安徽农业大学学报, 2015, 42(2): 310-315
Journal of Anhui Agricultural University
[DOI] 10.13610/j.cnki.1672-352x.20150302.024　网络出版时间：2015-3-2　15:27:00
[URL] http://www.cnki.net/kcms/detail/34.1162.S.20150302.1527.024.html

基于粒子群优化的支持向量机算法识别人类基因启动子

张　文，陈园园，张　瑾，骈　聪，李　琴，张良云*
(南京农业大学理学院，南京 210095)

摘　要：人类基因启动子识别是医学研究的基本需要。提取 DNA 序列碱基的 PZ 曲线特征、二核苷酸空间结构特征、保守信号似然得分，以及 K 联体似然得分，结合 GC 含量变化和非均匀指数，构建基于粒子群优化的支持向量机算法来识别人类基因启动子。利用粒子群优化支持向量机参数进行优化避免了人为选择的随机性，并且在分类问题中表现出较好的稳健性。对测试集的 10-折交叉检验结果为：敏感性为 92%，特异性为 91%，马修斯关联系数为 0.83。该结果表明，基于粒子群优化的支持向量机算法能有效识别启动子序列。

关键词：相位特异 PZ 曲线；粒子群优化；支持向量机；启动子预测
中图分类号：Q811.4　　**文献标识码：**A　　**文章编号：**1672–352X (2015)02–0310–06

Recognition of gene promoters in human beings based on the particle swarm optimized support vector machine algorithm

ZHANG Wen, CHEN Yuanyuan, ZHANG Jin, PIAN Cong, LI Qin, ZHANG Liangyun
(College of Science, Nanjing Agricultural University, Nanjing 210095)

Abstract: Recognition of gene promoters in human beings is a basic requirement for medical research. It was achieved through analysis of phase-specific PZ curves of nucleotide, spatial structure of nucleotide, conservative signal and K-mer likelihood score in DNA sequence, as well as GC content changes and in-homogeneity index. The support vector machine algorithm based-particle swarm optimization was proposed to identify human gene promoters. Using PSO algorithm to optimize the parameters of SVM can avoid the randomness of artificial selection and present better robustness in classification. The sensitivity, specificity and MCC tested by the 10-fold cross-validation were 92%, 91%, and 0.83, respectively. The result indicated that PSO-SVM method can be used to effectively identify promoter sequences.

Key words: phase-specific PZ curve; particle swarm optimization (PSO); support vector machine (SVM); promoter prediction

启动子对基因表达、细胞特异性的研究是十分重要的，是构建基因调控网络的一个核心问题。人类基因启动子研究是医学研究的基本需要，如哮喘、β 地中海贫血、鲁宾斯坦泰比综合症等疾病与启动子功能异常有关。启动子位于基因 5′端上游，通过 RNA 聚合酶的作用，使 mRNA 与模板 DNA 准确结合，具有转录起始的特异性。现有的启动子识别方法大致可分为 3 类：(1)基于信号的方法，通过 TATA、CAAT 盒等启动子元件来识别，但这些元件并不是决定启动子功能的唯一元件，假阳率较高。这类方法主要有 NNPP[1]、DragonPF[2]等。(2)基于内容的方法，这种方法主要是统计 DNA 序列 K 联体的频率，即组成成分特征来识别基因启动子，由于 DNA 序列 K 联体种类较多，仅六联体就有 4096 种，统计得出的特征维数过高不易于计算。这类方法主要有 Promoter Inspector[3]、Prometheus[4] 和 IDQD[5]等。

收稿日期：2014-10-22
基金项目：教育部博士点基金（20100097110040），中央高校基本科研业务费专项资金（KYZ201125）和江苏省自然科学基金（BK20140676, BK20141358）共同资助。
共同第一作者简介：张　文，硕士研究生。E-mail：xbpb2008@163.com
陈园园，博士，讲师。E-mail：chenyuanyuan@njau.edu.cn
*** 通信作者：**张良云，教授，博士生导师。E-mail：zlyun@njau.edu.cn

(3)基于 CpG 岛的方法，据统计大多数人类基因启动子和 CpG 岛相关，从而可以通过 CpG 岛对启动子进行识别，但并非所有的启动子与 CpG 岛相关，因此仅依靠 CpG 岛识别启动子也会导致假阳率较高。这类方法主要有 FirstEF[7]等。因此有必要改进现有的方法提高启动子预测的准确率并降低预测的假阳率。

Gangal 等开发的人类基因 RNA 聚合酶 II 启动子预测工具 Prometheus[4]，采用了支持向量机(support vector machine，SVM）技术，使用多项式核函数建立预测模型，输入包括 K 联体频率、CG 含量等特征，10-折交叉验证(10-fold cross validation)的敏感性为 86%，特异性为 87%，关联系数为 0.74，该方法并未对支持向量机的参数进行优化。Tao 等[6]采用遗传算法(genetic algorithm，GA)和粒子群优化(particle swarm optimization，PSO)算法相结合的方法侧重于优化特征来进行启动子预测，由于选取的特征有限，支持向量机的预测性能并未完全体现。作者使用有效的特征提取方法，采用粒子群优化算法优化支持向量机参数，采用径向基函数(Radial Basis Function，RBF)为核函数，建立 PSO-SVM 分类器来识别人类基因启动子，降低了假阳率并且提高了准确率。本文分以下几个部分阐述 PSO-SVM 方法识别人类基因启动子。

1　材料与方法

1.1　材料

本实验数据选取的序列都是人类 DNA 序列，启动子序列来自 Human Promoter Database (http://zlab.bu.edu/~mfrith/HPD.html)、外显子和内含子序列来自 Berkeley Drosophila Genome Project (http://www.fruitfly.org/seq_tools/datasets/Human/)。其中启动子选用 1076 条，外显子选用 890 条，内含子选用 1000 条共 2966 条。

1.2　特征选取

1.2.1　组成成分特征　DNA 序列可以看成是由 A、C、G、T 组成的字符串，每个字符表示一种核苷酸，K 个连续的核苷酸称作一个 K 联体(K-mer)，长度为 K 的 K 联体共有 4^K 种，K 联体的频率反映了序列碱基的偏好信息。统计序列 K 联体的频率，采用极大似然得分的方式对特征进行量化。设第 i 种 K 联体在训练正集和训练负集的频率分布为：

$$f^+_i = \frac{n^+_i}{\sum_{i=1}^{4^K} n^+_i}，\quad f^-_i = \frac{n^-_i}{\sum_{i=1}^{4^K} n^-_i}$$

其中 n^+_i, n^-_i 分别为训练正集和训练负集中的所有碱基序列第 i 种 K 联体出现的频数之和。对任一待测序列 X 计算其 K 联体(K=3，4，5，6)似然得分：

$$S_X = \sum_{i=1}^{4^K} m_i \ln\left(\frac{f^+_i}{f^-_i}\right)$$

m_i 为待测序列 X 的 K 联体频数，S_X 反映了待测序列 X 与正集和负集的相似程度，S_X 越大序列 X 越倾向于启动子。

1.2.2　CpG 岛特征　CpG 岛是一类长度超过 200 bp、C 和 G 含量大于 50%、双核苷酸 CG 出现的次数与估计出现的次数比(Obs/Exp)大于 60%的 DNA 序列。从已知的 DNA 序列统计发现，大约 60%的人类基因启动子与 CpG 岛相关，因此，它可做为人类基因启动子识别的一个特征(图 1 反映了 3 类 DNA 序列的 C、G 含量差异)。设 DNA 序列长度为 N，核苷酸 C、G 以及二联体 CG 的数目分别为 N_C、N_G、N_{CG}，则 C、G 含量 GC 和 Obs/Exp 比值 OE 计算如下：

$$GC = \frac{N_G + N_C}{N}，\quad OE = \frac{N_{CG} \times N}{N_C \times N_G}$$

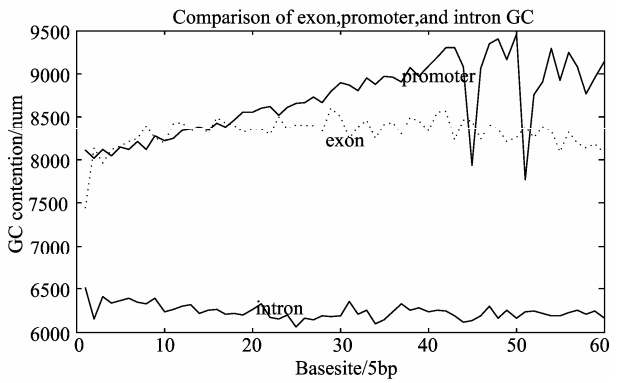


图 1　外显子、启动子和内含子（各取 3000 条）在 5 bp 间隔上的 GC 含量分布

Figure 1　Exon, promoter and intron (each selected 3000) GC content distribution in the interval 5 bp

1.2.3　信号特征　Narang[8]等人揭示了几个重要的启动子元件，如 TATA 盒、CAAT 盒和 GC 盒，它们分别对应于转录因子结合位点 TBP、CBF 和 SP1，出现在转录起始位点(TSS)周围相对固定的位置。统计 TATA 盒在-40 ~ -20 bp，CCAAT 盒在-140 ~ -80 bp，GC 盒在-140 ~ -80 bp、-80 ~ -20 bp、-20 ~ +40 bp 的保守模体序列出现的频率，和前一特征一样计算序列的模体频率的似然得分作为待测序列的信号特征。

1.2.4 二核苷酸空间结构特征 选取 Twist、Tilt、Roll、Shift、Slide 和 Rise[10] 6 类结构参数(具体数值参考文献[10])组成向量 v_j，j=1,2,⋯,6，其中前 3 个是角度参数，后 3 个是距离参数，v_j^i 表示向量 v_j 的第 i 个分量，统计长度为 N 的序列中二联体出现的频次 C_i，j=1,2,⋯,16，则序列的结构特征向量为：

$$V_j=\frac{1}{N-1}\sum_{i=1}^{16}C_i{v_j}^i$$

1.2.5 非均匀指数特征 非均匀指数(in- homogeneity index，IHI)反映了编码区与非编码区 DNA 序列的碱基成分差异性程度。对一段长度为 N 的核酸序列，令 N_a 为序列中对应碱基个数，a=1,2,3,4 分别表示碱基 A、T、G、C。将序列按碱基第 1、第 2、第 3 个位置分为 3 个子序列 N^l，l=1,2,3。令 N_a^l 表示第 l 个子序列碱基 a 的个数，非均匀指数定义为：

$$IHI=\sum_{i=1}^{3}\sum_{a=1}^{4}\left(N_a{}^l-\frac{N^lN_a}{N}\right)/\frac{N^lN_a}{N}$$

1.2.6 PZ 曲线特征 Z 曲线的提出开创了一个利用几何学方法分析和研究 DNA 序列的崭新领域，为了避免 Z 曲线特征中出现的退化现象，对 Z 曲线进行改进提出了 PZ 曲线[11]如下：

$$\begin{cases}x_n=uA_n+\sqrt{v}G_n-\left(uC_n+\sqrt{v}T_n\right)\\ y_n=\sqrt{v}A_n+uC_n-\left(uG_n+\sqrt{v}T_n\right)\\ z_n=uA_n+\sqrt{v}T_n-\left(uG_n+\sqrt{v}C_n\right)\end{cases}$$

其中 u 和 $\sqrt{v}$ 采用交叉检验的方法分别选取具有最佳分类准确率的值：5 和 $\sqrt{\frac{1}{11}}$。

1.3 方法

1.3.1 PSO 优化 SVM 参数 粒子群优化算法是一种基于群体智能理论的优化仿生算法，通过群体中粒子间的合作与竞争产生的群体智能进行搜索，用迭代来寻找最优解并用适应度评价解的质量。粒子群算法比遗传算法更为简单，它没有遗传算法的选择、交叉和变异等复杂操作，具有较低的时间复杂度和较快的收敛速度，被广泛应用在生物信息学领域的优化问题中，如 RNA 二级结构预测和蛋白质结构预测，作者将其应用到 SVM 参数寻优当中预测启动子。

设群体个数为 m，个体维数为 n，每个个体看作一个粒子对应 n 维空间的一个点，设第 i 个粒子位置为 $x_i(x_{i1},x_{i2},\cdots,x_{in})$，第 i 个粒子的速度(即第 i 次迭代的修正量)为 $v_i(v_{i1},v_{i2},\cdots,v_{in})$，PSO 通过循环迭代来完成搜索，每次循环可以得到该粒子的个体最优解 $p_i(p_{i1},p_{i2},\cdots,p_{in})$和全局最优解 $g_i(g_{i1},g_{i2},\cdots,g_{in})$，粒子通过下式迭代更新位置和速度：

$$v_i^{k+1}=\alpha v_i^k+c_1rand_1\left(p_i^{k-1}-x_i^{k-1}\right)+c_2rand_2\left(g_i^{k-1}-x_i^{k-1}\right),$$

$$x_i^{k+1}=x_i^k+v_i^{k+1},$$

其中 k 为迭代次数，$rand_1$ 和 $rand_2$ 为(0,1)之间的随机数，c_1 和 c_2 为学习因子，c_1 和 c_2 根据经验选取初始值 1.5 和 1.7。α 为惯性权重系数，用来控制粒子前一速度对现在速度的影响,本文采用线性递减策略，

$$\alpha=\alpha_{max}-\frac{\alpha_{max}-\alpha_{min}}{n_{max}}\times n,$$

$\alpha_{\min}$ 为初始权重，$\alpha_{\max}$ 为最终权重，$n_{\max}$ 为最大迭代次数，n 为当前迭代次数。

支持向量机(SVM)是在统计学习理论的基础上提出的机器学习方法，广泛应用于解决模式识别问题和函数拟合问题。SVM 的核函数参数与惩罚因子对预测效果有较大的影响，但理论自身未给出核函数参数与惩罚因子的最佳取值方法。

设训练集样本为(x,y)，$(x_i,y_i)\in R^l\times R$，i=1,2,⋯,n，l 为样本维数，n 为训练集样本数量。假设一个线性可分的二分类问题的线性判别函数为：$g(x)=wx+b$，w 和 b 为权向量和偏置量，从而分类决策函数为：$f(x)=sgn(wx+b)$。

根据间隔最大化(maximal-margin)原则，分类决策问题转化为对 w 和 b 寻优，使得间隔 $\frac{1}{2}\|w\|^2$ 最大，等价于求 $\frac{1}{2}\|w\|^2$ 的最小值，约束条件为：

$$y_i(wx_i-b)\geqslant 1-\xi_i$$

其中，ξ_i 是松弛变量，用于对误差的协调，$\xi_i\geqslant 0$。寻找最优决策函数即求解如下二次规划问题：

$$\min_{w,\xi_i}\left\{\frac{1}{2}\left(ww^T\right)+C\sum_{i=1}^{n}\xi_i\right\},$$

其中 C 为惩罚因子且 $C>0$，用于表示对误差较大样本的惩罚程度。对上述问题的求解转化为原问题的对偶问题，然后代入 $w=\sum_{i=1}^{n}a_iy_ix_i$ 得：

$$max\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jK\left(x_i,x_j\right)+\sum_{i=1}^{n}\alpha_i\right\},$$

约束条件为转变为：

$$\sum_{i=1}^{n}y_i\alpha_i=0\left(\alpha_i\in[0,C]\right),$$

其中 a_i 为待求解的 l 维非负拉格朗日乘子。$K(x_i,x_j)$ 为核函数，本文选取的核函数为径向基核函数RBF：

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

$1/\sigma^2$ 为核参数，求解对偶问题方法有选块法、SMO和分解法等。本文采用最快的二次规划优化算法SMO 求解 a_i 和 b，从而得到 SVM 最优分类决策函数：

$$f\left(x\right)=sgn\left\{\sum_{i=1}^{n}\alpha_i y_i K\left(x_i,x\right)+b\right\}$$

用 PSO 对 SVM 参数进行优化，主要是优化惩罚因子 C 和核函数参数 g=1/σ^2，粒子表示为 $x_i(C,g_i)$，粒子速度表示为 v_i=(v_{Ci},v_{gi})。PSO 优化 SVM 参数的目的是使 SVM 算法达到最大分类精度，所以把 SVM 算法的 10-折交叉检验精度作为 PSO 适应度评价解的质量，该算法的流程图如图 2 所示。

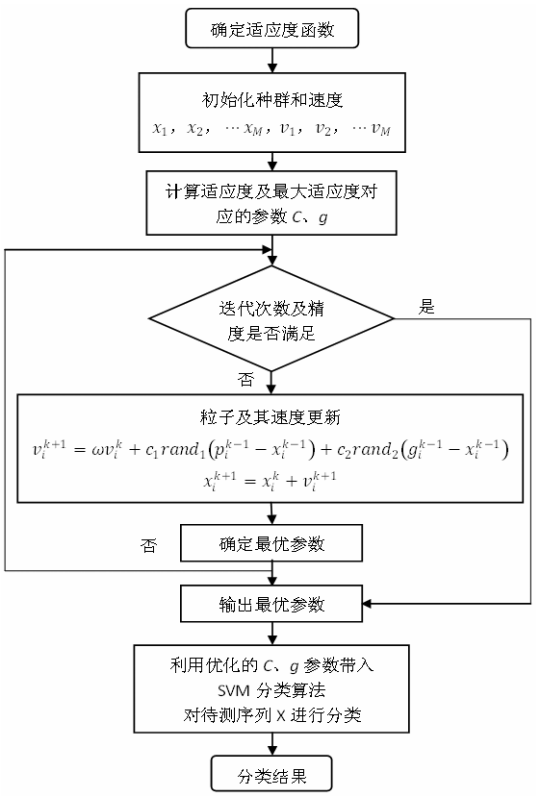


图 2　PSO-SVM 算法流程图

Figure 2　PSO-SVM algorithm flowchart

1.3.2　*PSO-SVM 分类器构建*　系统总分类器由 2 个二分类器组成，分别是启动子与外显子(Promoter-Exon)分类器和启动子与内含子(Promoter-Intron)分类器，每一个子分类器都通过 PSO-SVM 算法构建。Promoter-Exon 分类器和 Promoter-Intron 分类器均选取 PZ 曲线特征、非均匀指数、二核苷酸空间结构特征、GC 含量、六联体得分和信号特征为研究特征，分别为 9 维、1 维、6 维、2 维、3 维和 2 维，共 23 维。

1.3.3　*检验方法和评价标准*　本文采用的检验方法为 10-折交叉检验，即将数据集分成 10 份，轮换将其中 9 份作为训练数据，剩下的 1 份作为测试数据进行试验。把 10 次试验结果的正确率的平均值作为对算法精度的估计，一般需要进行多次 10-折交叉验证，再求其均值，作为对算法准确性的最终估计。本文主要采用以下 3 个指标对预测结果进行评价，分别是敏感性(sensitivity，Sn)、特异性(specificity，Sp)和马修斯关联系数(Mathews correlation coefficients，Mcc)：

$$Sn=\frac{TP}{TP+FN}\times 100$$

$$Sp=\frac{TN}{TN+FP}\times 100$$

$$MCC=\frac{\left(TP\times TN\right)-\left(FP\times FN\right)}{\sqrt{\left(TP+FP\right)\times\left(TN+FN\right)\times\left(TP+FN\right)\times\left(TN+FP\right)}},$$

其中，TP 表示预测正确的启动子序列数，FN 表示没有正确预测的启动子序列数，FP 表示没有正确预测的非启动子序列数，TN 表示正确预测的非启动子序列数。

2　结果与分析

实验环境为 Matlab7.11 版本，CPU 为 I5-2.5 GHz，内存 6 G，调用 LIBSVM[16]工具包，采用 PSO 优化的 SVM 算法[17]对数据进行训练预测。

2.1　不同 C、g 参数的分类准确率对比

不同的 C、g 参数对分类准确率的影响如表 1 所示(以 Promoter-Exon 分类器为例)。从表 1 可以看出，不同的参数 C、g 对分类准确率影响很大。

表 1　不同参数 *C*、*g* 的分类准确率对比

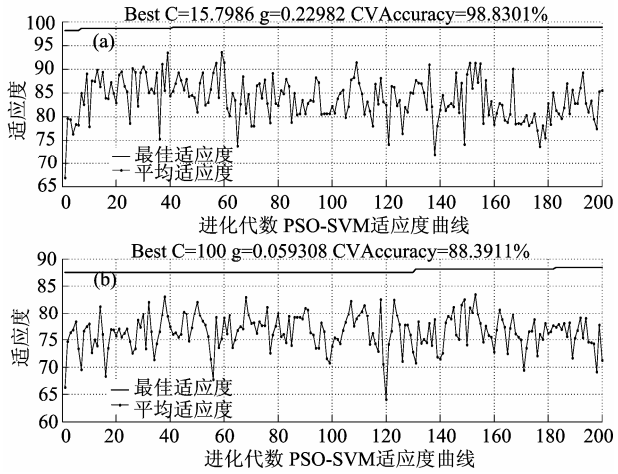
Table 1　Classification accuracy comparison of different parameters of *C* and g

参数 Parameter	*C*	*g*	准确性 Accuracy
1	2.18	1.02	0.575
2	92.72	12.18	0.662
3	64.67	0.048	0.923

2.2　PSO-SVM 二分类器的适应度曲线

构建 PSO-SVM 二分类器 Promoter-Exon 和 Promoter-Intron，经过实验分别得到 PSO-SVM 方法

的适应度曲线如图 3 的(a)和(b)所示。从图 3 可以看出，通过粒子群优化寻找最优参数后的最佳适应度比平均适应度有较大的提高，进而参数优化使得分类精度有明显改善。



(a) Promoter-Exon 分类器的适应度曲线:最优的参数 C、g 分别为 15.79 和 0.23，最佳分类准确率为 98.78%；(b) Promoter-Intron 分类器的适应度曲线: 最优的参数 C、g 分别为 100.00 和 0.06，最佳分类准确率为 88.39%

(a) The curve of Promoter-Exon classifier: the best parameters of C and g are 15.79 and 0.23, respectively, and the best accuracy is 98.78%; (b) The curve of Promoter-Intron classifier: the best parameters of C and g are 100.00 and 0.06, respectively, and the best accuracy is 88.39%

图 3 PSO-SVM 方法的适应度曲线

Figure 3 The fitness curve of PSO-SVM method for the prediction

2.3 PSO-SVM 二分类器的识别准确率

两个二分类器的识别结果如表 2 所示。从表 2 可以看出 PSO-SVM 方法对启动子和外显子分类效果最优，MCC 值为 0.993，这主要源于特征选取方面的优势，PZ 曲线特征体现的是嘌呤(A+G)和嘧啶(C+T)碱基、氨基(A+C)和酮基(G+T)碱基、强氢键(A+T)和弱氢键(G+C)碱基的分布信息，这些信息描述的是编码区碱基特性；非均匀指数特征体现的是蛋白质编码区密码子的碱基偏好性，而启动子序列并不编码蛋白质，从而使得外显子序列的 IHI 值与启动子序列的 IHI 值差异明显。从而启动子与外显子的分类准确率有较大提高。

表 2 二分类器实验结果

Table 2 Two classification experimental results

Classifier	*Sn*/%	*Sp*/%	*MCC*
Promoter-Exon	99.8	99.4	0.993
Promoter-Intron	88.6	92.7	0.812

2.4 PSO-SVM 方法和其他方法的分类准确率对比

文献[4]随机挑选了 100 个启动子序列和 100 个内含子序列，和现在通行的启动子预测软件进行了预测能力比较。本文采用同样的做法，但随机挑选 500 条序列，所得分类结果如表 3 所示。

从表 3 中数据可以看出本文的 PSO-SVM 算法实验结果优于表 3 中的其他方法，本文方法的主要改进在于选取有效特征提取方法，优化支持向量机的参数，采用了 RBF 径向基函数作为核函数，有效克服了表 3 其他 2 种机器学习方法在参数优化上的不足。

表 3 实验结果与其他方法对比

Table 3 Comparison of the experimental results in this article with those using other methods

Algorithm	*Sn*/%	*Sp*/%	*MCC*
NNPP(threshold 0.8)*	32	34	0.34
Prometheus*	86	88	0.74
This Article (PSO-SVM)	93	92	0.83

注：该数据来自文献[4]，NNPP 和 Prometheus 分别代表神经网络方法和支持向量机方法。

Note: The data comes from the literature [4], NNPP and Prometheus represent neural network and support vector machine, respectively.

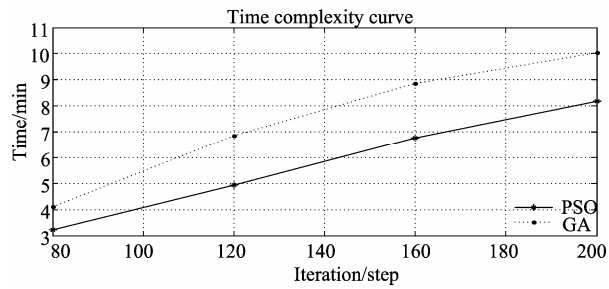


图 4 PSO-SVM 方法和其他优化方法的时间复杂度对比

Figure 4 The fitness curve of PSO-SVM method for the prediction

2.5 PSO 方法和其他优化方法的时间复杂度对比

遗传算法是一种常用的参数优方法，为了进一步说明本方法的可行性，故将粒子群优化方法(PSO)和遗传算法(GA)的时间复杂度做一对比如图 4 所示，横轴表示的是最大进化代数，纵轴表示的是参数寻优和训练的总时间。可以看出，粒子群优化方法的运算时间要小于遗传算法，从而在时间效率上要优于遗传算法。

这是由于粒子群优化方法与遗传算法的信息共享机制是不同的。在遗传算法中，染色体(chromosomes)互相共享信息，所以整个种群的移动

是比较均匀的向最优区域移动。而在粒子群优化方法中，只有全局最优解发送信息给其他粒子，是单向的信息流动，整个搜索更新过程是跟随当前最优解的过程。与遗传算法相比，所有的粒子可能更快的收敛于最优解。

3　小结与讨论

依据基因启动子区碱基的 PZ 曲线特征、二核苷酸空间结构特征、保守信号特征，以及 K 联体的似然得分，结合 GC 含量变化和非均匀指数，构建基于粒子群优化的支持向量机算法来识别人类基因启动子。在特征提取方法上，选取似然得分是本文的一个优势，对识别准确率的提高有很大帮助；在分类方法上本文选取群体智能优化算法优化支持向量机参数是一个较大创新，方法的稳健性和泛化性都有很大改观。

启动子和非启动子序列特征差异性的提取是影响识别效果的一个关键因素，本文选取的 GC 含量变化（如图 1）等特征有效体现了基因启动子与外显子和内含子等非启动子区的差异性；其次是构建分类器选取的分类算法，分类算法对分类结果的提高主要体现在该算法是否能够对提取出的特征有效拟合且在达到误差精度后收敛，应用基于 PSO 优化的 SVM 算法对人类启动子序列进行预测。结果表明，PSO-SVM 方法能够较好地识别启动子序列，优于预测方法 NNPP 和 Prometheus，从而可以选取人类特定的 DNA 长序列进行窗口滑动预测人类启动子。本文采用 SVM 构建了 2 个二分类器，没有把基因间序列作为另一分类进行识别，因此，启动子的预测有待进一步研究和探讨。

参考文献：

[1] Reese M G. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome[J]. Comput Chem, 2001, 26(1): 51-56.

[2] Bajic V B, Seah S H, Chong A, et al. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters[J]. Bioinformatics, 2002, 18(1): 198-199.

[3] Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by Promoter Inspector: a novel context analysis approach [J]. J Mol Biol, 2000, 297(3): 599-606.

[4] Gangal R, Sharma P. Human pol II promoter prediction: time series descriptors and machine learning[J]. Nucleic Acids Res, 2005, 33(4): 1332-1336.

[5] 吕军, 罗辽复. 人类 pol II 启动子的识别[J]. 生物化学与生物物理进展, 2005, 32 (12): 1185- 1191.

[6] Tao L, Chen H, Xu Y, et al. A new promoter recognition method based on features optimal selection[C]//Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on. IEEE, 2011: 1-4.

[7] Davuluri R V, Grosse I, Zhang M Q. Computational identification of promoters and first exons in the human genome [J]. Nat Genet, 2001, 29(4): 412-417.

[8] Narang V, Sung W K, Mittal A. Computational modeling of oligonucleotide positional densities for human promoter prediction [J]. Artif Intell Med, 2005, 35(1): 107-119.

[9] Bajic V B, Tan S L, Suzuki Y, et al. Promoter prediction analysis on the whole human genome[J]. Nat Biotechnol, 2004, 22(11): 1467-1473.

[10] Zeng J, Zhao X Y, Cao X Q, et al. SCS: Signal, context, and structure features for genome-wide human promoter recognition[J]. IEEE ACM Trans Comput Biol Bioinform, 2010, 7(3): 550-562.

[11] 李阳. 图形表示在 DNA 基因序列识别算法中的应用[D]. 长沙: 湖南大学, 2010.

[12] Abeel T, Saeys Y, Bonnet E, et al. Generic eukaryotic core promoter prediction using structural features of DNA[J]. Genome Res, 2008, 18(2): 310-323.

[13] 刘春卫, 罗健旭. 基于混合核函数的 PSO-SVM 分类算法[J].华东理工大学学报：自然科学版, 2014, 40(001): 96-101.

[14] 姜明辉, 袁绪川, 冯玉强. PSO-SVM 模型的构建与应用[J]. 哈尔滨工业大学学报, 2009, 41(2): 169-171.

[15] Zeng J, Zhu S, Yan H. Towards accurate human promoter recognition: a review of currently used sequence features and classification methods[J]. Brief Bioinform, 2009, 10(5): 498-508.

[16] Chang C C, Lin C J. LIBSVM: A library for support vector machine[J]. ACM Trans Intell Syst Technol, 2011, 2(3): 27.

[17] Faruto Y L. LIBSVM-Faruto Ultimate Version: a toolbox with implements for support vector machines based on libsvm[R/OL]. Software available at http://www. ilovematlab.cn, 2009.

[18] Santa L J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics[J]. Proc Natl Acad Sci, 1998, 95(4): 1460-1465.