一种基于互信息最大化的模型无关基因选择方法

魏莎莎[1] 陆慧娟[1] 安春霖[1] 郑恩辉[2] 金 伟[1]
(中国计量学院信息工程学院 杭州 310018)[1] (中国计量学院机电工程学院 杭州 310018)[2]

摘 要 针对大规模基因芯片高维度的基因表达数据存在大量无关和冗余特征可能降低分类器性能的问题,提出了一种基于互信息最大化方法(MMI)和与遗传算法的模型无关的基因选择方法来将特征选择转化为全局优化问题,其中的适应度函数定义为类间距离与类内距离之比,适应程度高。为了评价算法的性能,采用3个数据集进行了实验,结果表明MMIGA-Selection取得了较好的效果,在每个数据集上获得了较高的5折交叉验证正确率。MMIGA-Selection主要有两个优点:一是可以有效减少冗余基因;二是模型无关性,选择得出的特征子集可直接用于其他类型的分类器,分类精度较高。

关键词 互信息最大化,模型无关,遗传算法,基因选择
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.046

Model-free Gene Selection Method Based on Maximum Mutual Information

WEI Sha-sha[1] LU Hui-juan[1] AN Chun-Lin[1] ZHENG En-hui[2] JIN Wei[1]
(Department of Information Engineering, China Jiliang University, Hangzhou 310018, China)[1]
(Department of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China)[2]

Abstract The large number of irrelevant and redundant features in high dimensionality of large-scale gene chip expression data may reduce the performance of the classifiers. We proposed a model-free gene selection method based on the maximum mutual information (MMI) to transform feature selection into a global optimization problem. The fitness function was defined as the distance between the class and class in the ratio of the distance. In order to evaluate the performance of the algorithm, experiments were done in three data sets. Experimental results show that MMIGA-Selection obtains a better effect in every data set of the 5 fold cross validation accuracy. MMIGA-Selection has two main advantages. First, it can effectively reduce the redundant genes. Second, the model-free algorithm makes the feature subset directly apply to other types of classifier and obtains higher classification accuracy.

Keywords Maximum mutual information, Model-free, Genetic algorithm, Gene selection

1 引言

随着生物芯片的不断发展,大规模基因芯片(DNA微阵列)技术的应用成为功能基因组以及肿瘤诊断等研究中的重要监测手段[1],由于有些基因只在特定实验条件下表达,并且基因表达数据又具有维度高、样本小的特点,故需要对其进行特征选择,选取与分类紧密关联的基因,降低后期的生物学分析成本,降低机器学习的时间及空间复杂度,同时提高分类的正确率[2]。特征选择根据各个特征的重要程度,剔除一组特征中不相关的冗余特征后,挑选出对分类有意义的某些特征以降低特征空间维数。在模式识别、数据挖掘以及机器学习中,特征选择都非常关键[3]。

根据是否依赖机器学习算法,特征选择算法可以分为两大类,一类为Wrapper型算法,另一类为Filter型算法。Wrapper法与具体分类器结合,将分类器预测正确率作为评价基因组合好坏的标准,这种方法降维效果好,但计算代价大,效率低。Filter法如t-test[4]、信噪比[5]等泛化能力强,简单快速,但以单个基因蕴含的分类信息多少为标准,没有考虑基因之间的相互联系,其中分类信息高的并不一定是最优组合。

互信息通常用于描述两个随机变量间的统计相关性,用一个变量中包含另一个变量的信息多少表示两个随机变量之间的依赖程度,是信息论中的一个测度,一般用熵来表示[6]。同一分类系统的基因在统计学上并非独立而是相关的,这是用互信息进行筛选的基础。考虑在不同时间或不同条件下获取的每一个基因,确定基因之间的互信息就是要定义相似性

到稿日期:2013-11-20 返修日期:2014-02-26 本文受国家自然科学基金(61272315,60842009,60905034),浙江省自然科学基金(Y1110342,Y1080950),浙江省科技厅国际合作项目(2012C24030)资助。

魏莎莎(1989-),女,硕士生,CCF会员,主要研究方向为机器学习、数据挖掘,E-mail:weishasha5210@163.com;**陆慧娟**(1962-),女,博士,教授,CCF常务理事,主要研究方向为机器学习、模式识别和生物信息学等,E-mail:hjlu@cjlu.edu.cn(通信作者);**安春霖**(1988-),女,硕士生,CCF会员,主要研究方向为机器学习、数据挖掘;**郑恩辉**(1975-),男,博士,副教授,主要研究方向为模式识别和数据挖掘;**金 伟**(1989-),男,硕士生,主要研究方向为机器学习、模式识别。

测度来寻找变换关系，使得这两个基因间的相似性达到最大，从而确定该基因在分类系统中的重要程度。当信息量达到最佳配准时，互信息为最大，即互信息最大化。近来，利用互信息最大化法进行多模医学图像配准成为了医学图像处理领域的热点[7]，它是一种基于相关性的过滤特征选择方法，其特点是速度相对比较快，能够很快地排除很大数量的非关键性的噪声和无关基因，但得到的特征子集可能不是最优的，分类精度较低。

遗传算法(Genetic Algorithm,GA)[8]是搜索最优解的计算模型和方法，具有内在的隐并行性和很好的全局寻优能力，它将多个约束条件组合在一起，能自动获取和指导优化的搜索空间，自适应地调整搜索方向。目前有很多依赖于具体分类器(Wrapper 法)的 GA 算法在特征选择中得到应用[9]。在此基础上，李等[10]给出一个全新的考虑样本不平衡性和基因选择方法的稳定性的模型无关的基因选择方法，陆等[11]给出一个模型无关的遗传算法，但是将互信息最大化和遗传算法相结合提出模型无关基因选择算法还需进一步研究。

目前，裘等[3]提出了基于互信息和遗传算法的两阶段特征选择方法，它将归一化互信息与遗传算法结合，得到了较好的特征选择结果。在此基础上，由于互信息最大化在医学图像配准中不进行图像分割特征提取等预处理过程便能实现不同模态图像的配准，操作简单，应用广泛，因此本文提出了一种新的 Filter 型的基因选择算法：MMIGA-Selection。

互信息最大化与模型无关遗传算法组合，首先在数量较大的数据集中用互信息最大化法进行基因分组以及筛选，对特征进行排序，然后选择出排序在前的特征对遗传算法的部分种群进行初始化，使得遗传算法的初始种群中有较好的搜索起点，从而结合起来的算法只需较少的进化代数就可搜寻到较优的特征子集，将特征选择转化成组合优化问题[12]。MMIGA-Selection 将类间距离与类内距离之比作为种群适应度函数，不需要与具体分类器结合，具有模型无关性。MMIGA-Selection 结合了互信息最大化，拥有运算速度快、遗传算法分类精度高以及模型无关性的优点。

2 基于互信息最大化的基因分组与筛选

基因筛选是基因选择的一个前期过程，在基因选择和基因降维时先进行基因筛选能够提高计算效率，是选出具有代表性的精简基因子集的有效方法[13]。

熵表示任何一种能量在空间中分布的均匀程度，是信息论中一个非常重要的概念。其中，熵的大小跟能量分布均匀程度有关，能量越不确定且越均匀，熵就越大[14]。“信息熵”是 Shannon[15]将熵应用于信息处理所提出的概念。在互信息最大化中，信息熵主要用来衡量一个随机变量取值的重要性程度，衡量标准是看特征能够为分类带来多少信息，带来的信息越多，该特征越重要。

假设 X 来自于一个集合 S，且 X 是一个离散随机变量。X 的概率密度分布函数表示为 $p(x)$，则 X 的信息熵定义如下：

$$H(X)=-\sum_{x\in s}P(x)\log_2(P(x)) \tag{1}$$

已知一个来自于集合 T 的变量 Y，$p(x|y)$ 表示 Y 取值为 y 时 X 取值为 x 的概率，X 的不确定性用 $H(X|Y)$ 来衡量，如式(2)所示：

$$H(X|Y)=-\sum_{y\in T}p(y)\sum_{x\in s}p(x|y)\log_2(p(x|y)) \tag{2}$$

$p(x,y)$ 用来表示 X,Y 的联合概率密度，则它们的互信息量 $I(X;Y)$ 定义为：

$$\begin{aligned}I(X;Y)&=H(X)-H(X|Y)\\&=\sum_{x\in s}\sum_{y\in T}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}\end{aligned} \tag{3}$$

在基因筛选中互信息通常用于计算特征与特征之间的关系。在以上公式中，假设考虑特征 t 与类别 c 的分布，N 为基因总数，A 为类 c 中出现特征 t 的基因数，B 为非类 c 中出现特征 t 的基因数，C 为类 c 中不出现特征 t 的基因数，特征 t 与类 c 之间的互信息定义为：

$$\begin{aligned}I(t,c)&=\log\frac{p(t|c)}{p(t)}=\log\frac{p(t,c)}{p(t)\times p(c)}\\&\approx\log\frac{A\times N}{(A+C)\times(A+B)}\end{aligned} \tag{4}$$

如果 $I(t,c)=0$，那么特征 t 与类 c 相互独立。

在式(5)中提供关于类别信息的加权平均值来衡量一个特征在全局特征选择中的重要性：

$$I_{avg}(t)=\sum_{i=1}^{k}P(c_i)I(t,c) \tag{5}$$

特征选择后，尽可能多地保留关于类别的信息即达到互信息最大化：

$$MaxMI(t)=\sum_{i=1}^{k}P(c_i|t)\log\frac{P(c_i|t)}{P(c_i)} \tag{6}$$

则最大互信息量：

$$I^*=MaxMI(t) \tag{7}$$

用互信息最大化来进行特征选择，选择后的特征集合应该尽可能多地提供关于类别的信息。两个随机变量之间共有的信息量越大，说明两个变量之间的相关程度越高。如果两个变量完全不相关，则互信息量为 0。互信息最大化是选择相关程度最高的两个变量，进行循环迭代，得出每次信息量中相关程度最高的变量。

3 基于模型无关遗传算法的基因选择

通过互信息最大化的筛选，为后面遗传算法排除了较大数量的噪声，提高了效率。本文提出的基因选择方法具有模型无关性，将遗传算法中的种群适应度函数定义为类间距离与类内距离之比。

3.1 模型无关基因选择方法

现假设有来自 L 个不相交的类别的 n 个样本，在样本 i ($1\leqslant i\leqslant n$)上，可以用 $A_i=(a_{i1},a_{i2},\cdots,a_{ip})$ 来表示 p 个基因的表达值，称为样本 i 的基因表达谱。表达谱与样本一一对应。a_{ij} 为样本 i 上基因 j 的表达值，则基因表达矩阵为：$A=[A_1^T,A_2^T,\cdots,A_n^T]^T$。

不同样本的相同类别中的表达值间的距离称为类内距离，一般记为 compact；不同样本的不同类别中的表达值间的距离称为类间距离，一般记为 scatter。一个相对理想的基因一定是类间的距离比较大，类内的距离比较小，即 compact/scatter 越小，那么这个基因就越理想[10]。

给定基因表达矩阵 A，则表达谱的类质心矩阵为：

$$\overline{A}=[\overline{a}_{kj}]_{L\times P} \tag{8}$$

矩阵 $\overline{A}$ 中第 k 个行向量表示类别 k 中所有样本表达谱的

质心，其中：

$$\bar{a}_{kj}=\frac{1}{n_k}\sum_{i\in c_k}a_{ij} \tag{9}$$

$n_k=|C_k|$，C_k 为属于第 k 个类别的样本标号的集合。令样本 i 属于第 k 个类别，由基因表达矩阵 A 和类质心矩阵 $\bar{A}$ 得样本 i 的类内变化谱及类内变化矩阵为：

$$X_i=(x_{i1},x_{i2},\cdots,x_{ip}),X=[x_{ij}]_{n\times p} \tag{10}$$

其中，$x_{ij}=|a_{ij}-\bar{a}_{kj}|$，即 $X_i=|A_i-\bar{A}_k|$。

综上可得类内变化谱的质心矩阵为：

$$\bar{X}=[\bar{x}_{kj}]_{L\times P} \tag{11}$$

$$\bar{x}_{kj}=\frac{1}{n_k}\sum_{i\in c_k}x_{ij} \tag{12}$$

假设基因 j 是一个类间差别相对较大的理想基因，即 $\bar{a}_{kj}$ $(1\leqslant k\leqslant L)$ 间距离较大，则对于任意的基因 j，可以用下面定义的 $S(j)$ 直观地表示基因 j 上的样本的类间差别：

$$S(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}(\bar{a}_{kj}-\hat{a}_j)^2} \tag{13}$$

其中，$\hat{a}_j=\frac{1}{L}\sum_{k=1}^{L}\bar{a}_{kj}$。

3.2 遗传算法描述

遗传算法[16,17]是一种随机搜索策略，由美国 Michigan 大学 Holland[18]教授于 1975 年首次提出，是模拟达尔文生物进化论的遗传选择和自然淘汰的生物进化过程，通过选择遗传和变异等进化操作，实现各个个体适应性的提高，从而解决各种复杂问题。1967 年 Bagley[19]首次提出了自我调整优化的概念，让遗传算法这一术语得到传播。模式理论奠定了遗传算法的基础，由 Holland[20]在 1975 年研究得到。20 世纪 80 年代以后，遗传算法发展迅速，广泛地应用于组合优化机器学习等领域，能够快速有效地选择出理想特征组。

遗传算法首先需要编码，其中二进制编码是最常用的编码方式。每一个染色体需要编码长度为 K 的“01001…10010”串，其中“1”表示选择对应的特征，“0”表示不选择。

本文中的 MMIGA-selection 算法将适应度函数定义为 3.1 节中介绍的 compact/scatter 的值，则令样本的类内平均距离为：

$$d_0=\bar{x}_{kj} \tag{14}$$

样本的类间平均距离为：

$$d_1=S(j) \tag{15}$$

理想的基因应该具备类内距离尽可能小、类间距离尽可能大的特点。因此，适应度函数可定义为：

$$F=e^{d_1/(d_0+\Delta)} \tag{16}$$

这里，Δ 是一个足够小的正实数，在实验中设置为 10^{-10}。

在遗传算法中，由于存在着收敛速度慢的问题，过大的交叉变异率就容易导致“早熟”现象，即过早地收敛在某一个局部极值点上，形成局部最优。因此，在这里动态地调整交叉变异率，交叉率公式如下：

$$p_c=\begin{cases}\sin\left(\frac{\pi}{2}\times\frac{f_{\max}-f'}{f_{\max}-f_{avg}}\right), & f'>f_{avg}\\ p_0, & f'\leqslant f_{avg}\end{cases} \tag{17}$$

式中，f 表示变异染色体的适应度，f' 是两个交叉染色体适应度的最大值。p_0 是 0 和 1 之间的常数，可设置为 $p_0=0.1$，同理可得出变异率 p_m，其中将 p_0 换成 p_1，$p_1=0.01$。

4 MMIGA-Selection

4.1 算法分析

综上所述，MMIGA-Selection 算法的目的是得到分类精度更高的特征子集，同时提高遗传算法的计算效率，让遗传算法具有模型无关性，不依赖任何分类器。算法首先结合互信息最大化方法对基因进行初步筛选，去除大量噪声，为遗传算法提供一个比较理想的种群初始化环境。

然后对得到的结果进行排序，选择出类相关性最大的特征，遗传算法运行时，初始种群除了互信息最大化得出部分个体外，还包括由遗传算子和种群适应度函数随机产生的个体。在遗传算法迭代时，解码出选择的特征位，得出的适应度最高的个体所对应的特征子集，该特征子集即为特征选择的结果，是本文中 MMIGA-Selection 算法得到的一个全局优化的结果，分类精度较高。

4.2 算法描述

根据以上的分析，MMIGA-Selection 具体算法流程可描述如下：

输入：原始基因集 D，设置迭代次数 T，初始化种群染色体编码 p_0 与 p_1。

输出：优化后的特征基因子集 TD。

1）利用互信息最大化方法，对输入基因集进行筛选，形成选择后的基因子集 FD；

2）根据编码方案，将种群中的个体编码成二进制串，计算 f，f_{max} 和 f_{avg}；

3）由随机竞争法得出初始种群，适应度最大的染色体直接复制到下一代；

4）根据交叉率和变异率公式将种群中染色体随机两两配对，进行均匀交叉及变异；

5）判断是否达到最大迭代次数或连续若干代 f_{max} 不变；如果是，则跳到步骤 7），否则执行 6）；

6）重复步骤 3）－5）；

7）输出最优特征子集。

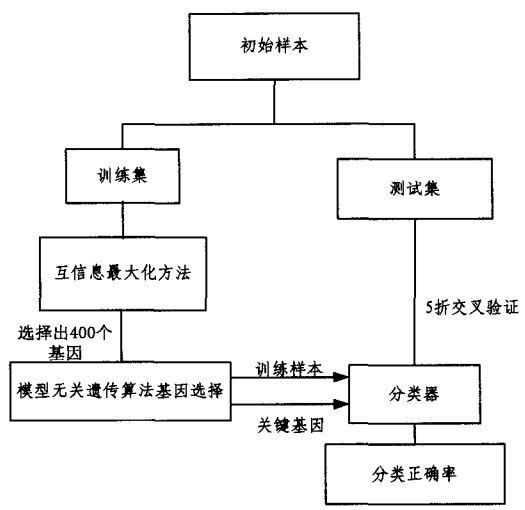

图 1 MMIGA-Selection 的流程图

5 实验与结果分析

为评估 MMIGA-Selection 算法的性能，本文对算法进行

了实验分析与仿真，分析的对象是常用的 Leukemia、Colon 与 SRBCT 这 3 个数据集。其中 Leukemia 数据集是 Golub 等人研究的两种急性白血病的基因表达谱，其中有 72 个急性白血病样本，每个样本含有 7130 个基因的表达数据；Colon 是 Alonp's 等人测定的有 62 个样本和 2000 个基因表达数据的数据集；SRBCT 数据集是 Khan 等人对儿童小圆蓝细胞瘤进行研究获得的 83 例样本中 2308 条基因的表达水平，包括 4 种类型。前两个数据集属于二分类问题，SRBCT 数据集属于多分类问题。数据集具体情况见表 1。

表 1　数据集

数据集	样本总数	基因数	类分布	
			类名	样本数
Leukemia	72	7130	ALL	24
			MLL	20
			AML	28
Colon	62	2000	Negative	40
			Positive	22
SRBCT	83	2308	Negtive	150
			Positive	120

实验利用 BPNN、ELM、SVM 以及 Bayes 作为分类器，以分类精度高低来评价特征子集。其中设 BPNN 隐层节点数为 100，ELM 隐层结点数为 300，激励函数为 sigmoid 函数，SVM 用 RBF 核函数，惩罚系数 0.125，gamma 系数 0.1255。

在分类之前，先将基因表达矩阵中的元素进行对数转换和标准化。

$$x_{ij}'=\frac{x_{ij}-\bar{x}_i}{\sqrt{\frac{1}{N-1}}\sqrt{\sum_{j=1}^{N}(x_{ij}-\bar{x}_i)^2}} \tag{18}$$

实验对二分类及多分类数据集进行特征选择，首先用互相息最大化方法选出了 400 个基因作为模型无关遗传算法的初始化环境。在 Leukemia 数据集上依次选择出了 207，17，50，12 个基因；在 Colon 数据集上选择出了 199，123，74，20 个基因；在 SRBCT 数据集上选择出了 250，108，62，18 个基因作为基因子集。

表 2 给出了 MMIGA-Selection 与 t-test/f-test 方法的统计检验。其中样本方差为：

$$S^2=\sum(X-\bar{X})^2/(n-1) \tag{19}$$

X 为检验量的样本，$\bar{X}$ 为平均值，n 为样本数。则可得 F 值为：

$$F=\frac{n_1S_1^2/(n_1-1)}{n_2S_2^2/(n_2-1)} \tag{20}$$

S_1^2、S_1^2、n_1、n_2 分别代表 F 检验中两个检验样本的样本方差和样本数。α 为 0.1。通过查表得 $F_{0.5}(14,14)=2.46$，则 $F\geqslant F_{\alpha/2}(n_1-1,n_2-1)$，因此两种方法之间存在显著的差异。从实验的结果可以得到，大部分情况下，MMIGA-Selection 优于 t-test/f-test。

表 2　两种基因选择方法的统计检验

数据集	选择方法	方差	F 值
Leukemia	t-test	0.000937	6.7825
	MMIGA-Selection	0.000138	
Colon	t-test	0.000661	2.8017
	MMIGA-Selection	0.000235	
SRBCT	f-test	0.002390	2.8712
	MMIGA-Selection	0.000080	

同时，t-test/f-test 属于 Filter 法，对 4 种分类器而言都满足 $F<F_{\alpha/2}(n_1-1,n_2-2)$，且由表 3 可以得到，MMIGA-Selection 3 个分类器的方差没有显著的差异，因此可以说 MMIGA-Selection 选择的关键基因是和模型无关的，泛化程度高。

表 3　不同分类器的方差检验

数据集 \ 分类器	BPNN	SVM	ELM	Bayes
Leukemia	0.0035471	0.0017979	0.005102	0.0035470
Colon	0.0032142	0.0047820	0.002533	0.0028413
SRBCT	0.0026488	0.0039063	0.004072	0.0026488

为了比较 t-test/f-test 法与 MMIGA-Selection 的性能，本文利用 BPNN、ELM、SVM 和 Bayes 4 个不同分类器在 3 个数据集上分别进行了 15 次 5 折交叉验证，最终分类精度取 15 次分类精度结果的平均值，实验结果如表 4－表 6 所列，对应的精度图见图 2－图 4。

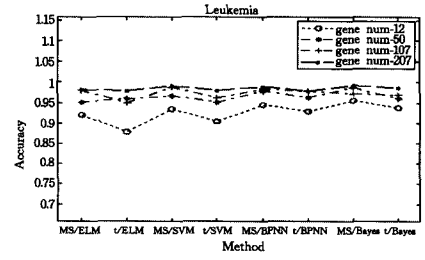

图 2　Leukemia 数据集上不同分类器分类精度结果

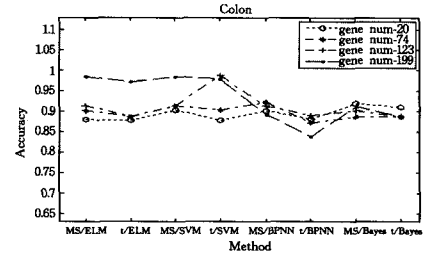

图 3　Colon 数据集上不同分类器分类精度结果

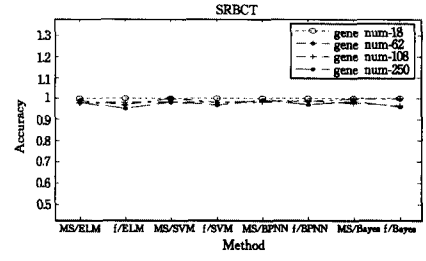

图 4　SRBCT 数据集上不同分类器分类精度结果

表 4　Leukemia 数据集上 5 折交叉验证精度

方法 \ 基因数	12	50	107	207
MMIGA-Selection/ELM	0.91958	0.95102	0.97998	0.98122
t-test/ELM	0.87720	0.95987	0.95001	0.97861
MMIGA-Selection/SVM	0.93429	0.96668	0.98812	0.99100
t-test/ SVM	0.90402	0.95123	0.96330	0.98001
MMIGA-Selection/BPNN	0.94467	0.97866	0.98324	0.98918
t-test/ BPNN	0.92887	0.96310	0.97812	0.97861
MMIGA-Selection/Bayes	0.95462	0.98774	0.97214	0.99110
t-test/Bayes	0.93712	0.96012	0.96889	0.98611

表 5　Colon 数据集上 5 折交叉验证精度

方法 \ 基因数	20	74	123	199
MMIGA-Selection/ELM	0.87872	0.90122	0.91223	0.9840
t-test/ELM	0.87800	0.88711	0.88711	0.9719
MMIGA-Selection/SVM	0.90321	0.91244	0.91222	0.9841
t-test/ SVM	0.87811	0.90244	0.9879	0.9793
MMIGA-Selection/ BPNN	0.90128	0.92221	0.91221	0.89221
t-test/ BPNN	0.88122	0.87162	0.88971	083871
MMIGA-Selection/ Bayes	0.91943	0.88702	0.90177	0.91117
t-test/ Bayes	0.90942	0.88712	0.88712	0.88710

表 6　SRBCT 数据集上 5 折交叉验证精度

方法 \ 基因数	18	62	108	250
MMIGA-Selection/ELM	1.0000	0.99122	0.97822	0.98110
f-test/ELM	1.0000	0.97012	0.98221	0.95330
MMIGA-Selection/SVM	1.0000	1.00000	0.98391	0.98357
f-test/ SVM	1.0000	0.98110	0.98101	0.97112
MMIGA-Selection/BPNN	1.0000	0.99001	0.98410	0.99212
f-test/ BPNN	1.0000	0.98704	0.99100	0.97222
MMIGA-Selection/Bayes	1.0000	0.99778	0.97912	0.98777
f-test/Bayes	1.0000	0.99811	0.96812	0.96222

由上述图和表可以看出，MMIGA-Selection 算法在不同分类器上都有较高的分类精度。由于较好的初始种群提供了较优的搜索起点，在选择过程中为遗传算法节省了时间，结合遗传算法本身特点，MMIGA-Selection 算法比 t-test/f-test 法相对更耗时，但是差别较小，在基因选择领域中可以接受。

结束语　本文提出了一种基于互信息最大化的模型无关遗传算法的特征选择方法。算法结合互信息最大化方法(Filter 法)对基因进行初步筛选，去除大量噪声以及不相关特征，为遗传算法提供一个更好的种群初始化环境。理想的搜索起点能够加快遗传算法的搜索速度。由于将类间距离与类内距离定义为遗传算法的适应度函数，使得算法与分类器无关，适应程度与运行效率相对较高。实验结果显示，MMIGA-Selection 在基因特征选择上取得了较好的结果，算法有两个主要优点：1)能够有效减少冗余基因；2)模型无关性，泛化程度高。

参 考 文 献

[1] Kang H N, Chen I M, Wilson C S. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia[J]. Blood, 2010, 115: 1394-1405

[2] 任江涛，黄焕宇，孙婧昊. 基于相关性分析及遗传算法的高维数据特征选择[J]. 计算机应用，2006，26(6)：1403-1405

[3] 裘国永，王娜，汪万紫. 基于互信息和遗传算法的两阶段特征选择方法[J]. 计算机应用研究，2012，29(8)：2903-2905

[4] Liu Hui-qing, Li Jin-yan. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns[J]. Genome Informatics, 2002, 13: 51-60

[5] Zhao Zheng, Wang Lei, Liu Huan. Efficient spectral feature selection with minimum redundancy[J]. Proceedings of the National Conference. 2010, 1: 673-678

[6] 王凌，陈震，危水根，等. 基于改进最大互信息法的 MR 切片图像配准[J]. 生物医学工程学杂志，2012，29(2)：201-205

[7] 杨虎，马斌荣，任海萍，等. 基于最大互信息的人脑 MR-PET 图像配准方法[J]. 北京生物医学工程，2001，20(4)：246-251

[8] Michalewicz Z. A Modified Genetic Algorithm for Optimal Control Problems[J]. Computers Math, 1992, 23(12): 83-94

[9] Huang Jin-jie, Cai Yun-ze, Xu Xiao-ming. A Hybrid Genetic Algorithm for Feature Selection Wrapper Based on Mutual Information[J]. Pattern Recognition Letters, 2007, 28(13): 1825-1844

[10] 李建中，杨昆，高宏，等. 考虑样本不平衡的模型无关的基因选择方法[J]. 软件学报，2006，17(7)：1485-1493

[11] Lu Hui-juan, Chen Wu-tao, Ma Xiao-ping, et al. Model-free Gene Selection Using Genetic Algorithms [J]. International Journal of Digital Content Technology and its Applications, 2011, 5(1): 195-203

[12] 陆慧娟. 基于基因表达数据的肿瘤分类算法研究[D]. 徐州：中国矿业大学，2012

[13] 王明怡. 微阵列数据挖掘技术的研究[D]. 杭州：浙江大学，2004

[14] 刘庆和，梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用，2011，47(12)：130-132

[15] Hu Y, Loizou P C. Speech enhancement based on wavelet thresholding the multitaper Spectrum [J]. IEEE Trans on Speech and Audio Processing, 2004, 12(1): 59-67

[16] Wang Zhi-teng, Zhang Hong-jun, Hang Ying. Fire distribution optimization based on quantum immune genetic algorithm[C]// 2011 International Conference of Information Technology Computer Engineering and Management Sciences. IEEE, 2011, 1: 95-98

[17] Jiang Feng-guo. The truss structural optimization design based on improved hybrid genetic algorithm[J]. Advanced Materials Research, 2011, 163-167: 2304-2308

[18] Holland J H. The psychology of vocational choice: A theory of personality types and model environments[M]. Oxford, England: Oxford University Press, 1965

[19] Bagley J D. The Behavior of Adaptive Systems which Employ Genetic and Correlation Algorithms[D]. The univerty of Michigan, 1967

[20] Holland P W. Discrete Multivariate Analysis: Theory and Practice[M]. MIT Press, 1975

(上接第 224 页)

[14] 苗东菁，石胜飞，李建中. 一种局部相关不确定数据库快照集合上的概率频繁最近邻算法[J]. 计算机研究与发展，2011，48(10)：1812-1822

[15] 袁培森，沙朝锋，王晓玲，等. 一种基于学习的高维数据 c-近似最近邻查询算法[J]. 软件学报，2012，23(8)：2018-2031

[16] Zhang Li-ping, Li Song, Li Lin, et al. Simple Continues Near Neighbor Chain Query of the Datasets Based on the R Tree[J]. Journal of Computaional Information Systems, 2012, 8(22): 9159-916

[17] 张丽平，李林，李松，等. 预定数据链规模的单纯型连续近邻链查询[J]. 计算机工程，2012，38(10)：51-53

[18] 张丽平，李松，郝忠孝. 球面上的最近邻查询方法研究[J]. 计算机工程与应用，2011，47(5)：126-129

[19] 张丽平，李松，郝晓红，等. J_n 粗糙 Vague 区域关系[J]. 浙江大学学报，2012，46(1)：122-128