Journal of Computer Applications
计算机应用,2015,35(10):2808-2812
ISSN 1001-9081
CODEN JYIIDU
2015-10-10
http://www.joca.cn

文章编号:1001-9081(2015)10-2808-05
doi:10.11772/j.issn.1001-9081.2015.10.2808

基于单核苷酸统计和支持向量机集成的人类基因启动子识别

徐文轩[1],张　莉[1,2*]

(1.苏州大学 计算机科学与技术学院系,江苏 苏州 215006;
2.江苏省计算机信息处理技术重点实验室(苏州大学),江苏 苏州 215006)

(*通信作者电子邮箱 zhangliml@suda.edu.cn)

摘　要:为高效地判别人类基因启动子,提出了一种基于单核苷酸统计和支持向量机集成的人类基因启动子识别算法。首先通过基因单核苷酸统计,从而将一个基因数据集分为C偏好和G偏好两个子集;然后分别对这两个子集提取DNA刚性特征、词频统计特征和CpG岛特征;最后采用多个支持向量机(SVM)集成的方式来学习这三种特征,并讨论了三种集成方式,包括单层SVM集成、双层SVM集成和级联SVM集成。实验结果表明所提算法能够提高人类基因启动子识别的敏感性和特异性,其中双层SVM集成的敏感性达到79.51%,且级联SVM集成的特异性高达84.58%。

关键词:CpG岛;DNA刚性;人类启动子识别;KL散度;单核苷酸统计;支持向量机

中图分类号:TP3-05;TP301.6　　**文献标志码**:A

Human promoter recognition based on single nucleotide statistics and support vector machine ensemble

XU Wenxuan[1], ZHANG Li[1,2*]

(1. *School of Computer Science and Technology, Soochow University, Suzhou Jiangsu* 215006, *China*;
2. *Provincial Key Laboratory for Computer Information Processing Technology (Soochow University), Suzhou Jiangsu* 215006, *China*)

Abstract: To efficiently discriminate the promoter in human genome, an algorithm for human promoter recognition based on single nucleotide statistics and Support Vector Machine (SVM) ensemble was proposed. Firstly, a gene dataset was divided into two subsets such as C-preferred and G-perferred subsets by using single nucleotide statistics. Secondly, DNA rigidity feature, word-based feature and CpG-island feature were extracted for each subset. Finally, these features were combined by using SVM ensemble learning. In addition, three ensemble ways were discussed, including single SVM ensemble, double-layer SVM ensemble and cascaded SVM ensemble. The experimental result shows that the proposed method can improve the sensitivity and specificity of human propoter recognition. Especially, the double-layer SVM ensemble can achieve the highest sensitivity of 79.51%, while the cascaded SVM ensemble has the highest specificity of 84.58%.

Key words: CpG-island; DNA rigidity; human promoter recognition; Kullback-Leibler divergence; nucleotide statistics; Support Vector Machine (SVM)

0　引言

在遗传学中,启动子(promoter)是DNA控制基因转录过程开始的一段区域,包括基因转录起始点和控制基因活性的基因[1]。启动子决定基因转录的方向、速度和精准度,因而启动子的识别对研究人类基因的表达调控具有十分重要的作用。自从1997年第一篇关于启动子识别算法[2]的评论文章发表以来,启动子识别技术就得到了快速发展,日渐成为十分重要的研究课题。

启动子识别的关键之一是如何提取更具识别性和分辨力的特征。信号、语义和结构特征是用来识别核心启动子区域的主要特征。CpG岛是一种典型的信号特征,已被广泛用于识别[3-5]。可以用来识别和预测启动子的语义特征[6]有N联体,KL散度(Kullback-Leibler divergence)被用来选择N联体并降低搜索空间大小。除了信号和语义特征,DNA三维结构特征[7]得到越来越多的关注,并且已被证明是一种有效的启动子识别特征[8]。相比信号和语义特征,DNA结构特征可以提供一些重要的补充信息。

启动子识别另一关键是选择合适分类模型。机器学习的一些方法已被用来识别启动子,比如:支持向量机(Support Vector Machine, SVM)[9]、隐马尔可夫模型[10-11]、相关向量机[12]、线性和二次判别式分析[13-14],以及神经网络[15-16]。SVM已被证明是一种有效的启动子识别算法[17],相比其他启动子识别中的机器学习算法,对于解决大量高维复杂的数据具有更好的性能,因此本文将SVM作为启动子识别的分类器。

收稿日期:2015-06-15;**修回日期**:2015-06-27。　**基金项目**:国家自然科学基金资助项目(61373093);国家级大学生创新创业训练计划项目(201410285032);江苏省自然科学基金资助项目(BK20140008, BK201222725);江苏省高校自然科学研究项目(13KJA520001);江苏省"青蓝工程"资助项目;苏州大学大学生课外学术科研基金资助项目(KY2014687B, KY2015544B, KY2015818B);苏州大学敬文书院"3I工程"项目(29)。

作者简介:徐文轩(1993-),男,江苏盐城人,CCF会员,主要研究方向:机器学习、模式识别;　张莉(1975-),女,江苏张家港人,教授,博士生导师,博士,CCF高级会员,主要研究方向:机器学习、模式识别。

徐文韬等[18]把 SVM 应用到启动子识别中,直接对碱基进行编码,没有提取更具分辨力的特征;智慧等[19]提出一种基于知识统计特征编码方法的双层 SVM 共识模型,没有考虑非启动子域;梅丽[17]提出基于两级 SVM 的启动子识别算法,考虑 CpG 岛和 KL 散度特征以及其他的非启动子域。

启动子发展的研究趋势是同时考虑启动子和其他的几种非启动子基因域,比如外显子(exon)、内含子(intron)和 3′UTR,并且同时考虑多种特征,但是目前还没有文献研究关于人类启动子的单核苷酸统计方案。本文将单核苷酸统计算法与多特征多支持向量机结合,同时考虑启动子与非启动子域,提出一种人类启动子识别算法。首先通过单核苷酸统计算法将序列分为具有不同性质的两个子集,即 G 偏好和 C 偏好集合,在不同的偏好子集上分别提取多种基因特征,包括 DNA 刚性特征、词频统计特征和 CpG 岛特征,这样有针对性地提取不同性质的基因特征具有更好的识别性与分辨力。最后对这些特征分别采用支持向量机集成的方式来学习,从而提高启动子的识别效率。

1　工作基础

1.1　特征提取

1.1.1　DNA 刚性

DNA 三维结构特征可以通过局部角参数(扭(twist)、卷(roll)和倾斜(tilt))以及平移参数(平移(shift)、滑动(slide)和上升(rise))来刻画。具有序列依赖性的 DNA 刚性(DNA rigidity)特征是一种属于 DNA 三维结构特征的重要物理属性[20]。所有的人类启动子中的 DNA 刚性参数值已被测量并应用于计算机启动子预测识别中[21]。

基于统计力学利用三核苷酸模型可以计算基因序列的刚性特征值[22]。本文以 6 碱基位长的序列为单位来计算序列中每个碱基位点的刚性特征值,其刚性值 f_l 可以通过叠加三个重叠的三核苷酸参数值得到,即

$$f_l = \sum_{i=1}^{4} t_i \tag{1}$$

其中:t_i 是在位置 i 的三核苷酸的刚性参数值。

根据文献[23]提供的换算表中 32 个三核苷酸刚性参数值,对于给定的序列长为 L 的 DNA 序列的刚性特征值向量为 $\boldsymbol{f} = [f_1, f_2, \cdots, f_{L-5}]^{\mathrm{T}}$。

1.1.2　基于 KL 散度的词频统计特征

基因可以看成一系列由字母 A(腺嘌呤:adenine)、C(胸腺嘧啶:cytosine)、G(鸟嘌呤:guanine)和 T(胞嘧啶:thymine)组成的文档集合,其中每个字母都代表一种核苷酸。N 个连续的核苷酸称为 N 联体(N-mer),共有 4^N 种 N 联体。N 联体的频率分布具有重要的生物学意义。由于统计出来的 N 联体频率值中有很多为 0,会大大增加分类器的计算复杂度,从而影响性能。KL 散度是一种统计学的度量方法,可以度量两个概率分布之间的差异。利用 KL 散度来选取有效的 N 联体,可以提取最有用的信息,简化计算过程。对于启动子和非启动子需要分开统计其频率。令 $\boldsymbol{f}_{pr}$ 为启动子中 N 联体出现的频率,$f_{pr}(i)$ 是其第 i 个分量,表示第 i 个 N 联体;$\boldsymbol{f}_{np}^{a}(a = 1,2,3)$ 为三种非启动子中 N 联体出现的频率序列,其中:$a = 1$ 代表外显子,$a = 2$ 代表内含子,$a = 3$ 代表 3′UTR。启动子和第 a 种非启动子的 KL 散度定义为:

$$D_a(\boldsymbol{f}_{pr}, \boldsymbol{f}_{np}^{a}) = \sum_{i=1}^{4^N} d_i^a \tag{2}$$

其中:

$$d_i^a = f_{pr}(i)\ln\frac{f_{pr}(i)}{f_{np}^{a}(i)};\ i = 1,2,\cdots,4^N \tag{3}$$

将 $d_i^a(i = 1,2,\cdots,4^N)$ 降序构成一个新的向量 $\boldsymbol{d}^a = [d_1^a, d_2^a, \cdots, d_{4K}^a]^{\mathrm{T}} \in \mathbf{R}^{4^N}$,并重新排列 $\boldsymbol{f}_{pr}$ 和 $\boldsymbol{f}_{np}^{a}$。为了获得 m_a 种最具识别性和分辨力的 N 联体,求解如下的最优化问题:

$$\min_{m_a} \quad \frac{\sum_{i=1}^{m_a} d_i^a}{D_a(\boldsymbol{f}_{pr}, \boldsymbol{f}_{np}^{a})} - \theta \tag{4}$$

$$\text{s.t.} \quad \frac{\sum_{i=1}^{m_a} d_i^a}{D_a(\boldsymbol{f}_{pr}, \boldsymbol{f}_{np}^{a})} \geqslant \theta;\ a = 1,2,3$$

其中:阈值 $\theta > 0$,在实验中设为 0.98。

求解优化问题(4)后,得到 $m_a(a = 1,2,3)$,就可以取前 m_a 个所对应 N 联体的频率构成区分启动子与第 a 种非启动子的显著特征;然后对每条基因都进行上述的显著特征提取,并保留这 m_a 种 N 联体。

1.1.3　CpG 岛

CpG 岛(CpG-island)是 DNA 上一段长度超过 200 bp 富含由磷酸二酯键相连的 C、G 碱基对的区域[21],其中 C + G 含量大于 50%,双核苷酸 CG 出现的次数与估计出现的次数之比(Obs/Exp)大于 60%。根据已知 DNA 数据统计显示,大约半数的哺乳动物启动子中含有 CpG 岛[24]。CpG 岛特征可以将这一比例提高至 72%[9]。

大量的研究表明,大约 70% 的人类基因启动子与 CpG 岛相关[25]。所以 CpG 岛特征可以作为识别人类启动子的重要特征。本文中采用两个重要的 CpG 岛特征,包括 C + G 的含量(GC_con)和双核苷酸 CG 的预测值与观测值之比(o/e):

$$GC\_con = \frac{n_{\mathrm{C}} + n_{\mathrm{G}}}{L} \tag{5}$$

$$o/e = \frac{n_{\mathrm{CG}} * L}{n_{\mathrm{C}} * n_{\mathrm{G}}} \tag{6}$$

其中:L 是序列长度,n_{C}、n_{G} 和 n_{CG} 分别是 C、G 核苷酸和 CG 双核苷酸在序列中的数量。

1.2　SVM

SVM 是一种基于统计学习理论的通用学习方法[26],SVM 可以实现结构最小化原则来提高学习机的泛化能力。在 SVM 中,核函数将原始样本映射到一个高维特征空间中,使得原始样本可以以非线性的方式可分。

给定一组训练样本$(\boldsymbol{x}_i, y_i), i = 1,2,\cdots,n$,其中:$\boldsymbol{x}_i \in \mathbf{R}^d$,$d$ 是特征的个数,$y_i \in \{-1, +1\}$ 是样本 $\boldsymbol{x}_i$ 的类标签,n 是样本个数。SVM 的对偶规划问题可以表示为:

$$\max \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{7}$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0;\ 0 \leqslant \alpha_i \leqslant C$$

其中:$C > 0$ 是惩罚因子,$k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 是核函数,α_i 是拉格朗日乘子。当 $0 \leqslant \alpha_i \leqslant C$ 时,对应的 $\boldsymbol{x}_i$ 称为无边界的支持向量。如果 $\alpha_i = C$,则对应的 $\boldsymbol{x}_i$ 为有界支持向量。二次规划(7)求解后,可以生成分类判别函数:

$$g(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x}) + b\right) \quad (8)$$

其中:sgn(·)表示符号函数;b 是阈值,可以由 $b = y_{sv} - \sum_{i=1}^{n}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x}_{sv})$ 算出;$(\boldsymbol{x}_{sv}, y_{sv})$ 为无边界支持向量。

2 基于单核苷酸统计和支持向量机集成的方法

本文提出的基于单核苷酸统计和支持向量机集成的方法如图 1 所示,包括三个步骤:单核苷酸统计、多种特征提取和 SVM 集成。本章主要介绍单核苷酸统计和 SVM 集成模型。

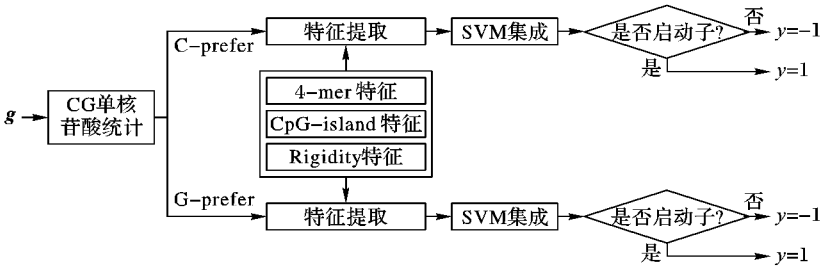

图 1 基于单核苷酸统计和支持向量机集成框架

2.1 单核苷酸统计

因为基因的复杂高维并包含大量的信息,所以通常提取的特征很难包含全部的生物学意义。同时,对于不同的 DNA 片段组成成分也不尽相同,在同一片段的不同位置的核苷酸组成也不同,因此核苷酸 A、G、C、T 的组成成分是非常重要的特征。本文称 C 含量大于 G 含量的序列为具有 C 偏好(C-prefer)序列;反之为 G 偏好(G-prefer)序列。由此可以把基因序列按照 C 和 G 的含量划分为两个子集。

单核苷酸统计是指用 C 和 G 的含量比来实现对序列 C 偏好或 G 偏好的统计分类。令 C 偏好 DNA 序列集合为 X_C 且 $|X_C| = n_1$,G 偏好 DNA 序列集合为 X_G 且 $|X_G| = n_2$,且 $n_1 + n_2 = n$。对这两个子集 X_C 和 X_G,分别进行多种特征提取。本文将在实验中对单核苷酸统计分类的有效性进行验证。

2.2 SVM 集成模型

SVM 已被证明是启动子识别的有效算法[27]。启动子识别中单种特征的识别效果不如多种特征,因此可以用集成处理多种特征。将多个 SVM 简单集成形成单层 SVM 集成。单层 SVM 集成虽未广泛用于启动子识别,却常用于其他领域。基于 SVM 两层集成的识别方法先后被提出:文献[19]利用双层 SVM 建立共识模型;文献[17]提出基于级联 SVM 的启动子识别算法,将 CpG 岛和 5 联体特征级联使用。

在基于单核苷酸统计和支持向量机集成框架中,SVM 集成部分可以分别考虑上面提到的三种集成形式,如图 2~4 所示。图 2 为单层 SVM 集成模型,共采用 5 个 SVM。CpG 岛和 DNA 刚性特征各训练 1 个 SVM;由于 4-mer 特征提取的特殊性,需要对不同的非启动子分别进行,因而需要三个分类器来处理三种非启动子和启动子的识别。直接采用多数投票的规则对 5 个 SVM 的输出进行融合。投票分为两次,首先对 4-mer特征的 3 个输出进行投票,对测试样本进行标记,+1 是启动子,-1 则是非启动子;然后对其投票后的结果结合其他两种特征的分类器的结果再次投票;根据最后的投票来判断样本是否为启动子。

图 3 为双层 SVM 集成模型,共采用了 6 个 SVM,第一层有 5 个分类器,第二层只有 1 个分类器。第一层 CpG 岛和 DNA 刚性特征各训练 1 个 SVM,4-mer 特征用 3 个 SVM。将 5 个 SVM 的输出作为新特征,重新训练第二层 SVM,即输入为 5 维。根据第二层输出对测试样本进行判别,若输出是 +1,则是启动子;否则是非启动子。

图 4 为级联 SVM 集成模型,共采用 4 个 SVM,第一层为单个分类器,第二层为 3 个分类器。文献[17]模型只有两种特征,缺乏灵活性。由于借鉴其模型,本文因此忽略 DNA 刚性。第一层是通过 CpG 岛进行识别。第二层根据 5 联体特征,将第一层判断为非启动子的序列作进一步识别。最终结果由 3 个 SVM 投票得出。

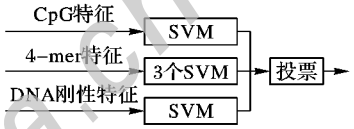

图 2 单层 SVM 集成模型

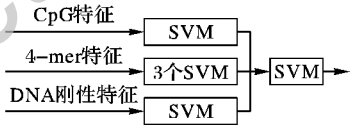

图 3 双层 SVM 集成模型

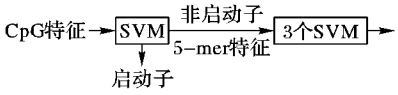

图 4 级联 SVM 集成模型

3 实验结果分析

本实验是为了验证本文算法框架的性能。本文算法的第一步是通过对基因序列集合进行单核苷酸统计,把训练集合分为 C-prefer 和 G-prefer 两个子集。因此通过单核苷酸统计的有效性评估,来论证本文算法框架的有效性和可行性。对于 SVM 的实现,本文运用了 Chih-Jen Lin 教授编写的 libsvm-2.89 工具箱(http://www.csie.ntu.edu.tw/~cjlin),选择径向基作为核函数,核函数为 $k(\boldsymbol{x}_i,\boldsymbol{x}) = \exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}\|^2)$ 中,核参数 $\gamma > 0$。本文采用十折交叉验证来对参数 C 和 γ 进行优化。

3.1 数据集

本文识别由 DBTSS[23] 数据库定义的转录起始点附近[-200,+50]位置的启动子与其他基因组区域相同长度的非启动子序列。对于基于统计模式识别的识别算法的实验验证,需要大量的已经准确注释的启动子非启动子序列。本文选择 DBTSS 数据库中的 30 964 条启动子序列作为训练集测试集,因为 DBTSS 中的基因序列兼具良好的覆盖性和良好的质量。本文在 EID 数据库中随机选择长度为 251 bp 的 10 000 条外显子和 10 000 条内含子以及在 UTRdb 数据库中随机选择 10 000 条 251 bp 长度的 3′UTR 序列构成非启动子数据集。

10 次分别从启动子、外显子、内含子和 3′UTR 数据集中随机选取8 000条样本序列,其中4 000 条作为训练,4 000 条作为测试集。测试集、训练集中启动子、外显子、内含子和 3′UTR 序列样本的比例为1∶1∶1∶1。启动子为 +1 类,非启动子为 -1 类,正负样本集不均衡。

3.2　评价方法

本文使用由 Bajic[28] 提出的评价准则敏感性 S_n、特异性 S_p 和平均条件概率 ACP 来评估本文算法,它们的定义如下:

$$S_n = \frac{TP}{TP + FN} \tag{9}$$

$$S_p = \frac{TN}{TN + FP} \tag{10}$$

$$ACP = \frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) \tag{11}$$

其中:TP 代表正确识别正集的样本数,TN 代表正确识别负集样本数,FP 代表错误识别正集的样本数,FN 代表错误识别负集的样本数。

3.3　单个 SVM 模型

本节实验通过把单核苷酸统计运用到单种特征上来验证单核苷酸统计的有效性。分类器采用单个 SVM 模型。1.1 节给出了三种关于启动子的特征:CpG 岛、N 联体和刚性特征,这些特征分别用于训练 SVM。其次进行单核苷酸统计,重新提取其 CpG 岛、N 联体和刚性特征,分别再训练 SVM。这里令 $N=4$。表 1 显示了单个 SVM 采用不同特征的分类性能,其中"+CG"表示进行了单核苷酸统计。实验结果显示单核苷酸统计的融入提高了分类器对于单种特征集的分类性能。融入了单核苷酸统计的分类器性能提高。在 4-mer 特征,SVM + CG 把敏感性提高 4.42%,特异性提高 1.39%,且平均条件概率提高2.96%。在 CpG 岛上的性能提高幅度不是很大,敏感性提高了 3.53%,特异性只提高了 0.27%,而平均条件概率提高了1.29%。在刚性特征上,敏感性得到了很大幅度的提高,达到 8.89%;特异性提高了 2.55%;平均条件概率提高了 5.57%。对这三个特征来说,刚性特征在融入了单核苷酸统计后,性能得到较大幅度的提高,N-联体特征次之。

表 1　单核苷酸统计分类在三个单种特征集上的性能　%

评价准则	4-mer		CpG 岛		Rigidity	
	SVM	SVM + CG	SVM	SVM + CG	SVM	SVM + CG
S_n	70.83	75.25	68.11	71.64	52.5	61.39
S_p	81.27	82.66	81.22	81.49	79.51	82.06
ACP	70.38	73.34	69.58	70.87	63.18	68.75

3.4　SVM 集成模型

本节实验通过把单核苷酸统计运用到 SVM 集成上来验证单核苷酸统计的有效性。分类器采用 2.2 节给出的三种 SVM 集成模型:单层 SVM 集成、双层 SVM 集成和级联 SVM 集成。

表 2 显示了三种集成 SVM 的分类性能,其中"+CG"表示进行了单核苷酸统计。实验结果表明结合了单核苷酸统计的 SVM 集成能够进一步提高分类的性能。不管是在敏感性和特异性还是平均条件概率性能上,进行了单核苷酸统计的集成比不进行具有更好的效果;双层 SVM + CG 的敏感性和平均条件概率是所有方法中最高的,而级联 SVM + CG 的特异性最高;单层 SVM + CG 的性能比较稳定,在敏感性、特异性和平均条件概率上均排在所有方法的第二位。

表 2　SVM 集成模型性能比较　%

集成模型	S_n	S_p	ACP
级联 SVM	65.25	82.94	70.89
级联 SVM + CG	65.27	**84.58**	72.79
双层 SVM	72.36	82.09	71.84
双层 SVM + CG	**79.51**	83.65	**75.73**
单层 SVM	71.89	83.18	73.05
单层 SVM + CG	76.86	84.23	75.72

4　结语

本文同时考虑启动子和 DNA 中的非启动子域,提出了一种人类启动子识别框架。本文的框架特色在于结合了单苷酸统计和 SVM 集成进行启动子识别。为了评估本文算法的性能,实验部分对比分析了结合了单苷酸统计的 SVM 分类器以及 SVM 集成分类器。实验结果表明,融合了单苷酸统计后,不管是单个 SVM 还是 SVM 集成均在性能上有不同程度的提升。当然,也验证了 SVM 集成比单个 SVM 的效果要好。此外,单层 SVM 集成的性能比较稳定。

因为基因数据复杂高维的特性,以及实验样本有限,特征的种类有限等局限性,因此,在接下来的研究中,将考虑更多具有代表性的基因序列和更具分辨力的特征提取方式。

参考文献:

[1] BAJIC V B, CHONG A, SEAH S H, *et al.* An intelligent system for vertebrate promoter recognition [J]. IEEE Intelligent Systems, 2002, 17(4): 64 - 70.

[2] FICKETT J W, HATZIGEORGIOU A G. Eukaryotic promoter recognition [J]. Genome Research, 1997, 11(5): 861 - 878.

[3] UMESH P, DUBEY J K, KARTHIKA R V, *et al.* A novel sequence and context based method for promoter recognition [J]. Bioinformation, 2014, 10(4): 175 - 179.

[4] ZENG J, ZHAO X, CAO X, *et al.* SCS: signal, context, and structure features for genome-wide human promoter recognition [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(3): 550 - 562.

[5] DENG J, LIANG H, ZHANG R, *et al.* Methylated CpG site count of dapper homolog 1 (DACT1) promoter prediction the poor survival of gastric cancer [J]. American Journal of Cancer Research, 2014, 4(5): 518 - 527.

[6] HUANG W L, TUNG C W, LIAW C, *et al.* Rule-based knowledge acquisition method for promoter prediction in human and Drosophila species [J]. Scientific World Journal, 2014, 2014(2014): 1 - 14.

[7] FUJII S, KONO H, TAKENAKA S, *et al.* Sequence-dependent DNA deformability studied using molecular dynamics simulations [J]. Nucleic Acids Research, 2007, 35(18): 6063 - 6074.

[8] GAN Y, GUAN J, ZHOU S. A comparison study on feature selection of DNA structural properties for promoter prediction [J]. BMC Bioinformatics, 2012, 13:4.

[9] ANWAR F, BAKER S M, JABID T, *et al.* Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach [J]. BMC Bioinformatics, 2008, 9(1): 414 - 418.

[10] WU J, XIE J. Hidden Markov model and its applications in motif findings [M]// Statistical Methods in Molecular Biology. New York: Humana Press, 2010: 405 - 416.

[11] ZHAO X, ZHANG J, CHEN Y, *et al.* Promoter recognition based on the maximum entropy hidden Markov model[J]. Computers in

Biology & Medicine, 2014, 51(15): 73 - 81.
[12] LI Y, LEE K, WALSH S, *et al.* Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a relevance vector machine [J]. Genome Research, 2008, 16(3): 414 - 427.
[13] LIU G, LIU J, CUI X, *et al.* Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae [J]. Journal of Theoretical Biology, 2012, 293(1): 49 - 54.
[14] LU J, LUO L. Prediction for human transcription start site using diversity measure with quadratic discriminant [J]. Bioinformation, 2008, 2(7): 316 - 21.
[15] WANG J, UNGAR L H, TSENG H, *et al.* MetaProm: a neural network based meta-predictor for alternative human promoter prediction [J]. BMC Genomics, 2007, 8(1): 374 - 13.
[16] BURDEN S, LIN Y, ZHANG R. Improving promoter prediction for the NNPP2.2 algorithm: a case study using escherichia coli DNA sequences [J]. Bioinformatics, 2005, 21(5): 601 - 607.
[17] MEI L. Human promoter recognition algorithm[D]. Dalian: Liaoning Normal University, 2010. (梅丽. 人类启动子识别算法研究[D]. 大连: 辽宁师范大学, 2010.)
[18] XU W, YE Z, YU X. The human promoter recognition based on SVMs[J]. Auhui Agricultural Science Bulletin, 2006, 12(13): 64 - 66. (徐文韬, 叶子弘, 俞晓平. 基于支持向量机(SVMs)的人类核心启动子的识别[J]. 安徽农学通报, 2006, 12(13): 64 - 66.)
[19] ZHI H, LI T. Applying novel knowledge-based encoding methods and dual SVM to human Pol Ⅱ promoter recognition [J]. Journal of Harbin Medical University, 2012, 46(1): 23 - 26. (智慧, 李通化. 应用新的基于知识编码方法及双层SVM识别人类Pol Ⅱ启动子[J]. 哈尔滨医科大学学报, 2012, 46(1): 23 - 26.)
[20] GODDARD N L, BONNET G, KRICHEVSKY O. Sequence dependent rigidity of single stranded DNA [J]. Physical Review Letters, 2000, 85(11): 2400 - 3.
[21] 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.
[22] BAJIC V B. Comparing the success of different prediction programs in sequence analysis: a review [J]. Brief Bioinform, 2000, 1(3): 214 - 228.
[23] YAMASHITA R, SUZUKI Y, WAKAGURI H, *et al.* DBTSS: database of human transcription start sites, progress report 2006 [J]. Nucleic Acids Research, 2006, 34(Database issue): 86 - 89.
[24] LI W, KOU Q, WEI L, *et al.* Plant promoter recognition based on analysis of base bias and SVM [J]. Journal of Liaoning Normal University: Natural Science, 2012, 35(2): 183 - 187. (李文举, 寇秋波, 韦丽华, 等. 基于碱基偏好分析和SVM的植物启动子识别[J]. 辽宁师范大学学报: 自然科学版, 2012, 35(2): 183 - 187)
[25] SAXONOV S, BERG P, BRUTLAG D L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters [J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(5): 1412 - 1417.
[26] VAPNIK V, CORTES C. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273 - 297.
[27] GANGAL R, SHARMA P. Human pol II promoter prediction: time series descriptors and machine learning [J]. Nucleic Acids Research, 2005, 33(4): 1333 - 6.
[28] BAJIC V B. Comparing the success of different prediction programs in sequence analysis: a review [J]. Brief Bioinform, 2000, 1(3): 214 - 228.

(上接第2807页)

样本,尤其是在实验1中,能够有效找到不一致样本集中的不一致样本,由此说明本文的不一致性决策算法不仅是可行的,而且还是有效的,这对提高算法的抗噪声能力、预防临床医生对疾病的误判有着积极的意义。

参考文献:

[1] CENTERA M M, LOWAST A J, LORTET-TIEULENTB J, *et al.* Intemational variation in prostate cancer incidence and mortality rates[J]. European Urology, 2012, 61(6): 1079 - 1092.
[2] ZHANG J. Study on the characterristics of three-dimensional magnetic resonance spectroscopy of prostatic diseases[D]. Wuhan: Wuhan University, 2012. (张建伟. 前列腺疾病磁共振波谱成像(MRSI)特点的研究[D]. 武汉: 武汉大学, 2012.)
[3] NIAF E, ROUVIERE O, MEGE-LECHEVALLIER F, *et al.* Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI[J]. Physics in Medicine and Biology, 2012, 57(12): 3833.
[4] ZHOU T, LU H, CHEN Z, *et al.* Prostate tumor recognition based on two-stage ensemble SVM[J]. Optics and Precision Engineering, 2013, 21(8): 2138 - 2145. (周涛, 陆惠玲, 陈志强, 等. 基于两阶段集成支持向量机的前列腺肿瘤识别研究[J]. 光学精密工程, 2013, 21(8): 2138 - 2145.)
[5] MENG Z, HUANG B. Comparative study of attribute reduction in inconsistent incomplete decision system[J]. Control and Decision, 2011, 26(6): 867 - 872. (蒙祖强, 黄柏雄. 不一致不完备决策系统中属性约简的比较研究[J]. 控制与决策, 2011, 26(6): 867 - 872.)
[6] LI F, LIU Q, YE M, *et al.* Approaches to knowledge reductions in inconsistent decision tables[J]. Control and Decision, 2006, 21(8): 857 - 862. (李凡, 刘启和, 叶茂, 等, 不一致决策表的知识约简方法研究[J]. 控制与决策, 2006, 21(8): 857 - 862.)
[7] QIAN W, YANG B, XU Z, *et al.* Rule extraction algorithm based on discernibility matrix in inconsistent decision table[J]. Computer Science, 2013, 40(6): 215 - 218. (钱文彬, 杨炳儒, 徐章艳, 等. 基于差别矩阵的不一致决策表规则获取算法[J]. 计算机科学, 2013, 40(6): 215 - 218.)
[8] YIN L, GUI S, YANG C, *et al.* Core set analysis in inconsistent decision tables[J]. Information Sciences, 2013, 241(20): 138 - 147.
[9] HUANG G, LIU Y. Calculation and translation for core of information entropy and algebra reductions in inconsistent decision table [J]. Chinese Journal of Computer Systems, 2008, 29(2): 308 - 312. (黄国顺, 刘云生. 不一致决策表信息熵约简与代数约简的核计算与转化[J]. 小型微型计算机系统, 2008, 29(2): 308 - 312.)
[10] HUANG G, LIU Y. Inconsistency analysis and translation of different types of attribute reduction for inconsistent decision tables[J]. Chinese Journal of Computer Systems, 2008, 29(4): 703 - 708. (黄国顺, 刘云生. 不一致决策表各种属性约简的不一致性分析与转化[J]. 小型微型计算机系统, 2008, 29(4): 703 - 708.)
[11] CHEN D, ZHANG X, LI W. On measurements of covering rough sets based on granules and evidence theory[J]. Information Sciences, 2015, 317(1): 329 - 348.