基于 SVM 的甲壳动物线粒体基因分析方法

徐启华[1], 耿　帅[2], 肖　晓[2], 申　欣[3]

(1. 淮海工学院 电子工程学院, 江苏连云港 222005; 2. 中国矿业大学 信息与电气工程学院, 江苏徐州 221116; 3. 淮海工学院 海洋学院, 江苏连云港 222005)

摘要: 甲壳动物线粒体基因组蕴涵了物种进化历程中重要的遗传信息, 如何有效地利用这些保留在基因组中的基因序列和基因顺序信息, 是甲壳动物线粒体基因组研究的一个重点方向。为了进一步探讨甲壳动物稳定、可靠的系统发育关系, 本文利用支持向量机的分类功能实现了甲壳动物线粒体基因组基因区与基因间区、编码区与非编码区的准确分类和预测, 同时为了提高分类学习机的泛化能力, 使用了交叉验证方法和粒子群算法优化选取支持向量机相关训练参数。通过 MATLAB 仿真分析的方法, 对 10 种甲壳动物线粒体基因组序列的基因区和基因间区进行分类, 以及对 5 种甲壳动物进行线粒体基因组序列中编码区和非编码区的分类, 获得了较好的分类准确率。仿真结果表明本文方法是可行的和有效的, 能够出色地应用于甲壳动物线粒体基因组序列的研究分析。

关键词: 甲壳动物; 线粒体; 基因分析; 支持向量机; 交叉验证; 粒子群算法

中图分类号: P735　　**文献标识码**: A　　**文章编号**: 1000-3096(2015)05-0054-08

doi: 10.11759/hykx20130731001

甲壳动物是后生动物中形态结构和栖息环境差异、多样性最高的动物类群。加之它们在海、淡水域环境的优势地位、巨大产量、生态意义和在渔业生产中的重要价值, 因而在国际上受到多方面的重视, 使甲壳动物的研究近十几年来取得了飞跃的发展[1]。然而, 甲壳动物在节肢动物门内的分类地位和甲壳动物本身的系统发生问题, 长期以来是学术争论的焦点, 迄今仍然是悬而未有定论[2]。并且由于甲壳动物形态结构高度复杂、种属形态标记少、近缘种和同属种同域分布普遍存在, 但遗传特征却非常不同, 给以形态和生理特征为主的分类鉴定和系统发生关系研究带来了困难, 同时也制约了甲壳动物水平的整体提高[3]。

甲壳动物线粒体基因组在核苷酸组成与偏好、基因组成、基因间区、基因排列、编码区域非编码区等方面具有自身的独立性, 因此, 现有的基因预测方法, 在分析甲壳动物线粒体基因组及构建分子系统演化关系时常常不能得到稳定、可靠的结果[3] 。

支持向量机(SVM, support vector machine)是一种新兴的基于统计学习理论的机器学习方法。近几年来, SVM 已经被用来解决生物信息中的很多问题: 蔡春等[4]、赵丹[5]给出了一般基因序列的特征提取方法并归一化处理特征向量, 通过交叉验证的思想选择惩罚参数和核函数宽度参数, 实现了基因序列的准确分类; Niijima 等人利用 SVM 进行致病基因的识别, 识别出一小部分有用于理解癌症的层机制基因, 有助于设计出不太昂贵的治疗试验[6-7]; 刘建丽等[8]、徐健等[9]提出了基于 SVM 的人类基因序列分类方法; 另外, 也有学者利用 SVM 进行了蛋白质结构预测以及识别由 DNA 转录到 mRNA 时翻译成蛋白质的起始位点等研究工作, 取得了优于聚类分析、神经网络、隐马尔科夫模型等方法的效果[8]。

本文采用 SVM 技术对甲壳动物线粒体基因进行研究, 根据已有的 61 个甲壳动物线粒体基因组数据, 对其基因区和基因间区进行精确分类, 并对基因区中的编码区和非编码区进行精确区分, 实现甲壳动物线粒体基因组基因区与基因间区、编码区与非编码区的准确预测。这将对探讨和解决甲壳动物系统学中的诸多相关问题具有积极的意义。

1　基本概念

1.1　SVM 基本原理

支持向量机是 Vapnik 提出的一种基于统计学习理论的新型机器学习方法, 已经成为机器学习界的

收稿日期: 2014-02-21; 修回日期: 2014-05-28

基金项目: 江苏省海洋资源开发研究院开放课题(JSIMR09C07); 江苏省海洋生物技术重点建设实验室开放课题(2009HS12)

作者简介: 徐启华(1962-), 男, 陕西山阳人, 博士, 教授, 研究方向为智能控制, 电话: 13851280983, E-mail: xuqh@hhit.edu.cn

研究新热点，并在模式识别、回归分析、函数逼近、信号处理等领域得到了成功应用[10]。支持向量机根据结构风险最小化原则，可以尽量提高学习机的泛化能力，同时，通过将优化问题转化为求解一个凸二次规划的问题所得的解是唯一的全局最优解，这样就避免了神经网络的局部极值问题。另外，它巧妙地解决了维数问题，使得其算法的复杂度与样本维数无关。

采用支持向量机进行模式分类的主要思想是：通过事先选择的非线性映射将输入向量映射到高维特征空间，在这个空间中构造满足分类要求的线性最优超平面来分割训练样本集，并且使训练样本集中的点距离该最优超平面尽可能地远。

分类的具体算法如下[11]：

训练样本为 $T=\{(x_i,d_i)|i=1,2,\cdots,n\}$，其中 $x_i\in R^l$ 是第 i 个输入模式，$d_i\in\{+1,-1\}$是对应的期望输出。

首先用一非线性映射 $\varphi(x)=[\varphi_1(x),\varphi_2(x),\cdots,\varphi_N(x)]^T$ 把输入数据从原空间映射到 N 维特征空间，在特征空间中构造最优分类超平面

$$\sum_{j=1}^{N} w_j\varphi_j(x)+b=0$$

其中 $w=[w_1,w_2,\cdots,w_N]^T$ 表示把特征空间连接到输出空间的线性权值向量，b 表示偏置。上述问题即为在满足条件

$$d_i\left[\sum_{j=1}^{N} w_j\varphi_j(x)+b\right]\geqslant 1-\xi_i$$

时最小化 $\|w\|$ 的问题，这里 ξ_i 为松弛变量。写成标准形式为

$$\left.\begin{aligned}&\min J(w,\xi)=\frac{1}{2}w\cdot w+c\sum_{i=1}^{n}\xi_i\\ &\text{s.t.}\quad d_i\left[\sum_{j=1}^{N} w_j\varphi_j+b\right]\geqslant 1-\xi_i\\ &\xi_i\geqslant 0,i=1,2,\cdots,n\end{aligned}\right\}\tag{1}$$

c 为影响分类精度的可调参数。用 Lagrang 乘子法求解这一优化问题，建立 Lagrang 函数

$$\begin{aligned}L(w,b,\xi,\alpha,\beta)=&\frac{1}{2}w\cdot w+c\sum_{i=1}^{n}\xi_i-\\ &\sum_{i=1}^{n}\alpha_i\left\{d_i[w^T\varphi(x_i)+b]-1+\xi_i\right\}-\sum_{i=1}^{n}\beta_i\xi_i\end{aligned}\tag{2}$$

其中 α_i,β_i 为非负的 Lagrang 乘子。求解此优化问题，可以得到

$$\left.\begin{aligned}&\sum_{i=1}^{n}\alpha_i d_i=0\\ &w=\sum_{i=1}^{n}\alpha_i d_i\varphi(x_i)\\ &c-\alpha_i-\beta_i=0\end{aligned}\right\}\tag{3}$$

把(3)代入(2)，由式(1)描述的问题转化为最大化下面的泛函

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j d_i d_j[\varphi^T(x_i)\cdot\varphi(x_j)]$$

SVM 并不直接求解高维特征空间的点积 $\varphi^T(x_i)\cdot\varphi^T(x_j)$，而是用原空间的核函数代替它。核函数是满足 Mercer 条件的对称函数

$$K(x,x_i)=\varphi(x)^T\cdot\varphi(x_i)=\sum_{j=1}^{N}\varphi_j(x)\varphi_j(x_i)$$

最终，优化问题转化为

$$\left.\begin{aligned}&\max Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j d_i d_j K(x_i,x_j)\\ &\text{s.t.}\quad\sum_{i=1}^{n}\alpha_i d_i=0\\ &0\leqslant\alpha_i\leqslant c,i=1,2,\cdots,n\end{aligned}\right\}\tag{4}$$

把它整理成以 α_i 为变量的标准形式二次优化问题，可以方便地求解。求出各 α_i 后，即可得到 w，并可根据全部支持向量(support vector)得到 b 的平均值

$$b=\frac{1}{I_s}\sum_{i\in I}\left[d_i-\sum_{j=1}^{n}d_j\alpha_j K(x_i,x_j)\right]$$

I 为支持向量的集合，I_s 为支持向量的个数。

样本训练完成后，获得数据分类的最优超平面、支持向量和相应参数构成的分类器。未知测试样本 x 按下式进行分类预测

$$y(x)=\text{sgn}\left[\sum_{i=1}^{n}d_i\alpha_i K(x,x_i)+b\right]\tag{5}$$

选择不同的核函数，可构造不同的 SVM，常用的核函数有多项式核函数、Gauss 核函数、Sigmoid 核函数等，本文采用 Gauss 核函数

$$K(x,x_i)=\exp\left(-\frac{|x-x_i|^2}{2\sigma^2}\right)$$

其中 σ 是核函数参数，有时为了书写方便也用 $g=1/2\sigma^2$ 来表示核函数参数。

1.2 交叉验证方法

对于 SVM 训练中的惩罚参数 c 和核函数参数 g 的选择，本文选择使用交叉验证(CV, Cross Validation)方法[12]，可以在某种意义下得到最有参数，同时有效地避免了过学习和欠学习问题。CV 是用来验证分类器或回归器性能的一种统计分析方法，其基本思想是在某种意义下将数据进行分组，一部分作为训练集，另一部分作为验证集；其方法是用训练集数据对分类器或回归器进行训练，再利用得到的模型对验证集数据进行测试，以得到的分类准确率或平均均方根误差作为评价分类器或回归器的性能指标。

最常用的 CV 方法是 K-fold Cross Validation (K-CV)方法，将原始数据均分成 $K(K\geqslant 3)$组，将每个子数据集分别作为一次验证集，同时其余的 K-1 组子数据集作为训练集，用得到的 K 个模型对验证集的分类准确率或平均均方根误差的平均数作为此分类器或回归器的性能指标。将 c 和 g 限制在一定范围内，对其进行划分网格进行搜索，采用上述的 K-CV 方法可以得到不同的 c 和 g 对应的分类正确率或平均均方根误差，选择最大分类正确率或最小平均均方根误差对应的 c 和 g 即是要选择的参数值。

1.3 粒子群算法

粒子群算法[13](PSO, Particle Swarm Optimization)是 Kennedy 和 Eberhart 于 1995 年最早提出的一种源于鸟类捕食行为的优化算法，鸟类在捕食时，每只鸟找到食物的最简单有效的方法是找到距离食物最近的鸟然后搜索其周围区域。

PSO 算法中每个粒子都代表一个潜在解，对应一个由自适应度函数决定的适应度值。PSO 算法首先要初始化一群粒子，每个粒子都由位置、速度和适应度值三部分组成。粒子在可行解空间中运动，通过跟踪个体极值 *Pbest* 和群体极值 *Gbest* 来更新个体位置。

假设在一个 D 维搜索空间中，$X=(X_1,X_2,\ldots,X_n)$是由 n 个粒子构成的种群，$X_i=(x_{i1},x_{i2},\cdots,x_{iD})^T$ 是第 i 个粒子在空间中的位置，对应的适应度值由建立的适应度函数决定。定义 $V_i=(V_{i1},V_{i2},\cdots,V_{iD})^T$ 表示第 i 个粒子的速度，个体最优值和种群的全局最优值分别用 $P_i=(P_{i1},P_{i2},\cdots,P_{iD})^T$ 和 $P_g=(P_{g1},P_{g2},\cdots,P_{gD})^T$ 表示。

所谓优化就是一个迭代的过程，粒子通过个体最优值和全局最优值根据下面的公式来更新自身的位置和速度[13]。

$$V_{id}^{k+1}=wV_{id}^{k}+c_1r_1\left(P_{id}^{k}-X_{id}^{k}\right)+c_2r_2\left(P_{gd}^{k}-X_{id}^{k}\right) \quad (6)$$

$$X_{id}^{k+1}=X_{id}^{k}+V_{id}^{k} \quad (7)$$

式中，d=1,2,…,D; i=1,2,…,n; w 和 k 分别为惯性权值和当前迭代次数; c_1、c_2 是非负的加速度因子; r_1、r_2 是[0,1]上的随机数。在粒子群优化过程中，一般会对位置和速度设置搜索范围$[-X_{\max}, X_{\max}]$、$[-V_{\max}, V_{\max}]$，这样可以提高搜索的效率。

本文将使用交叉验证方法和粒子群算法进行优化选择支持向量机的惩罚参数和核函数宽度，大大地改善了学习机的泛化能力。

2 SVM 基因分类方法

2.1 SVM 训练属性选择

对于一个给定的 DNA 序列，可以利用滑动窗口算法计算特定短序列在这个给定 DNA 序列中出现的频率，本项目取短序列的长度为 2~6。

假设 l 表示滑动窗口的长度(以 l=3 为例)，滑动窗口最初位于 DNA 序列的开始点，此时得到第一个长度为 l 的字符串。滑动窗口依次向右移动一个字符，直至到达 DNA 序列的尾端[1]，如图 1 所示。

图 1 利用滑动窗口截取短序列示意图

Fig.1 Schematic diagram of short sequence intercepted with the sliding window

根据上述滑动窗口方法能够得到一系列窗口短序列，设 W 表示给定的一段 DNA 序列，S 表示某一特定的短序列，$N(S)$表示短序列 S 在这一段 DNA 序列 W 中出现的次数，则短序列 S 出现的频率 f 可由下式得到[8]

$$f(S)=N(S)/C(S) \quad (8)$$

其中 $C(S)=length(W)-length(S)+1$ 表示 DNA 序列 W 中长度与短序列 S 相同的短序列的数量，$length(W)$为给定 DNA 序列 W 的长度，$length(S)$为特定短序列 S 的长度。

通常用*f(S)*表示短序列在一个 DNA 序列中出现的频率，而对于一种甲壳动物的线粒体基因组序列，可能是一个含有 t 个 DNA 序列的 DNA 全序列集 $E=\{W_i|i=1,2,\cdots,t\}$，如果短序列 S 在 W_i 中出现至少一次，则称S在W_i中出现。设$h(S)$表示短序列S在含有t个DNA序列的DNA全序列集E中出现的次数，则$h(S)/t$表示短序列 S 在整个 DNA 集 E 中的出现率。因此，可以用下式来表示短序列 S 在整个 DNA 序列中出现的频率[8]

$$F(S)=\left(\sum_{i=1}^{t}N_i(S)\Big/\sum_{i=1}^{t}C_i(S)\right)\cdot(h(S)/t) \tag{9}$$

以基因区和基因间区的分类为例。为了区分基因区和基因间区，引入相对差的概念，设 $F_1(S)$表示短序列S在基因区序列出现的频率，$F_2(S)$表示短序列S在基因间区序列出现的频率，则在短序列 S 在基因区和基因间区序列中出现的相对差为 $D(S)=[F_1(S)-F_2(S)]/[F_1(S)+F_2(S)]$。显然，如果 $D(S)$越大，即短序列 S 在基因区和基因间区序列中出现的相对差越大，对基因区和基因间区序列的分类越有利，因此，选择相对差大的短序列作为训练属性。

本文所取短序列的长度为 2~6，所有的短序列(候选训练属性)的个数为 5456 个。按短序列的长度(2、3、4、5、6)分为五组，分别计算出各组短序列对应的相对差 D，并按照 D 值的绝对值由大到小的数序排列。选择每组的前 10 个序列作为训练属性来构造训练集，即选择的训练属性向量可以表示为$A=(S_1,S_2,\cdots,S_{50})$。

对编码区和非编码区进行分类时，训练属性的选择与上述做法类似。

2.2 SVM 数据集的构造

对于甲壳动物的线粒体基因组而言，每一种甲壳动物线粒体基因组包括的基因区序列和基因间区序列数目可能是不相同的。为了说明共同性，假设某一类甲壳动物的线粒体基因组序列中含有 m 个基因区序列和 n 个基因间区序列。

用于 SVM 训练和测试的数据集根据上节选择的训练属性来构造。对于每一个基因区序列，其样本的类标号记作+1，训练例的输入向量按照下式计算得到

$$x(i)\_positive=(p_1/q_1,p_2/q_2,\cdots,p_{50}/q_{50}),i=1,2,\cdots,m \tag{10}$$

其中，$p_j(j=1,2,\cdots,50)$是短序列 S_j 在第 i 个基因区序列中出现的次数；q_j(j=1,2,⋯,50)是第 i 个基因区序列中长度与短序列 S_j 相同的短序列的数量。

同样，对于每个基因间区序列，其样本的类标号记作−1，也可以用同样的方法得到训练例的输入向量如下

$$x(i)\_negative=(r_1/o_1,r_2/o_2,\cdots,r_{50}/o_{50}),i=1,2,\cdots,n \tag{11}$$

其中，$r_j(j=1,2,\cdots,50)$是短序列 S_j 在第 i 个基因间区序列中出现的次数；$o_j(j=1,2,\cdots,50)$是第 i 个基因间区序列中长度与该短序列相同的短序列的数量。

由于有些基因间区序列的长度很短，甚至长度只有 1 个碱基，本文将所有长度小于 10 个碱基的基因间区序列忽略不计。假设被忽略的基因间区序列个数为 n_1 个，则这一类甲壳动物线粒体基因组的有效训练样本数为 $m+(n-n_1)$，每个样本有 50 个训练属性。

类似地，可以给出编码区和非编码区序列训练例输入向量的计算公式。

2.3 SVM 的训练和预测

应用台湾大学林智仁教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的 Libsvm 工具箱，编程实现 SVM 的训练和预测，主要步骤如下:

①根据不同要求将数据集分为训练集和测试集，两类的数目可以指定，选择的过程是随机的;

②利用 Matlab 自带的归一化函数 mapminmax 将构造好的训练集和测试集数据进行归一化处理，将所有数据归一化到区间[0,1]上;

③采用交叉验证或粒子群优化算法来选择最优的 SVM 参数 c 和 g;

④利用得到的最佳参数 c 和 g 进行 SVM 训练，得到一个 SVM 分类器;

⑤使用得到的分类器对测试集进行输出预测，得到其分类正确率。

3 基因分类实例仿真

3.1 数据预处理

从 Genbank 数据库中可以得到 61 种甲壳动物线粒体基因组序列数据。针对每一组基因全序列，根据其基因描述信息包含的每一段基因区在全序列中的起始位置和终止位置，利用编写的专门程序来分别提取出基因区和基因间区，将得到的基因区和基因间区序列保存成有关格式的数据。不失一般性，我们选择中华绒螯蟹 *Eriocheir sinensis*、长腕寄居蟹 *Pagurus longicarpus*、日本蟳*Charybdis japonica*、日本绒螯蟹 *Eriocheir japonica*、斑节对虾 *Penaeus monodon*、加州美对虾 *Farfantepenaeus californiensis*、

细角滨对虾 *Litopenaeus stylirostris*、日本囊对虾 *Marsupenaeus japonicus*、凡纳滨对虾 *Litopenaeus vannamei* 和中国明对虾 *Fenneropenaeus chinensis* 10 种甲壳动物的线粒体基因序列进行基因区和基因间区序列的分类计算。

同样，根据 Genbank 文件上的基因描述信息包含的每一段蛋白质编码区在全序列中的起始位置和终止位置，利用编写的专门程序来分别提取出蛋白质编码区和非蛋白质编码区，将得到的蛋白质编码区和非蛋白质编码区序列保存成有关格式的数据。选择中华绒螯蟹、中国龙虾 *Panulirus stimpsoni*、端足 *Onisimus nanseni*、等足 *Eophreatoicus* sp.14 FK-2009 和中国明对虾 5 种甲壳动物的线粒体基因序列进行蛋白质编码区和非蛋白质编码区序列的分类计算。

3.2 训练属性选择

以斑节对虾为例，采用滑动窗口算法实现训练属性的选择，分别选择长度为 2、3、4、5、6 的短序列各 10 个，构成 50 维的训练属性集。

图 2~图 4 给出了长度为分别 2、3、4 的短序列的相对差(取前 10 个作为训练属性)。

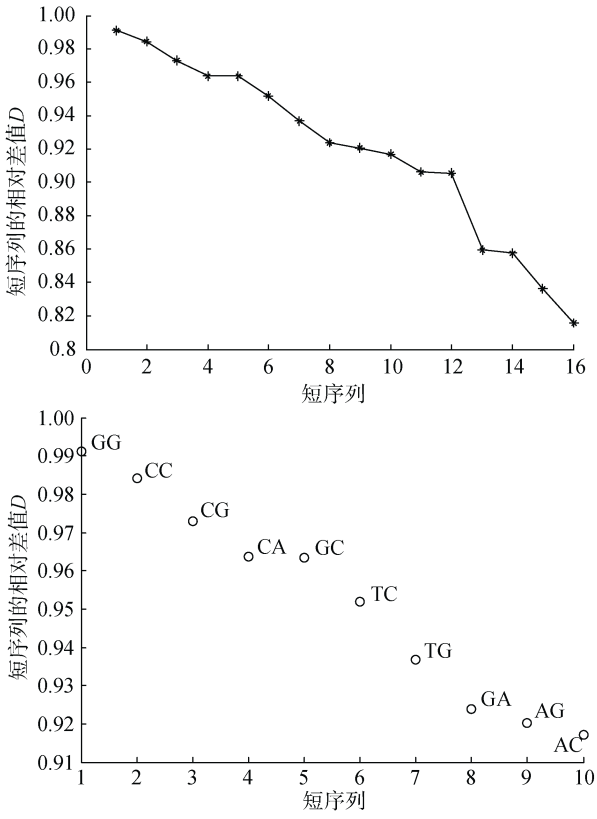

图 2　长度为 2 的短序列的相对差(取前 10 个作为训练属性)

Fig.2　Relative difference of the short sequences with length of 2 (taking the top 10 as the training property)

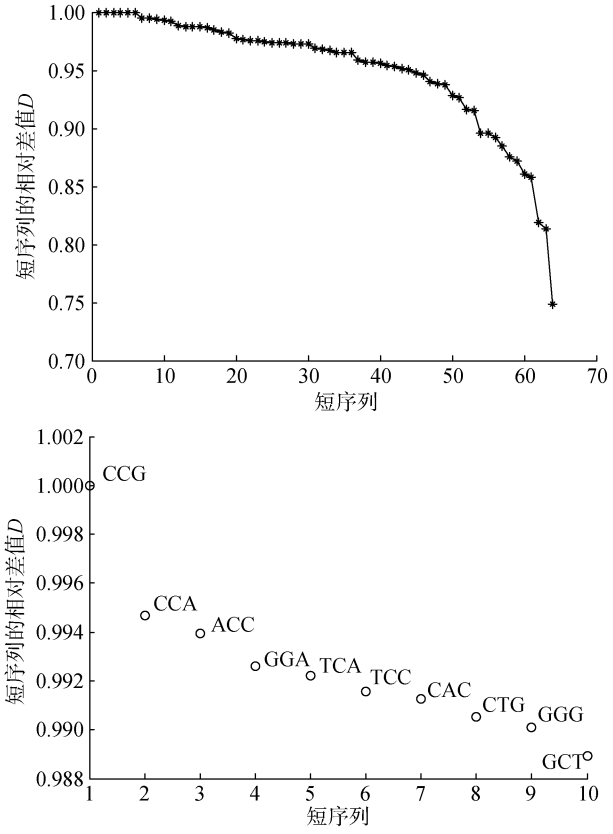

图 3　长度为 3 的短序列的相对差(取前 10 个作为训练属性)

Fig.3　Relative difference of the short sequences with length of 3 (taking the top 10 as the training property)

由图可知，每一组数据中所选的不同长度的 10 个短序列都是最能反映基因区和基因间区的不同，可以作为训练属性来构造训练样本集和测试样本集，能够较好地实现基因区和基因间区的分类。

3.3 基因区与基因间区的分类结果

用训练集训练 SVM 分类器时，使用交叉验证或粒子群优化算法优化对分类准确度影响较大的惩罚因子 c 和核函数宽度相关参数 g，然后用测试样本集来测试所得分类器的分类准确率。仍以斑节对虾为例。

图 5 给出了交叉验证方法得到的 SVM 最优参数 c 和 g，图 6 是在采用上述参数时分类器的分类准确率。图 7 给出了粒子群算法得到的 SVM 最优参数 c 和 g，图 8 是在采用上述参数时分类器的分类准确率。

类似地，可以得到每一种方法下 4 次运行的结果。由于训练集样本和测试集样本是通过随机选择构成的，所以每次执行的结果可能会存在差异，但求其平均值作为最后的分类准确率会更有信服力。

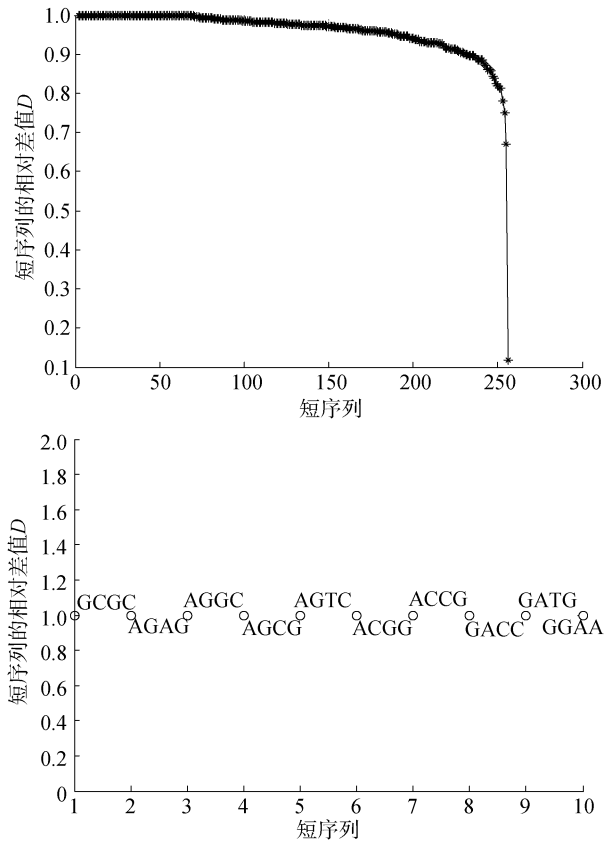

图 4　长度为 4 的短序列的相对差(取前 10 个作为训练属性)

Fig.4　Relative difference of the short sequences with length of 4 (taking the top 10 as the training property)

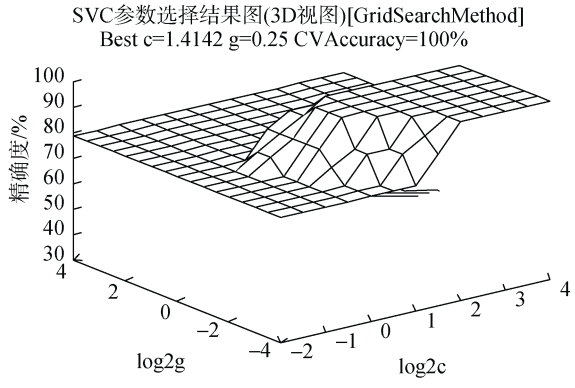

图 5　交叉验证方法得到的最优参数 *c* 和 *g*

Fig.5　The optimum parameters for c and g gained with the cross-validation method

```
打印粗略选择结果
Best Cross Validation Accuracy = 100% Best c = 1.31951 Best g = 0.25
打印精细选择结果
Best Cross Validation Accuracy = 100% Best c = 1.41421 Best g = 0.25
Accuracy = 83.3333% (5/6) (classification)
```

图 6　分类器分类准确率(交叉验证方法寻优参数 *c* 和 *g*)

Fig.6　The classification accuracy (selecting the optimum c and g with the cross-validation method)

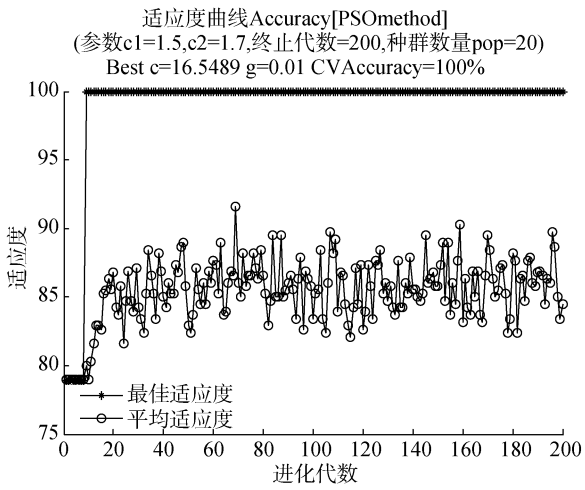

图 7　粒子群算法得到的最优参数 *c* 和 *g*

Fig.7　The optimal parameters c and g obtained by Particle swarm optimization

```
Best Cross Validation Accuracy = 100% Best c = 17.8837 Best g = 0.01
Accuracy = 100% (6/6) (classification)
```

图 8　分类器分类准确率(粒子群寻优参数 *c* 和 *g*)

Fig.8　The classification accuracy (selecting the optimum c and g with the PSO method)

10 种甲壳动物基因区和基因间区的分类结果见表 1, SVM 训练时所用的样本数以及最后预测的样本个数见表 1。

表 1 给出了 10 种甲壳动物基因区和基因间区的分类结果，每一类基因组数据分别随机选择不同的训练集样本和测试集样本处理 4 次，得到 4 个分类正确率，最后求其平均值，且每一组数据训练 SVM 分类器时的参数选择都采取交叉验证方法和粒子群优化算法两种方法。

由表中结果可以得知，本文研究的方法用于实现甲壳动物线粒体基因区和基因间区分类能够得到比较高的分类精度，满足预期目标。

3.4　编码区和非编码区的分类结果

蛋白质编码区和非蛋白质编码区的分类结果见表 2, SVM 训练时所用的样本数以及最后预测的样本个数见表 2。

表 2 给出了 5 种甲壳动物编码区和非编码区的分类结果，相似地，每一类基因组数据分别随机选择不同的训练集样本和测试集样本处理 4 次取其分类正确率的平均值，同样采取交叉验证方法和粒子群优化算法(PSO 算法)两种方法来优化选取每一组数据训练 SVM 分类器时的参数。

表 1　基因区和基因间区的分类结果

Tab.1　The classification results of the gene region and gene intergenic region

物种名称 (训练样本数/预测样本数)	优化方法	分类正确率 (%)				平均分类正确率(%)
斑节对虾(19/6)	交叉验证	100	83.33	100	100	95.83
	PSO	100	100	100	100	100
长腕寄居蟹(28/11)	交叉验证	100	100	100	90.91	97.73
	PSO	90.91	100	81.82	90.91	90.91
中华绒螯蟹(25/11)	交叉验证	72.73	90.91	90.91	90.91	86.36
	PSO	72.73	81.82	81.82	100	84.09
日本囊对虾(14/8)	交叉验证	100	100	100	100	100
	PSO	100	100	100	100	100
凡纳滨对虾(19/6)	交叉验证	100	83.33	100	100	95.83
	PSO	100	83.33	100	100	95.83
中国明对虾(18/8)	交叉验证	100	100	87.50	100	96.87
	PSO	100	100	87.50	100	96.87
日本绒螯蟹(25/15)	交叉验证	93.33	93.33	100	86.67	93.33
	PSO	80	86.67	100	93.33	90
细角滨对虾(19/8)	交叉验证	87.50	100	100	87.50	93.75
	PSO	100	62.50	75	100	84.37
加州美对虾(15/8)	交叉验证	87.50	87.50	87.50	100	90.62
	PSO	100	100	100	100	100
日本蟳(13/5)	交叉验证	100	100	100	80	95
	PSO	100	100	80	80	90

表 2　蛋白质编码区和非蛋白质编码区的分类结果

Tab.2　The classification results of the protein coding region and non-coding region

物种编号 (训练样本数/预测样本数)	优化方法	分类正确率 (%)				平均分类正确率(%)
中华绒螯蟹(24/13)	交叉验证	84.61	92.31	100	92.31	92.31
	PSO	84.61	92.31	92.31	76.92	86.54
中国明对虾(24/13)	交叉验证	84.61	84.61	92.31	92.31	88.46
	PSO	92.31	76.92	84.61	84.61	84.61
端足 *Onisimus nanseni*(24/13)	交叉验证	76.92	92.31	84.61	92.31	86.54
	PSO	84.61	84.61	84.61	84.61	84.61
等足 *Eophreatoicus* sp.14 FK-2009(24/13)	交叉验证	84.61	100	100	84.61	92.31
	PSO	69.23	76.92	100	84.61	82.69
中国龙虾(24/13)	交叉验证	84.61	100	92.31	92.31	92.31
	PSO	84.61	92.31	84.61	92.31	88.46

由表中结果可以得知，本文方法能够用于甲壳动物线粒体编码区和非编码区的分类，且分类精度比较理想，满足预期要求。

4　结束语

本文系统地研究了基于SVM的甲壳动物线粒体基因组基因区与基因间区、编码区与非编码区的准确分类方法，为进一步探讨甲壳动物稳定、可靠的系统发育关系打下基础，可为甲壳动物各个纲(亚纲)之间的系统发生关系及物种鉴定提供基因学的证据，为甲壳动物资源的保护和利用、水产养殖种的杂交培育提供基础和保障。通过 MATLAB 环境下的仿真分析，表明本文研究的 SVM 基因序列分析方法能够用于甲壳动物线粒体基因序列的分析，取得了理想

的分类精度，表明了该方法的可行性和有效性。

参考文献:

[1] 耿帅，韩磊，孟娇. 基于 Matlab 的甲壳动物线粒体 DNA 的点突变检测方法[J]. 淮海工学院学报, 2012, 21(1): 25-28.

[2] 刘瑞玉. 现生甲壳动物 (CRUSTACEA) 最新分类系统[M]//中国甲壳动物学会.甲壳动物学会论文集，北京：科学出版社, 2003: 78-88.

[3] 申欣. 软甲纲动物和星虫动物线粒体基因组特征及分子进化研究[D]. 北京：中国科学院研究生院, 2008.

[4] 蔡春，苗立峰，邓乃扬. DNA 序列特征提取方法研究[J]. 北京联合大学学报：自然科学版, 2008, 22(4): 70-72.

[5] 赵丹. 基于 SVM 分类机的 DNA 序列分类方法[D]. 南昌大学, 2010.

[6] Niijima S, Kuhara S. Gene subset selection in kernel-induced feature space[J]. Pattern recognition letters, 2006, 27(16): 1884-1892.

[7] Chen Z, Li J, Wei L, et al. Multiple-kernel SVM based multiple-task oriented data mining system for gene expression data analysis[J]. Expert Systems with Applications, 2011, 38(10): 12151-12159.

[8] 刘建丽，刘椿年. 基于 SVM 的人类基因序列分类方法[J]. 北京工业大学学报, 2008, 34(8): 884-890.

[9] 徐健，李柏年，张孔生，等. 基于 SVM 分类机的一种 DNA 序列判别方法[J]. 安徽理工大学学报：自然科学版, 2009, 3: 58-61.

[10] Haykin S, Network N. A comprehensive foundation [J]. Neural Networks, 2004, 2.

[11] 徐启华，师军. 基于支持向量机的航空发动机故障诊断[J]. 航空动力学报, 2005, 20(2): 298-302.

[12] 史峰，王小川，郁磊等. MATLAB 神经网络 30 个案例分析[M]. 北京：北京航空航天大学出版社, 2010.

[13] 史峰，王辉，郁磊等. MATLAB 智能算法 30 个案例分析[M]. 北京：北京航空航天大学出版社, 2011.

Analysis method for crustacean mitochondrial gene based on SVM

XU Qi-hua[1], GENG Shuai[2], XIAO Xiao[2], SHEN Xin[3]

(1. School of Electronic Engineering, Huaihai Institute of Technology, Lianyungang 222005, China; 2. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China; 3. Marin School, Huaihai Institute of Technology, Lianyungang 222005, China)

Received: Feb.,21,2014

Key words: crustacean; mitochondria; gene analysis; support vector machine; cross validation; particle swarm optimization

Abstract: Crustaceans mitochondrial genome contains important genetic information in the course of the species evolution, so it is a priority research direction for the crustacean mitochondrial genome to effectively use the gene sequence and the order information reserved in the genome. To further explore the phylogenetic relationship of stablility and reliablility of the crustaceans, the classification function of the support vector machine was used to realize the accurate classification and prediction of both the gene region with gene intergenic region and coding region with non-coding region in the crustacean mitochondrial genome. In addition, in order to improve the generalization ability of the classification learning machine, the cross-validation method and particle swarm optimization algorithm were selected to optimize the training parameters of support vector machine. Through the method of simulation analysis in MATLAB, a better classification accuracy is obtained between the gene region and gene intergenic region of 10 species of crustaceans, and the excellent result is also gained between the coding region and non-coding region of 5 kinds of crustaceans. The simulation result shows that this method is feasible and effective and it can be well used to investigate and analyze the mitochondrial genome of crustaceans.

(本文编辑：梁德海)