信息,造价昂贵,因此样本数通常在数十到数百,这样一个典型的"高维小样本"问题,使得传统统计方法不再可行,于是机器学习方法如决策树[2]、神经网络和支持向量机[3]在该领域得到了广泛重视和应用。

特征选择对于基因表达数据分析有双重意义:(1)消除"维数灾(curse of dimensionality)"对学习机器的影响,提高学习机器的泛化性能和计算效率;(2)消除不相关特征和冗余特征,获得更有价值的信息,具体来说,就是获得对于分类问题有最佳判别能力的基因,这些基因也被认为具有生物学意义,生物学家或医学专家可进一步设计针对性的实验对其展开研究,并获得对疾病的产生和发展机制的深刻理解。由于基因芯片上包含成千上万个基因,从生物学和医学研究的角度来说,寻找少量的关键基因非常重要,在癌症和肿瘤基因表达谱数据分析中,特征选择又常被称为基因选择(gene selection)[4]。

机器学习中常用的特征选择方法分为过滤(filtering)和封装(wrapper)两大类方法。前者不依赖于特定的分类器,仅依据样本固有的信息,选择能达到最佳类可分性的特征,如基因选择中常使用的S2N(Signal to Noise)法和BW(Between-categories to Within-category sums of squares)算法[5];后者通过评估分类器的性能,如最大程度提高分类的准确率,得到该准确

基金项目:福建省自然科学基金(the Natural Science Foundation of Fujian Province of China under Grant No.2009J05153)。
作者简介:杨帆(1982-),博士研究生,研究方向:数据挖掘,机器学习方法和生物信息学;王华珍(1975-),博士,讲师,研究方向:智能计算,数据挖掘,机器学习和智能控制;米红(1962-),教授,博士生导师,研究方向:数据挖掘,系统工程。
收稿日期:2009-07-14 修回日期:2009-08-17

率下的“最优”特征子集,或者是最大化学习机器的泛化性。封装方法中,SVM 能很好地应对高维小样本数据,与其他方法相比能在基因表达数据上得到较好的泛化性能,因此人们认为,利用 SVM 来选择合适的特征子集能够取得相对较好的效果,其中基于支持向量机的递归特征消除(SVM-RFE)是目前国际上最主流的基因选择方法之一[4-6]。

SVM 最初被设计用于二分类问题,因此 SVM-RFE 方法对于多分类必须要加以改进[5-6]。常用的方法是将 m 个类别的多分类问题分解成 m 个“一对多”的二分类问题,针对每个二分类问题训练 SVM 模型,并进行特征选择,对各子 SVM 的特征选择的结果进行组合后消除 1 个判别能力最弱的基因,并重复该过程。

将从 Pareto 最优的角度说明这种设计方法并不能达到所谓最优的效果,所选出的特征不是 Pareto 意义上的“最优”,不同类别所选出的特征并不具备可比性,据此提出了基于类别的特征选择的概念,并在此基础上对常用的多分类 SVM-RFE 方法进行了改进。

2 基于支持向量机的递归特征消除

SVM 是针对二分类问题设计的,算法在高维空间中寻找一个最大间隔超平面作为决策函数将训练样本分开[3]。在 p 维原始输入空间中,给定 N 个训练数据$\{x_k,y_k\}\in \mathrm{R}^p\times\{-1,+1\}$,其中 x_k 表示训练样本,y_k 表示类标,SVM 首先通过核函数 $\Phi(x)$ 将训练样本投影到一个高维空间,并计算决策函数 $f(x)=\langle w,\Phi(x)\rangle+b$。

对于线性可分的样本集,采用线性核函数,并使得:

$$y_k[(w\cdot x_k)+b]-1\geqslant 0,k=1,\cdots,N \tag{1}$$

该约束条件使得经验误差趋向于零,体现了经验风险最小化准则的思想。在样本集线性可分的情况下,分类间隔等于 $2/\|w\|$,寻找最大间隔超平面等价于使 $\|w\|^2$ 最小,在约束条件下最小化目标泛函:

$$J=\frac{1}{2}\|w\|^2 \tag{2}$$

最小化 $\|w\|^2$ 使函数集 VC 维尽量小,具有较好的泛化能力,体现了结构风险最小化准则的思想。

如果样本集不是线性完全可分的,一些样本点无法正确分类,此时增加松弛变量 $\xi_k\geqslant 0$,约束条件变为:

$$y_k[(w\cdot x_k)+b]-1+\xi_k\geqslant 0,k=1,\cdots,N \tag{3}$$

同时,优化的目标变为:

$$J=\frac{1}{2}\|w\|^2+C\sum_{k=1}^{N}\xi_k \tag{4}$$

其中常数 $C>0$,控制错分样本的惩罚程度。根据拉格朗日理论,SVM 模型的求解可化为不等式约束下凸二次函数寻优的问题,存在唯一的全局最优解。权向量的形式为 $w=\sum_{k=1}^{N}\alpha_k^* y_k\Phi(x_k)$,其中,$\alpha^*$ 通过求解对偶二次规划问题得到:

$$\max_{\alpha} W(\alpha)=\sum_{k=1}^{N}\alpha_k-\frac{1}{2}\sum_{k,l=1}^{N}\alpha_k\alpha_l y_k y_l(\langle\Phi(x_k)\cdot\Phi(x_l)\rangle+\frac{1}{C}\delta_{k,l})$$

$$\text{s.t.}\ \sum\alpha_k y_k=0,0\leqslant\alpha_k\leqslant C,\forall k \tag{5}$$

$$\delta_{k,l}=\begin{cases}1 & \text{if } k=l\\ 0 & \text{if } k\neq l\end{cases}$$

通常只有很少数的 α_k^* 不为零,只有极少数支持向量(support vector)体现在决策函数中,由此可知,权值 w 是支持向量的线性组合。

考虑第 i 个特征对目标函数 J 的影响,由泰勒展开可得:

$$\Delta J(i)=\frac{\partial J}{\partial w_i}\Delta w_i+\frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2+\cdots \tag{6}$$

在目标函数 J 的最优点上,一阶项为零,因此只考虑二阶项,同时假设样本集线性可分,可得:

$$\Delta J(i)=(\Delta w_i)^2 \tag{7}$$

当移除第 i 个特征时,$\Delta w_i=w_i$,因此第 i 个特征对目标函数的影响大小为 $c_i=(w_i)^2$。当使用 SVM 作为特征排序的依据时,可采用 w_i^2 作为各个特征重要性的度量,w_i^2 越大,则该特征越重要,反之亦然。经验研究表明,对于基因表达数据这样的高维小样本数据来说,采用线性核即可取得良好的分类效果,因此基因选择中使用的 SVM 一般均采用线性核[4]。

算法 1 SVM-RFE

输入:N 个训练样本的初始训练集

$X_0=\{x_k\}_N,x_k\in\mathrm{R}^p,Y=\{y_k\}_N,y_k\in\{-1,+1\}$

输出:按重要性降序排列的特征序列 F。

初始化:$F=[\ ]$,特征子集 $S=[1,\cdots,p]$。

当 $S\neq\phi$ 时,重复以下步骤:

(1) 使用训练数据 $X=X_0[:,S]$生成线性 SVM;

(2) 根据权值 w_i 计算排序准则 $R_i=(w_i)^2,i\in S$;

(3) 找到排在最后的特征的序号:$t=\arg\min_i(R)$,其中 $R=\{R_i\}$,将其加入特征序列 $F=[t,F]$;

(4) 得到余下的特征子集 $S=S\backslash t$。

3 Pareto 最优与多分类特征选择

多分类问题的特征选择一直是机器学习中的一个难题,目前还没有完善的解决方法。从多目标优化(Multi-objective Optimization Problems,MOPs)的角度来说,可以把多分类特征选择表述成:寻找一个特征子集,能同时最大化学习机器在各个类别上的泛化性。而特征子集的寻找,是以排序准则 R 为依据的,如果假定只要通过排序准则 R 选择合适的特征,就可以最优化各子分类问题的解,并且假定,R 越大,特征越优,则实质上已经将特征选择问题转化为目标函数为 R 的最优化问题。

各个类别之间存在质的区别,满足该条件的最优特征子集在大多数场合下不存在,而且难以找到,因此不管是封装法还是过滤法,都是选取那些在所有类别上平均效果最好的特征。例如 BW 法,可以看成是 S2N 法在多分类问题上的扩展。

对于 m 类别的多分类问题,SVM-RFE 首先将分类问题分解成 m 个“一对多”的子分类问题,建立 m 个 SVM 子分类器,

$$\min\ J_k=\frac{1}{2}\|w\|+C_k\sum_{i=1}^{n}\xi_{ik} \tag{8}$$

$$\text{s.t.}\quad y_i(w_k\cdot x_i+b_k)\geqslant 1-\xi_{ik}$$

$$\xi_{ik}\geqslant 0,i=1,\cdots,n,k=1,\cdots,m$$

在重新训练 SVM 的过程中,得到 m 个特征重要性的序列 $R(k)=\{w_{ik}^2,i=1,2,\cdots,s\}$,其中 s 为当前所选特征子集的大小,$k=1,2,\cdots,m$。目前绝大部分研究中的做法是选取某个规则对这 m 个特征重要性序列进行融合。例如文献[6]对每一类选出的特征子集进行融合,对第 k 类按照其序列 $R(k)$ 选取一个特

征子集 S_k,取下一轮训练时的特征子集为:$S=\bigcup_{k=1}^{c} S_k$;更多的研究对每一类的特征重要性序列进行融合,例如 X Zhou 等[6]对每一个特征在所有类中的重要性求平均,建立新的排序规则:$R(i)=\sum_{k=1}^{m} w_{ik}^2$,X Chen 等[7]在采用 SVM-RFE 对文本数据进行特征选择时,取每一个特征在所有类中的重要性的最大值为其最终的重要性,即 $R(i)=\max_k w_{ik}^2$。

不难发现,这些方法在递归的每一步中得到的特征子集都是为整个分类问题设计的,对于每一个子分类问题都不是最优的。对于文献[6]中描述的方法,想象一种可能情形:如果为每一个子分类器消除掉的特征都不相同,那么求并集的过程将得到全集,此时递归过程永远不会终止。X Zhou 等提出的方法实质是将 m 目标优化问题通过平均加权法转化为单目标优化问题:$\min J=\sum_{k=1}^{m} J_k$,并通过对单目标优化问题的敏感性分析得出新的排序准则。包括文献[7]方法在内,对于每一个子分类器来说,这些方法所得到的特征子集都不是最优的。进一步的,由于递归特征消除过程的"嵌套性(nested)",在不断重新训练 SVM 模型并进行特征消除的过程中,这种偏差可能会不断放大。

从 MOPs 的角度来看,单目标优化问题的最优解可以简单地定义,而多目标优化的结果却是一组均衡解,即所谓的 Pareto 最优解。多数情况下各个子目标可能是相互冲突的,某子目标的改善可能引起其他子目标性能的降低,即同时使多个目标均达到最优解一般是不可能的。在多分类问题中,各子目标函数之间实际上同样存在冲突和矛盾,因此,最优特征子集的寻找也会存在矛盾和冲突。从这个角度来看,上述方法(包括 BW 法在内)都是不合理的。下面简要介绍 Pareto 最优解[8]的相关概念,并用 2 个例子来说明目前基因选择研究中存在的这个问题。

定义 1 Pareto 支配(Pareto dominance)　称向量 $\boldsymbol{u}=(u_1,\cdots,u_k)$ 支配 $\boldsymbol{v}=(v_1,\cdots,v_k)$,记为 $\boldsymbol{u}\prec\boldsymbol{v}$,当且仅当 $\exists j\in\{1,\cdots,k\}:u_j<v_j$,$\forall i\in\{1,\cdots,k\}:u_i\leqslant v_i$。

定义 2 Pareto 最优(Pareto optimality):决策变量 $x\in\Omega$ 成为 Ω 上的 Pareto 最优解(Pareto optimal solution)的充要条件是,当且仅当不存在 $y\in\Omega$,使得 $\boldsymbol{v}=F(y)=(f_1(y),\cdots,f_k(y))$ 支配 $\boldsymbol{u}=F(x)=(f_1(x),\cdots,f_k(x))$。

定义 3 Pareto 最优解集(Pareto-optimal set):Pareto 最优解集由所有 Pareto 最优解组成,记为 $P_S=\{x\in\Omega|$不存在 $y\in\Omega:F(y)\prec F(x)\}$。

Pareto 最优解是不存在比这个解至少一个目标方案更好而其他目标非劣的解,也就是不可能优化其中部分目标而使其他目标不劣化。因此,Pareto 最优解集的元素,彼此之间无法进行一般意义上的性能优劣的比较。可以仿照 Pareto 最优的概念描述想要寻找到满足何种性质的基因:给定排序准则 R,称基因 g_i 支配基因 g_j,记为 $g_i\prec g_j$,当且仅当 $\exists k_0\in\{1,\cdots,c\}:R(i,k_0)>R(j,k_0)$,$\forall k\in\{1,\cdots,c\}:R(i,k)\geqslant R(j,k)$,那么基因 $g\in F$ 成为 F 上的 Pareto 最优基因的充要条件是,当且仅当不存在 $g'\in F$,满足 $g'\prec g$。

从这个角度来看,目前国内外的基因选择研究,将该问题化为多目标优化问题,而实际采取的求解方法却无法得到 Pareto 意义下的最优解。

例 1　考虑一个 3 分类问题,有 4 个特征 A、B、C 和 D。给定排序准则 R,就每个子分类问题生成 3 个排序 R_1、R_2 和 R_3,4 个特征的目标值分别为 $A=(0.50,0.80,0.95)$,$B=(0.10,0.20,0.10)$,$C=(0.31,0.29,0.40)$,$D=(0.60,0.81,0.61)$,在三维空间中位置如图 1 所示。不难看出,其中 A 和 D 是 Pareto 最优特征,B 和 C 不是。

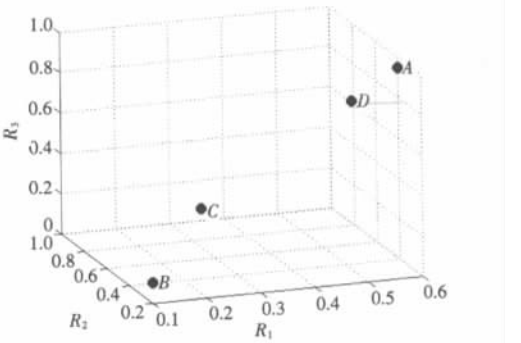

图 1　一个 3 分类问题中的 Pareto 最优特征的示例

例 2　Scott A.Armstrong 等 2001 年发表在 Nature Genetics 上的白血病(Leukemia)数据集,原始数据集有 72 个样本,11 225 个基因,3 个子类:急性髓细胞性白血病(Acute Myelogenous Leukemia,AML),急性淋巴细胞性白血病(Acute Lymphoblastic Leukemia,ALL),混合谱系白血病(Mixed-Lineage Leukemia,MLL)。采用线性 SVM-RFE(惩罚系数取 C=100)进行特征选择,每次消去 10%的基因,当基因数小于 100 时,每次消去 1 个特征,经过 135 次特征消除后,剩下 3 个基因 A(第 5547 号基因)、B(第 6887 号基因)和 C(第 10038 号基因),在 3 个子分类器上得到的排序值约为(0.85,0.86,0.42),(0.72,0.74,0.32),(0.40,056,0.66)。记文献[6,8]所示的两种多分类 SVM-RFE 算法分别为 RFE1 和 RFE2,得出的最终排序为:

RFE1:A,C,B

RFE2:A,C,B

显然,从 Pareto 最优的观点来看,A 和 B 是非劣解的,而 C 被 A 支配。但是在两种算法的输出结果中,C 的排序高于 B,根据递归特征消除的策略,此时要消去特征 B,而不是 C。这与 Pareto 最优解集的观点相矛盾。进一步地分析发现:(1) B 对于第 1 类和第 2 类的贡献不如 A 和 C 显著,但对于第 3 类的分类意义特别显著,是第 3 类非常重要的一个判别基因,从知识发现的角度来看,B 是所关心的基因;(2) A 和 C 对于第 1 类和第 2 类的判别意义都很显著,且 $A\prec C$,从消除冗余属性的角度来看,应该消除 C。因此,此时 RFE2 和 RFE1 的结果都不理想。

4　基于类别的多分类 SVM-RFE 过程

由以上分析可知,目前已有的多分类基因选择方法存在不合理之处,所选中的基因往往不满足 Pareto 最优的定义。多分类 SVM-RFE 算法目的在于为多个子类寻找一个共同的最优特征子集,其本质是一个多目标优化问题,而实际上采取的做法却并不满足 Pareto 最优的概念。同时,求解 Pareto 最优基因子集是一个非常困难的过程,最终得到的基因子集会包含很多基因,达不到研究的本质目的。

如果换个视角来看待基因选择问题,会发现,实际上这种共同的最优特征子集未必存在。目前很多研究的做法实质上假

设共同的最优特征一定存在，寻优的方向也只考虑这些特征，而不考虑一些可能对于某些单类的识别更好的特征。该文视基因选择为一个多目标优化问题是非必要的，可以直接将基因表达谱数据的 m 分类问题按“一对多”方式分解成 m 个子问题，分别求解最优特征子集，即将其分解成 m 个单目标优化问题，从而避免求解 Pareto 最优子集的难题。

利用线性 SVM-RFE 来完成这一设计，即生成 m 个 SVM 子分类器，分别进行递归特征消除，为每个子分类问题得到递归过程中每一步的排序，然后消除排序最靠后的基因。

算法 2 多分类 SVM-RFE

(1) 将 m 类分类问题分解为 m 个“一对多”的二分类问题；

(2) 对每一个二分类问题调用算法 1，得到 m 个基因排序。

很自然地，与多分类 SVM-RFE 求解出特征子集相比，这 m 个优化特征子集更能够提升每一个子分类器的性能，最大程度上针对每一个子类找到紧致的相关子集。性能提升后的子分类器可以视为“子类专家”，整个分类问题的性能也将得到提升。

当获得 m 个基于类别的基因排序后，就获得了针对每个类的基因的重要性信息，但是在研究中，还必须根据重要的基因构建分类模型，以供诊断之用，并评价所选基因的整体可靠性。由于各个子问题所选取的特征子集不同，原问题被分解到 m 个不同的特征子空间中，在各子空间中样本点的位置也都不同，因此各子分类问题已经变成完全不同的分类问题。对于 SVM 子分类器来说，即使采用相同的核函数，这时它们输出的决策值也是无法直接比较的。此时如何组合在各个不同特征空间中训练出的 SVM 模型，成为要解决的问题之一。在不同的特征空间中训练 SVM 的子分类器，并通过组合这些子分类器对基因组织样本进行分类，目前此类研究未见报道。

虽然在不同特征空间中子分类器的输出决策值无法直接比较，但是从概率的意义来看，却可以估计出每个“一对多”的子问题中，样本属于该类的似然概率，因此提出使用后验概率输出支持向量机 (Posteriori Probability SVM)[9-10]，通过计算测试样本属于各子类的后验概率，依据“Max-Win”的策略，选择可能性最大的子类作为输出。

Wahba 和 Platt[9]提出通过直接从 SVM 输出的判别函数映射得到后验概率，从而避开对类条件概率的密度估计。进一步地，Platt[10]用含有两个参数 A 和 B 的 Sigmoid 函数来逼近后验概率：

$$\Pr(y=1|x) \approx P_{A,B}(f) \equiv \frac{1}{1+\exp(A\cdot f+B)} \tag{9}$$

其中 $f=f(x)$，实际情况下用交叉验证得到 $f(x)$ 的估计值 f_i。对于给定的训练样本集给定 N 个训练数据 $\{x_k,y_k\}\in R^n\times\{-1,+1\}$，设正类样本数为 N^+，负类样本数为 N^-。Platt 指出可通过求解以下最大似然估计问题来得到最优的参数 (A^*,B^*)：

$$\max_{z=(A,B)^T} F(z)=-\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i)) \tag{10}$$

其中，$p_i=P_{A,B}(f_i)$，$t_i=\begin{cases}\dfrac{N^+ +1}{N^+ +2}, y_i=+1\\ \dfrac{1}{N^- +2}, y_i=-1\end{cases}$。

据此可求出参数 A、B，进而得到二分类 SVM 的后验概率输出。具体求解过程在此不再赘述，具体实验中，采用 Libsvm 实现后验概率的求解。

5 实验结果分析

通过实验来比较 SVM 在 RFE1、RFE2 以及该文所提出的算法 (记为 RFE3) 所选出的特征子集上的性能。实验平台采用 Matlab7.5，SVM 的实现和修改均采用 Libsvm2.9，RFE 过程所训练的 SVM 模型采用线性核，惩罚系数取 C=100，选取 8 个国际公用的多分类基因表达谱数据集[11] (http://www.gems-system.org)，数据描述见表 1。由于 SVM-RFE 的过程非常耗时，选取不同大小的特征子集为{1,2,3,4,5,6,7,8,9,10,20,30,40,50,60,70,80,90,100,200,300,400}，每次递归中消除约 10%的特征，当剩下的特征数目小于等于 100 时，每次消除 1 个特征。实验方式采取 10 重交叉，将每一折数据作为测试集，其他数据作为训练集，为了避免“选择偏差 (selection bias)”，在训练集上进行特征选择，在特征子集上训练 SVM，对测试数据进行预测，记录 10 重交叉的平均准确率。重复该过程 100 次，取最后的平均值。

表 1 8 个多分类基因表达谱数据集的详细描述

数据集	诊断任务	数据集具体内容			
		样本数	基因数	类别	基因数/样本数
Brain Tumor 1	5 种脑部肿瘤	90	5 920	5	66
Brain Tumor 2	4 种脑神经胶质瘤	50	10 367	4	207
Leukemia 1	3 种急性白血病亚型	72	5 327	3	74
Leukemia 2	4 种急性白血病亚型	72	11 225	3	156
Lung Cancer	4 种肺部癌症肿瘤	203	12 600	5	62
SRBCT	儿童小轮蓝细胞瘤	83	2 308	4	28
Tumor 1	9 种人类肿瘤	60	5 726	9	95
Tumor 2	11 种人类肿瘤	174	12 533	11	72

图 2~图 9 展示了在 8 个数据集上使用“一对多”线性核 SVM 的平均预测准确率随特征子集大小改变的变化曲线，横轴代表不同大小的特征子集，为了显示方便，将其在坐标轴上均匀排列，纵轴为平均准确率，其中曲线“---”表示 SVM 在 RFE1 所求的特征子集上的分类效果，“-·-”表示在 RFE 2 上的分类效果，曲线“-*-”表示基于类别的 SVM-RFE 算法，记为 RFE3。

观察实验结果，可以得出以下结论：

(1) 从总的效果上说，除 Brain Tumor 1 和 Leukemia 2 数据外，RFE3 在其他 6 个数据集上总的效果都显著优于 RFE1 和 RFE2；在所有的 8 个数据集上，包括 Brain Tumor 1 和 Leukemia 2，当特征数小于 10 时，RFE3 的效果均显著优于其他算法，其中 Tumor 2、SRBCT、Lung Cancer、Leukemia 1、Leukemia 2 上，当仅取 10 个特征时的分类器平均准确率高达 91%、99%、93%、95.2%、94.7%，而 RFE1 和 RFE2 中性能最好的为 80%、95.1%、88.5%、93%、93.6%。少部分基因恰恰是基因表达研究中最关注的，具有良好判别能力的基因很可能具有很强的生物学意义。

(2) 对于 Brain Tumor 2 和 Tumor 1 这两个难分的数据集，RFE1 和 RFE2 的性能很差，而 RFE3 保持了相对较好的性能，例如在 Tumor 1 数据上，取 1 个基因时 RFE1 和 RFE2 的结果约为 21%和 19%，而 RFE3 约为 36%，当特征数小于 80 时差别非常明显，随着特征子集的增大，两者性能接近，但 RFE3 依然显著优于 RFE1 和 RFE2。在 Tumor 2 上，当取 1 个基因时，RFE1 和 RFE2 的结果仅仅约为 19.5%和 25%，而 RFE3 的结果高达 56%，这说明对于某些数据集来说，确实不存在对于

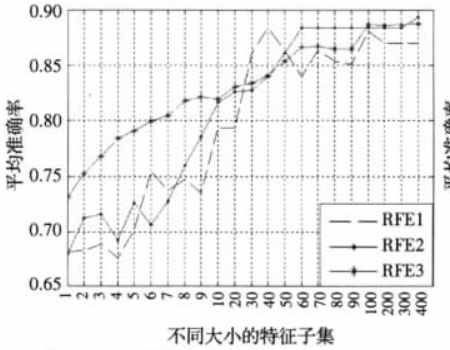

图 2　三种不同 RFE 算法在 Brain Tumor 1 数据上的效果

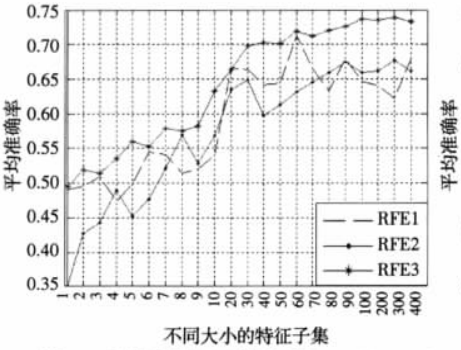

图 3　三种不同 RFE 算法在 Brain Tumor 2 数据上的效果

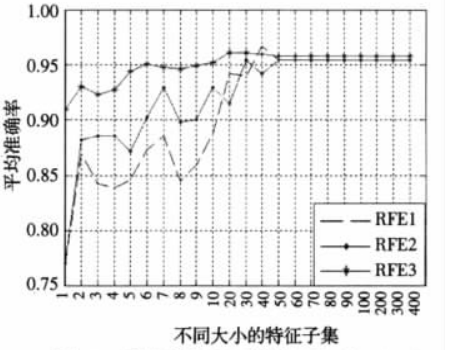

图 4　三种不同 RFE 算法在 Leukemia 1 数据上的效果

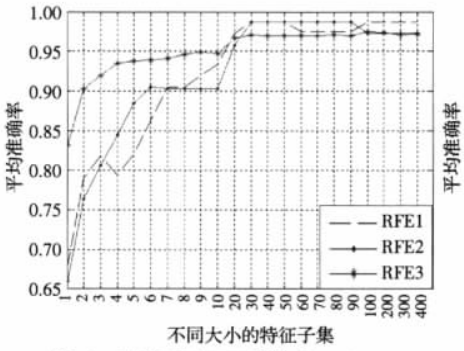

图 5　三种不同 RFE 算法在 Leukemia 2 数据上的效果

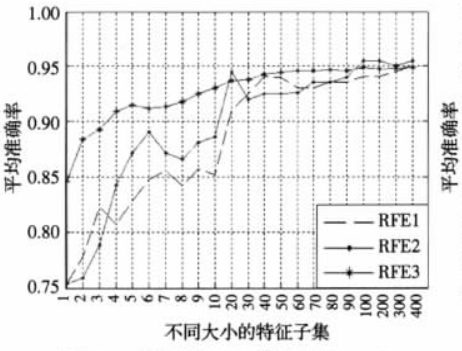

图 6　三种不同 RFE 算法在 Lung Cancer 数据上的效果

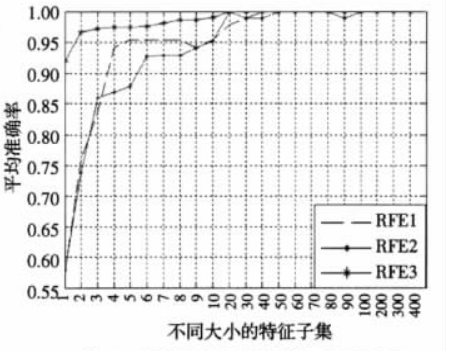

图 7　三种不同 RFE 算法在 SRBCT 数据上的效果

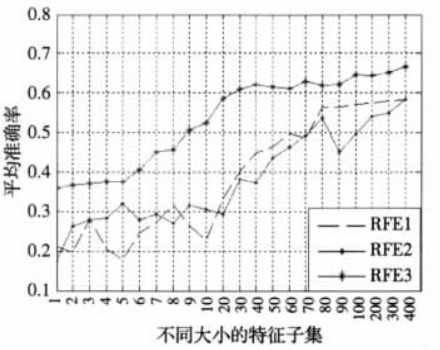

图 8　三种不同 RFE 算法在 Tumor 1 数据上的效果

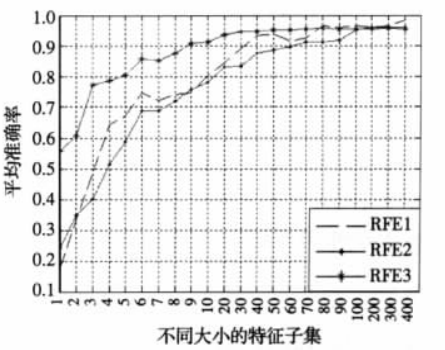

图 9　三种不同 RFE 算法在 Tumor 2 数据上的效果

所有类别均有良好判别能力的基因，此时根据 RFE1 和 RFE2 算法求出来的“统一”的基因子集,无法认定其具有一定的生物学意义。

(3) 随着特征数的逐渐增大,三种方法的性能也逐渐接近，这是由于此时所求出的特征子集的交集在逐步增大;同时,很多基因都加入到特征子集中,由于基因之间的高度相关性,分类器所接收的信息量已足够大,因此三种算法的性能慢慢接近。

6　结论

从 8 个公用数据集上的实验结果,可以得出以下两个结论：

(1) 从基因表达数据分析的角度上来说,对于多分类的基因表达谱数据,分类别的特征选择的效果优于“全局寻优”。其根本原因在于,“全局寻优”假定存在部分基因在所有的类别中均有优良判别能力。不排除在某些基因表达数据集上,可能有这样的极少数基因存在,但由于基因表达谱数据的样本数极其有限,很难找到这样的基因;同时,针对单个类别来寻找判别基因,更加符合逻辑,也具有很强的现实使用价值。因此,提出对多分类基因表达数据进行分类别的特征选择，具有针对性,且化繁为简,降低了原问题的求解难度。

(2) 从算法的角度上来说,SVM-RFE 在本质上仍然是一个针对二分类问题的特征选择方法，在应用到多分类问题上,仍然有绕不开的难题和瓶颈;多分类 SVM 采用“分解-组合”的思路,多分类 SVM-RFE 算法在此基础上遵循“分解-组合-分解”的思路，从对信息的处理方式来看，其步骤比多分类 SVM 更多,理论上的难度也大得多。所提出的分类别的 SVM-RFE 算法简化了这一过程,将其还原为“分解-组合”,降低了这一过程的难度,具有现实使用价值,可推广到其他应用中。

参考文献：

[1] Valafar F.Pattern recognition techniques in microarray data analysis a survey[J].Annals of the New York Academy of Sciences, 2002,980(1):41-64.

(下转 30 页)

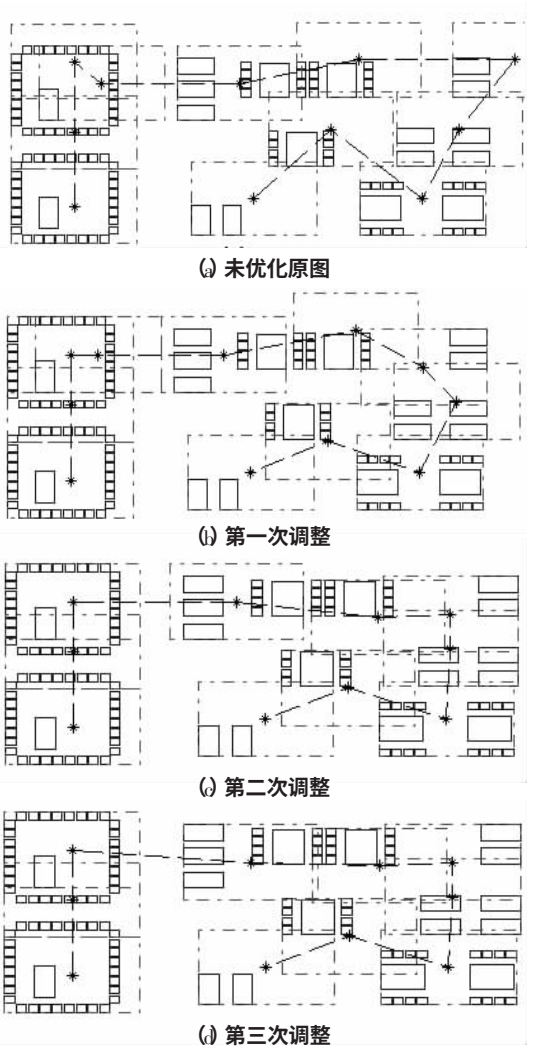

图 5 迭代三次的路径变化

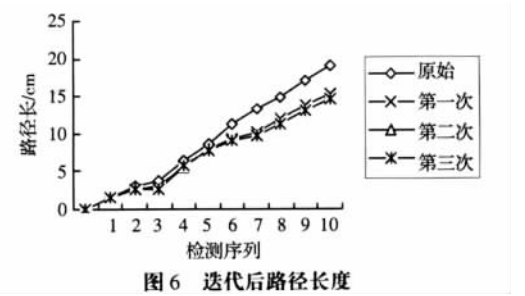

图 6 迭代后路径长度

5 结论

针对带覆盖约束的平面光学路径规划问题，提出了一种采用迭代的优化方法。通过分析约束边界线段与前后窗体的位置关系，证明在单窗口矩形约束范围内获得最小路径和的点在边界线段的某个位置。采用这种单窗口的调节方法，可以迭代解决多窗口的最小路径问题。最后，针对具有 124 个检测对象图进行实验，证明了该方法的有效性。该算法虽然针对多窗口访问路径的定位问题具有收敛性，但未证明该收敛结果为最优解。

参考文献：

[1] Malamas E N,Petrakis E G M,Zervakis M.A survey on industrial vision systems,applications and tools[J].Image and Vision Computing,2003,21(2):171-188.

[2] 王勇,蔡自兴,周育人,等.约束优化进化算法[J].软件学报,2009,20(1):11-29.

[3] Arkin E M,Hassin R.Approximation algorithms for the geometric covering salesman problem[J].Discrete Applied Mathematics,1994,55:197-218.

[4] Hochbaum D S,Maass W.Approximation schemes for covering and packing problems in robotics and VLSI[J].Lecture Notes in Computer Science,1984,166:55-62.

[5] Iwano K,Raghavan P,Tamaki H.The traveling cameraman problem,with applications to automatic optical inspection[J].Lecture Notes in Computer Science,1994,834:29-37.

[6] Park T H,Kim H J,Kim N.Path planning of automated optical inspection machines for PCB assembly systems [J].International Journal of Control,Automation,and Systems,2006,4(1):99-104.

[7] Park T H,Kim H J.Path planning of automatic optical inspection machines for PCB assembly systems[C]//Proceedings 2005 of IEEE International Symposium on Computational Intelligence in Robotics and Automation.Espoo:IEEE Press,2005:249-254.

[8] 罗兵,章云.基于蚁群算法的 SMT 自动光学检测路径规划[J].仪器仪表学报,2006(S2):94-97.

(上接 5 页)

[2] Breiman L.Classification and regression trees [M].Monterey,CA:Wadsworth and Brooks,1984.

[3] Cristianini N,Shawe-Taylor J.An introduction to support vector machines and other kernel-based learning methods[M].[S.l.]:Cambridge Univ Pr,2000.

[4] Guyon I,Weston J,Barnhill S,et al.Gene selection for cancer classification using support vector machines[J].Machine Learning,2002,46(1):389-422.

[5] Duan K B,Rajapakse J C,Wang H,et al.Multiple SVM-RFE for gene selection in cancer classification with expression data[J].IEEE Transactions on Nanobioscience,2005,4(3):228-234.

[6] Zhou X,Tuck D P.MSVM-RFE:Extensions of SVM-RFE for multi-class gene selection on DNA microarray data[J].Bioinformatics,2007,23(9):1106.

[7] Chen X,Zeng X,van Alphen D.Multi-class feature selection for texture classification[J].Pattern Recognition Letters,2006,27(14):1685-1691.

[8] Sawaragi Y,Nakayama H,Tanino T.Theory of multi-objective optimization[M].New York:Academic Pr,1985.

[9] Wahba G.Support vector machines,reproducting kernel Hilbert spaces and the randomized GACV[M]//Advances in Kernel Methods-Support Vector Learing.Massachusetts:MIT Press,1999:69-88.

[10] Platt J.Probabilistic outputs for support vector machines and comparison to regularized likelihood methods[M]//Smola A J,Bartlett P L,Schölkopf B,et al.Advances in Large Margin Classifiers.Cambridge,MA:MIT Press,2000.

[11] Statnikov A,Aliferis C F,Tsamardinos I,et al.A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis[J].Bioinformatics,2005,21(5):631-643.