文章编号:1004-9037(2014)01-0083-07

基于聚类和微粒群优化的基因选择方法

刘金勇[1] 郑恩辉[1] 陆慧娟[2]

(1.中国计量学院机电工程学院,杭州,310018; 2.中国计量学院信息科学学院,杭州,310018)

摘要:在高维的基因表达谱数据中,只有少量基因对分类诊断其作用,而且还存在大量冗余的与癌症分类诊断无关的噪声基因,这些都会导致分类性能的下降。通过基因选择选取与分类紧密关联的基因,不仅能够剔除与疾病无关的基因,减少机器学习算法的时间复杂度和空间复杂度,提高分类的正确率,而且选出的特征基因可以作为肿瘤基因诊断和肿瘤药物治疗靶标确定的依据,降低后期生物学分析成本。本文提出一种基于聚类和粒子群算法(Particle swarm optimization,PSO)的基因选择方法,在 PSO 算法进行搜索之前,先对基因进行聚类,并对聚类结果进行选择,将被选中的簇的中心作为 PSO 的初始值,每个被选中的簇作为一个搜索空间,并利用极限学习机(Extreme learning machine,ELM)的分类精度作为特征选择的适应评价标准。该算法不仅有效地利用了聚类算法对基因进行初步归并的能力,也利用了 PSO 算法的全局优化能力,克服了传统 PSO 算法早熟、局部收敛速度慢的缺点,因此它能够高效地完成最优基因子集的确定,同时提高癌症分类正确率。

关键词:基因表达谱数据;基因选择;微粒群优化;极限学习机

中图分类号:TP391.4 **文献标识码**:A

Gene Selection Based on Clustering Method and Particle Swarm Optimization

Liu Jinyong[1], *Zheng Enhui*[1], *Lu Huijuan*[2]

(1. College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou, 310018, China;
2. College of Information Engineering, China Jiliang University, Hangzhou, 310018, China)

Abstract: Gene expression data has a high application value for understanding the pathogenesis, disease diagnosis and gene-level drug development. However, the microarray data usually contains thousands of genes with a small number of samples, which causes serious curse of dimensionality and deteriorates the diagnosis accuracy. Moreover, it gives raise to difficulty to a lot of classifiers, and cuts down the cost of medical diagnosis. A new gene selection method is proposed, which is based on clustering and particle swarm optimization (PSO). Firstly, partition the genes using clustering algorithm and the useful are clusters selected for classification. Then the wrapper selection method based on particle swarm optimization(PSO) and extreme learning machine(ELM) is used to select the compact gene subset with high classification accuracy from the genes selected before. This method take advantages of clustering and PSO algorithm, and it can perform better in classification than other classical methods.

Key words: gene expression data; gene selection; particle swarm optimization; extreme learning machine

引 言

基因表达谱数据又叫做微阵列数据,它是利用基因芯片技术测得的高通量基因在不同生理阶段的表达数据值。其中“基因表达”是指细胞在声明过程中,把储存在 DNA 中的遗传信息经过转录和翻译,转变成具有生物活性的蛋白质分子。“基因

基金项目:国家自然基金(60905034,61272315)资助项目;浙江省自然基金(Y1080950,1110342)资助项目。

收稿日期:2013-09-30;**修订日期**:2013-12-25

表达水平”是指某个基因在一定时间内控制产生的蛋白质的量，它表明了细胞当前的生理状态[1,2]。通过合理的方法对这种高通量的基因表达数据进行分析，可以得到哪些基因之间存在调控关系、不同样本之间哪些基因的表达水平发生了变化、不同的生理阶段如何影响基因的活动。

基因选择是采用某种优化算法从基因表达谱数据的所有属性中选择出一个最具有疾病识别能力的基因子集的过程[3,4]。选择出的基因子集在肿瘤识别过程中发挥着至关重要的作用。基于基因信息排序的过滤法[5]和依赖具体分类器选取基因的缠绕法[6,7]是两种主要的基因选择方法。基于排序的过滤法如、信噪比[8]、信息增益[9,10]等具有简单快速的特点，但它们都是按照单个基因蕴含的分类信息多少为标准的，没有考虑基因之间的相互联系，而含有分类信息高的基因组合并不一定是最优的组合[11]。缠绕法与具体分类器（如支持向量机（Support vector machine，SVM），ELM 等）结合，将分类器预测正确率作为评价基因组合好坏的标准，这种方法可以找出最优的基因组合，同时最小化基因子集，但算法每次评价一个基因组合都要进行分类器训练，时间复杂度较高，而且选择出的基因子集在其他类型的分类器中的泛化能力不高。

粒子群优化算法是一种新兴的基于群体智能的启发式全局搜索算法，通过粒子间的竞争和协作以实现在复杂搜索空间中寻找全局最优点。本文中使用粒子群算法来进行特征选择，但是由于大多数 PSO 算法在应用的过程中，其初始化都是随机的，不能保证初始群体粒子的合理分布，在 PSO 搜索过程中，就容易出现大部分粒子均被相同的局部极值所限制时，导致当前的粒子群失去多样性，陷入局部最好解，出现“早熟现象”，最终影响最优解的搜索。为了解决这一缺陷，在 PSO 算法进行搜索之前，先对基因进行聚类，并对聚类结果进行初步选择，将被选中的簇的中心作为 PSO 的初始值，每个被选中的簇作为一个搜索空间，并利用 ELM 的分类精度作为特征选择的适应评价标准。

这种将基因聚类提取基因先验信息并耦合进 PSO 算法进行特征选择的思想，由于其初始化都是固定的，且符合数据本身特点，在很大程度上代表了本来的数据，这样就保证了初始群体粒子的合理分布，在 PSO 搜索过程中，不容易出现大部分粒子均被相同的局部极值所限制的情况，避免出现局部最好解，最终得到的解便是最优解，且限制搜索范围能够减少搜索的时间，减少时间复杂度。

1 相关知识点

1.1 熵与信息增益

令 X 为随机变量，X 的不同取值 x_i，$i=1,2,\cdots$ 对应着不同的概率 $P(x_i)$，$i=1,2,\cdots$，那么 X 的信息熵定义为

$$H(X)=-\sum_i P(x_i)\log_2(P(x_i)) \tag{1}$$

对于分类系统来说，类别 $C(c_1,c_2,\cdots c_l)$ 是变量，因此分类系统的熵就可以定义为

$$H(C)=-\sum_{i=1}^{l} P(c_i)\log_2(P(c_i)) \tag{2}$$

式(2)表示特征的变化越大，它所代表的信息也就越多。特别地，对于两类分类问题，信息熵可以表示为

$$H(C)=-P(c_1)\log_2(P(c_1))-P(c_2)\log_2(P(c_2)) \tag{3}$$

基因表达谱数据的信息增益是对每一个基因而言的，对应特定的含有 n 种情况的基因 X，它所对应的条件熵为

$$H(C\mid X)=-\sum_{j=1}^{n} P(x_j)\sum_{i=1}^{l} P(c_i\mid x_j)\log_2(P(c_i\mid x_j)) \tag{4}$$

其中 $P(c_i\mid x_j)$ 代表基因 x_j 属于类别 c_i 的条件概率。该基因 X 为整个分类系统所带来的信息增益，可以用原系统的信息熵与基因 X 固定之后的条件熵之间的差值，用式(5)表示。

$$IG(X)=H(C)-H(C\mid X) \tag{5}$$

$IG(X)$ 便代表了基因表达谱数据中每个基因的信息增益，信息增益值越大，则该基因代表的分类信息就越多。

1.2 微粒群优化算法

PSO 从这种模型中得到启示并用于解决优化问题。PSO 中，每个优化问题的潜在解都是搜索空间中的一只鸟，称之为粒子。所有的粒子都有一个由被优化的函数决定的适值，每个粒子还有一个速度决定它们飞翔的方向和距离，然后粒子们就追随当前的最优粒子在解空间中搜索。

PSO 初始化为一群随机粒子（随机解），然后通过迭代找到最优解。在每一次迭代中，粒子通过跟踪两个极值来更新自己；第一个就是粒子本身所找到的最优解，这个解称为个体极值；另一个极值是整个种群目前找到的最优解，这个极值是全局极值。另外也可以不用整个种群而只是用其中一部分作为粒子的邻居，那么在所有邻居中的极值就是

局部极值。

假设在一个 D 维的目标搜索空间中，有 N 个粒子组成一个群落，其中第 i 个粒子表示为一个 D 维的向量

$$\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iD}),\ i=1,2,\cdots,N \tag{6}$$

第 i 个粒子的“飞行”速度也是一个 D 维的向量

$$\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{iD}) \tag{7}$$

第 i 个粒子迄今为止搜索到的最优位置称为个体极值，记为

$$\boldsymbol{p}_{\text{best}}=(p_{i1},p_{i2},\cdots,p_{iD}) \tag{8}$$

整个微粒群迄今止搜索到的最优位置为全局极值，记为

$$\boldsymbol{g}_{\text{best}}=(p_{g1},p_{g2},\cdots,p_{gD}) \tag{9}$$

在找到这两个最优值时，粒子根据式(10,11)来更新自己的速度和位置

$$v_{id}=w*v_{id}+c_1r_1(p_{id}-x_{id})+c_2r_2(p_{gd}-x_{id}) \tag{10}$$

$$x_{id}=x_{id}+v_{id} \tag{11}$$

式中：c_1 和 c_2 为学习因子，也称加速常数，r_1 和 r_2 为[0,1]范围内的均匀随机数。式(6)右边由 3 部分组成，第 1 部分为“惯性”部分，反映了粒子的运动“习惯”，代表粒子有维持自己先前速度的趋势；第 2 部分为“认知”部分，反映了粒子对自身历史经验的记忆，代表粒子有向自身历史最佳位置逼近的趋势；第 3 部分为“社会”部分，反映了粒子间协同合作与知识共享的群体历史经验，代表粒子有向群体或邻域历史最佳位置逼近的趋势，根据经验，通常 $c_1=c_2=2$。v_{id} 是粒子的速度，$v_{id}\in[-v_{\max},v_{\max}]$，$v_{\max}$ 是常数，由用户设定用来限制粒子的速度。r_1 和 r_2 是介于[0,1]之间的随机数。

1.3　极限学习机

极限学习机(Extreme learning machine，ELM)是 Huang 等[12-14]提出来的一种单隐层前馈神经网络。ELM 的隐藏层参数(输入层权重和输入层偏置)都是随机产生，并且一旦产生就固定下来，输出层权重是通过最小二乘解计算出来的。给定样本 $\{(x_i,y_i)\}_{i=1}^{N}\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{id}]^{\mathrm{T}}\in\boldsymbol{R}^d$，$\boldsymbol{y}_i=[\boldsymbol{y}_{i1},\boldsymbol{y}_{i2},\cdots,\boldsymbol{y}_{im}]^{\mathrm{T}}$，$i=1,2,\cdots,N$ 设定隐层节点数为 L，且激活函数为 $g(x)$ 的标准单隐层前馈神经网络可以用数学模型表示出来

$$o_j=\sum_{i=1}^{L}\beta_ig_i(x_j)=\sum_{i=1}^{L}\beta_iG(a_i,b_i,x_j) \tag{12}$$

式中：β_i 为连接第 i 个隐层节点与输出神经元的输出权值，a_i 为连接输入神经元与第 i 个隐层节点的输入权值，b_i 为第 i 个隐层节点的偏置，o_j 为第 j 个输入样本的输出值，$j=1,\cdots,N$。

如果含有 L 个隐层节点，且激活函数为 $g(x)$ 的单隐层前馈神经网络可以零误差逼近于 N 个训练样本，即存在 β_i，a_i 和 b_i，使得

$$y_j=\sum_{i=1}^{L}\beta_ig(x_i,a_i,b_i),\quad j=1,2,\cdots,N \tag{13}$$

成立，并且

$$\sum_{j=1}^{L}o_j-y_j=0 \tag{14}$$

其中式(13)可以表示为

$$\boldsymbol{Y}=\boldsymbol{H}\boldsymbol{\beta} \tag{15}$$

其中 $\boldsymbol{H}$ 称为 ELM 的隐层输出矩阵

$$\boldsymbol{H}=\begin{bmatrix}h(x_1)\\ \vdots\\ h(x_N)\end{bmatrix}=\begin{bmatrix}G(a_1,b_1,x_1) & \cdots & G(a_L,b_L,x_1)\\ \vdots & \vdots & \vdots\\ G(a_1,b_1,x_N) & \cdots & G(a_L,b_L,x_N)\end{bmatrix}_{N\times L} \tag{16}$$

$$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}}\\ \cdots\\ \boldsymbol{\beta}_L^{\mathrm{T}}\end{bmatrix}_{L\times m}\qquad \boldsymbol{Y}=\begin{bmatrix}\boldsymbol{y}_1^{\mathrm{T}}\\ \vdots\\ \boldsymbol{y}_N^{\mathrm{T}}\end{bmatrix}_{N\times m}$$

如果隐层节点个数 L 和样本数量 N 是相等的，那么可以很容易的知道式(13,14)是成立的，但是当 $L<N$ 时，单隐层前馈神经网络并不能零逼近于 N 个训练样本。这时式(15)可以表示为

$$\boldsymbol{Y}=\boldsymbol{H}\boldsymbol{\beta}+\boldsymbol{E} \tag{17}$$

$\boldsymbol{E}=[e_1,e_2,\cdots,e_N]^{\mathrm{T}}$ 被称为是训练误差。这里训练一个 ELM 也即计算训练误差 $\boldsymbol{E}$ 的最小范数

$$\min\boldsymbol{E}=\min\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{Y} \tag{18}$$

因此，通过式(18)并利用最小二乘的方法计算得到输出权重

$$\bar{\beta}=\arg\min_{\beta}\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{Y}=\boldsymbol{H}^{*}\boldsymbol{Y} \tag{19}$$

其中 $\boldsymbol{H}^{*}$ 为矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆。

2　基于聚类与 PSO 的基因选择

在基因组学中，聚类算法是研究基因间相互关系的最基本手段。聚类算法能够将那些具有相似功能特点的基因聚在一起，根据聚类的结果，可以预测未知基因的功能，寻找基因之间的调控关系以及发现共同的模式。其中比较流行的启发式方法是 K-means 方法。在对基因进行聚类时聚类数目的选择有两种方法，一种是随机选取，但是这种选

取方法没有任何的针对性,需要迭代的次数较多,计算量也比较大;另外一种是根据某种准则选取,对基因表达谱数据的基因进行聚类时,需要结合数据本身的特点,包括数据的类别信息和冗余度信息等,如针对包含两种类别的样本进行聚类的时候,基因可以分为与样本类别相关的 2 簇以及 1 簇对样本分类无关的冗余基因,所以可以确定聚类数目为 3。

将基因表达谱数据分成训练集和测试集,通过信息增益的方法选择前 n 个信息熵最大的基因。接下来便是利用聚类算法对初选的基因子集进行聚类,聚类方法采用 K-means 方法,聚类数目按照上述方法给出,如对于两类样本,聚类数目选择为 3,k 类样本则对应聚类数目为 $k+1$。借助分类器对各簇的基因分类性能进行分析,将具有高分类性能的簇选择出来,排除对分类影响较小的簇,这些被选择的簇所包含的基因构成一个冗余度较低的特征基因子集。

最后将被选中的簇的中心作为初始位置,每一簇作为一个搜索空间,利用 PSO 进行 Wrapper 式的特征选择,本文采用 ELM 来评价基因优劣。根据 ELM 分类器返回的验证集上的准确率评价每个粒子的适应度值,通过不断更新 PSO 中群体粒子的位置和速度来搜索全局最优解。

对选取的基因的评价,利用 ELM 分类器计算 PSO 和聚类算法选择出来的特征基因的适应度,评价函数为

$$\text{fitness}(i)=\alpha\cdot\text{Accruacy}+(1-\alpha)\cdot(1/\text{geneNum}) \tag{20}$$

在特征选择过程中,应该选择样本测试精度高、基因个数少的粒子,即要选择适应度值最大的那个粒子,所选择的基因是依赖于 ELM 分类器的。在 PSO 中,一个粒子代表选择的一组基因子集,粒子在搜索过程中通过基因子集在分类器中评价,即 PSO 的适应值函数,更新个体最好位置和全局最好位置,直到达到最大迭代次数得到一组最优基因子集,最后利用分类器得到测试准确率,较好的即为提取到的关键基因。其中基因子集的评价函数的过程中,样本测试精度通过 ELM 分类器来完成的。选择出特征基因之后采用 ELM 建立分类模型,然后根据建立的模型测试分类正确率。

算法步骤描述如下:

(1) 利用信息增益方法,对原始基因进行过滤,形成精简的基因子集 FS;

(2) 利用 K 均值聚类方法对 FS 进行聚类,将 FS 聚类为规定的簇数;

(3) 使用 ELM 判断每一簇中基因的分类性能,并选择具有较高分类性能的簇中的基因作为特征基因子集 FSC;

(4) 将 FSC 的聚类中心作为 PSO 的初始化位置,每一个簇作为单独的搜索空间进行 PSO 搜索;

(5) 对选取的基因的评价,如果满足要求的指标,则基因的选择过程结束,接下来进行步骤 6;如果不满足要求,则采用式(10)和式(11)进行最优值和粒子位置和速度的更新,重新进行特征选择;

(6) 选择出特征基因之后采用 ELM 建立分类模型,然后根据建立的模型测试分类正确率。

3 实验结果分析

为了验证算法的有效性,本文在 3 个基因表达数据集上进行仿真实验。白血病(Leukemia)、结肠癌(colon)、小圆蓝细胞(SRBCTs)[16]。试验中用到的所有仿真都是在 Matlab 2010a 中实现的,所用计算机的配置为酷睿双核 2.5 GHz,2 GB 内存。由于基因表达谱数据的样本数非常少,所有的实验都采用 K-折交叉验证的方法,其中 k 值均选择为 5。

在开始特征选择算法之前,把数据集进行归一化处理,使得样本的每一维特征向量的均值为 0,方差为 1。接下来对基因表达谱数据使用信息增益的方法进行初步的选择,确定初选的基因子集,一般选择前 200 个信息增益值最大的基因构成候选基因子集。然后对这个子集进行 k 均值聚类,最佳的 k 值根据样本类别数而定,3 个数据集分别为 3,3 和 5。对基因表达谱数据的训练集的基因进行聚类后,将 ELM 作用于每个簇,得到该簇中最优基因子集的分类性能。图 1～3 分别为白血病、结肠癌和小圆蓝细胞数据集的每个聚类簇中不同数目的基因组合得到的最好分类精度。其中 ELM 分类器使用的激活函数为 sigmoid 函数,最优隐藏层节点数通过递增的方式从 1 增加到与训练样本数相等。

从图 1～3 中可以看出,随着选择基因数目的增多,分类正确率都基本呈现先增加后减少的趋势。在白血病数据集中单独使用簇 1 和簇 3 对样本进行分类获得的分类精度较簇 3 高;结肠癌数据集中簇 2 和簇 3 获得较簇 1 高的分类精度;小圆蓝细胞中簇 1 相较其他簇获得的分类精度比较低。所以根据 2 节中的描述,白血病数据集中簇 2 被视为冗余基因的集合,结肠癌数据集中簇 1 被视为冗

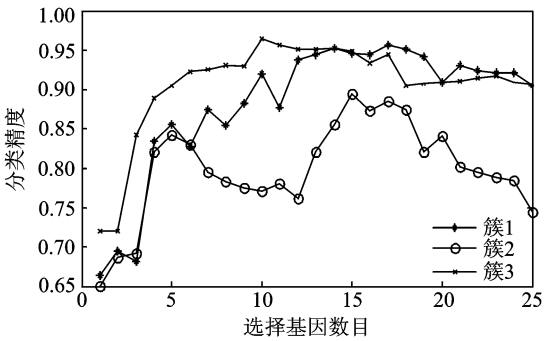

图 1　白血病数据不同簇中不同基因子集的分类性能

Fig. 1　The Leukemia dataset′s classification accuracy when sub-genes come from different clusters

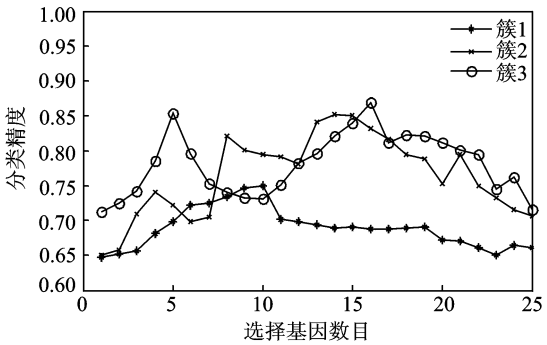

图 2　结肠癌数据不同簇中不同基因子集的分类性能

Fig. 2　The colon dataset′s classification accuracy when sub-genes come from different clusters

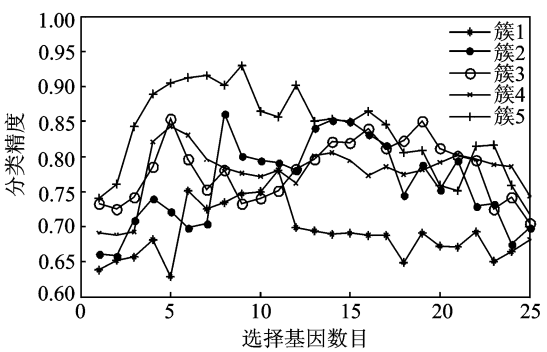

图 3　小圆蓝细胞数据不同簇中不同基因子集分类性能

Fig. 3　The SRBCT dataset′s classification accuracy when sub-genes come from different clusters

余基因的集合，小圆蓝细胞中簇 1 被视为冗余基因的集合。

通过以上对基因聚类结果的分析，分别得到了 3 个数据集的备选特征基因子集。再用 PSO 结合 ELM 搜索最优基因子集。在搜索过程中 PSO 算法的种群规模设置为 10，最大迭代次数为 100，初始学习因子 $c1=2$，$c2=2$，惯性权重 $w=0.8$；ELM 分类器使用 sigmoid 激活函数，最优隐藏层节点数通过递增的方式获得（从 1 增加到与训练样本数相等），适应度调节参数 α 设置为 0.8。

当适应度值达到一定的阈值或者不在变化的时候，或者迭代达到最大次数时，搜索过程便结束。最大适应度值时对应的被搜索到的基因也就是该算法获得的最优关键基因。

适应度的设定是分类精度和被选择基因个数的综合指标，适应度值越高，说明分类精度越高，被选择的基因个数越少。分类精度越高则表明癌症的诊断率越高，被选择的基因个数越少表明获得的靶向基因越精确。

对使用本文方法选择后的样本进行分类，最优关键基因的个数和分类精度的结果在表 1 中给出。其中平均分类精度是 30 次结果的平均值。

表 1　关键基因个数与分类精度

Table 1　Relationship between key genes′ number and classification accuracy　%

样本集	基因个数	平均分类精度	最优分类精度
白血病	3	100	100
结肠癌	3	93.62	100
小圆蓝细胞	9	100	100

从分类结果可以看出使用该方法获得的关键基因子集能够获得非常高的分类精度，且能够尽可能多的消除冗余基因。在白血病数据集上，只需要 3 个关键基因，就可以获得 100％的分类精度；在小圆蓝细胞数据集上只需要 9 个关键基因即可以获得 100％的分类精度；在较难分类的结肠癌数据集上，只需要 3 个关键基因便可以获得 93.62％的平均分类精度，但是在分类效果较好的时候能够获得 100％的分类精度。3 个数据集上的关键基因可以用表 2 描述。

表 2　本文方法选择的关键基因描述

Table 2　Description of the genes selected by the proposed method

数据集	分类精度	基因名称	基因描述
白血病	100％	X03934	GB DEF = T-cell antigen receptor gene T3-delta
		U05259	MB-1 gene
		M63138	CTSD Cathepsin D
结肠癌	93.6％	L11706	Human hormone-sensitive gene
		T51023	HEAT Shock Protein Hsp 90-Beta
		H20709	Myosin Light Chani Alkali
		FGFR4	Fibroblast growth factor
小圆蓝细胞	100％	FVT1	Follicular lymphoma variant translocation
		IGF2	Human DNA for insulin-like growth factor
		⋮	⋮

经查阅相关资料[2-3]可以发现，使用此方法获得的特征基因，确实是该基因表达谱数据对应的癌症的关键基因。因此该方法是具有很强的适用价值，不仅能够提高癌症的诊断率，而且能够有效获得靶向基因，为生物医学提供诊断的依据。

最后将本文所提方法与几种经典方法以及第 3 章中提出的 PSO-Selection 方法进行比较，包括分类精度和选择基因个数以及算法的耗时 3 个方面，比较结果见表 3～5。

表 3 本文方法与其他方法在白血病数据集上的比较

Table 3 Comparison of the proposed method with other method on Leukemia dataset

特征选择方法	基因子集个数	分类精度/%	耗时/s
T-statistic	82	98.61	3.3
信噪比	5	95.88	3.6
PSO-Selection	43	98.61	11.3
PSO-ELM	7	100	874.2
GA-SVM	7	100	3 158.6
PSO-SVM	7	100	1 069.9
本文方法	3	100	497.8

表 4 本文方法与其他方法在结肠癌数据集上的比较

Table 4 Comparison of the proposed method with other method on SRBCT dataset

特征选择方法	基因子集个数	分类精度/%	耗时/s
T-statistic	35	91.61	2.8
信噪比	36	91.5	2.9
PSO-Selection	69	91.78	10.1
PSO-ELM	4	93.63	627.8
GA-SVM	4	93.65	2 501.4
PSO-SVM	4	93.65	871.2
本文方法	3	93.62	301.5

表 5 本文方法与其他方法在小圆蓝细胞数据上的比较

Table 5 Comparison of the proposed method with other method on colon dataset

特征选择方法	基因子集个数	分类精度/%	耗时/s
T-statistic	18	100	3
信噪比	18	100	3.2
PSO-Selection	43	100	10.8
PSO-ELM	15	100	715.6
GA-SVM	15	100	2 604.7
PSO-SVM	15	100	903.5
本文方法	9	100	302.8

从表 3～5 中可以看出，本文方法与其他 6 种方法相比，使用最少的基因子集便可以获得与经典的 Wrapper 方法近似的分类精度，在 3 个数据集上，白血病和小圆蓝细胞均获得 100% 的分类精度，在比较难分类的结肠癌数据集上获得的分类精度只比最优秀的特征选择方法低了 0.03%。从算法的耗时上分析，可以看出本文方法虽然远高于 T-statistic、信噪比和 PSO-Selection 方法，但是选择的基因子集个数比这 3 种方法少很多，且分类精度普遍高于这 3 种方法；而与 PSO-ELM，GA-SVM，PSO-SVM 方法相比[15,16]，在基本上没有降低分类精度的前提下，大大地降低了算法的耗时。虽然本文方法比 PSO-ELM 方法多了基因聚类以及聚类后簇的选择过程，但是在进行 PSO 搜索之前已经将搜索范围缩小，且使用聚类中心作为 PSO 的初始位置使得搜索过程更快趋于最优，这两个因素都使得本文方法的耗时远低于 PSO-ELM 方法。

4 结束语

为了有效降低基因表达谱数据基因之间的冗余度，本文提出了一种基于聚类和粒子群算法的基因选择方法。因为聚类算法可以根据基因的功能将具有相同功能的基因聚成一簇，不同功能的基因聚在不同的簇，通过合理的预处理，含有大量噪声的信息簇被移除，而具有高贡献度的基因簇的基因子集构成候选特征基因作为 PSO 的搜索空间。从实验结果可以看出，本文方法能够成功选择较少数目但是有较高分类率的基因子集。

参考文献：

[1] 黄德双. 基因表达谱数据挖掘方法研究[M]. 北京：科学出版社，2009.
Huang Deshuang. Research on data mining of gene expression[M]. Beijing: Science Press, 2009.

[2] 郑继平. 基因表达调控[M]. 合肥：中国科学技术出版社，2012.
Zheng Jiping. Regulation of gene expression [M]. Hefei: Chinese Science and Technology Press, 2012.

[3] 杨华. 基于粒子群算法的特征选择方法研究[D]. 长沙：湖南大学，2010.
Yang Hua. Research on significant genes selection method based on PSO algorithm[D]. Changshang: Hunan University, 2010.

[4] Golub T R, Slonim D K, Tamayo P, et al. Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286:531-537.

[5] Liu H Q, Li J Y, Wong L. A comparative study on feature selection and classification methods using Gene expression profiles and proteomic patterns[J].

Genome Informatics, 2002, 13:51-60.

[6] Zhao Z, Wang L, Liu H. Efficient spectral feature selection with minimum redundancy[C]// Proceedings of the National Conference on Artificial Intelligence. Atlanta, Georgia, USA: [s. n.], 2010, 1:673-678.

[7] Chris D, PengH C. Minimum redundancy feature selection from microarray gene expression data [J] J Bioinform Comput Biol, 2005, 3(2):185-205.

[8] 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.

[9] 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.
Liu Qinghe, Liang Zhengyou. Optimized approach of feature selection based on information gain [J]. Computer Engineering and Application, 2011, 47(12):130-132.

[10] 任江涛，孙婧昊，黄焕宇. 一种基于信息增益及遗传算法的特征选择算法[J]. 计算机科学，2006，10(33)：193-196.
Ren Jiangtao, Sun Jinghao, Huang Huanyu. Feature selection based on information gain and GA [J]. 2006, 10(33): 193-196.

[11] Leung Y K, Hung Y. A multiple filter multiple wrapper approach to gene selection and microarray data classification[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(1):108-117.

[12] Huang G B, Ding X J, Zhou H M. Optimization method based extreme learning machine for classification[J]. Neurocomputing, 2010, 74:155-163.

[13] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006(70):489-501.

[14] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: a new learning scheme of feedforward neural networks[C]// Proceedings of International Joint Conference on Neural Networks(DCNN2004). Budapest, Hungary: [s. n.], 2004:25-29.

[15] 陆慧娟. 基于极限学习机集成的肿瘤基因表达数据分类[D]. 徐州：中国矿业大学，2013.
Lu Huijuan. Research on tumor gene expression data classification[D]. Xuzhou: China Mining University, 2013.

[16] 郑馨，王勇，汪国有. EM 聚类和 SVM 自动学习的白细胞图像分割算法[J]. 数据采集与处理，2013，28(5)：614-619
Zheng Xin, Wang Yong, Wang Guoyou. White blood cell segmentation using expectation-maximization and automatic support vector machine learning [J]. Journal of Data Acquisition and Processing, 2013, 5(28): 614-619.

作者简介：刘金勇(1986-)，男，硕士研究生，研究方向：人工智能与模式识别；郑恩辉(1975-)，男，副教授，研究方向：模式识别与人工智能；陆慧娟，女，教授，研究方向：机器学习，E-mail：hzliujy@126.com。