

## 基于模糊支持向量机的剪接位点识别

孙波, 李小霞, 李铖果

(西南科技大学 信息工程学院, 四川 绵阳 621010)

(sunboxinan@163.com)

**摘要:**为了提高模糊支持向量机(FSVM)对剪接位点的识别精度,提出一种计算样本隶属度的新方法。将样本到两聚类中心的距离比值作为样本的初始隶属度,采用K近邻(KNN)方法计算样本的紧密度,最后将初始隶属度与紧密度的乘积作为样本的最终隶属度,这样既提高了支持向量的隶属度,又降低了噪声样本的隶属度。将此方法应用到剪接位点的识别中,对组成性5'和3'剪接位点的识别精度分别达到了94.65%和88.79%,与经典支持向量机相比,3'剪接位点的识别精度提高了7.94%。

**关键词:**模糊支持向量机;隶属度;紧密度;剪接位点识别;选择性剪接

**中图分类号:** TP18; TP391.4 **文献标志码:** A

## Recognition of splice sites based on fuzzy support vector machine

SUN Bo, LI Xiao-xia, LI Cheng-guo

(School of Information Engineering, Southwest University of Science and Technology, Mianyang Sichuan 621010, China)

**Abstract:** In order to improve the splice site recognition accuracy of Fuzzy Support Vector Machine (FSVM), a new method for computing the membership degree of sample was proposed. The initial membership was defined as the distance ratio of the sample to the two cluster centers of positive and negative samples, K-Nearest Neighbor (KNN) was adopted to compute the tightness of the samples, and the multiplication of the tightness and the initial membership degree was used as the ultimate membership. It will not only improve the membership degree of support vector, but also reduce the membership degree of noise sample. This method was applied to recognize the splice site, and the experimental results show that the recognition accuracy of constitutive 5' and 3' splice site reaches 94.65% and 88.79% respectively. Compared with the classical support vector machine, the recognition accuracy of constitutive 3' splice site increases by 7.94%.

**Key words:** Fuzzy Support Vector Machine (FSVM); membership degree; tightness; splice site recognition; alternative splice

### 0 引言

早在1977年研究人员就发现真核生物蛋白质编码基因被“间隔序列”所打断,形成外显子—内含子镶嵌排列的结构,外显子和内含子的边界称为剪接位点。在转录过程中,内含子在剪接位点处被切除,外显子重新连接起来组成一个成熟的信使RNA(Message RNA, mRNA)。许多基因遵循可变剪接的模式,一个基因可以被剪接成多个成熟的mRNA序列。可变剪接增加了蛋白质组的多样性,并且是很多组织特异性发育过程的重要调控机制<sup>[1]</sup>。可变剪接可以概括为四种基本类型:可变5'或者可变3'剪接位点型(即外显子缩短或者延长型)、盒式外显子、互斥外显子和内含子保留。其中可变5'或者可变3'剪接位点型所占的比例最大<sup>[2]</sup>,本文预测的可变剪接位点就是取自这两种类型。

近年来支持向量机(Support Vector Machine, SVM)在模式识别和生物序列分析中得到了广泛的应用,但对噪声样本(或野值)却非常敏感,导致获得的分类面并不是真正的最优分类面。针对这一情况,Lin等学者<sup>[3-4]</sup>提出了模糊支持向量机(Fuzzy SVM, FSVM)理论,将模糊技术应用于支持向量机中,试图减小或消除它们对最优决策面的影响<sup>[5]</sup>。在FSVM

理论中,隶属度函数的设计是一个关键问题。目前隶属度函数的构造方法有很多,张翔等人<sup>[6]</sup>通过构造最小包围球,使位于最小包围球外部的噪声样本隶属度相对较小,从而减少噪声样本对分类器的影响。一般模糊隶属度是在原始空间中根据样本点的相互距离及到类中心的距离创建的,李雷等人<sup>[7]</sup>考虑样本之间的紧密程度,在特征空间中利用混合核函数建立了一种新的模糊隶属度。唐浩等人<sup>[5]</sup>提出结合K近邻法和样本到类中心距离的隶属度函数,计算了每个样本与其他样本之间的紧密程度,赋予紧密度低的样本较小的隶属度从而将噪声样本从支持向量中区分出来。这些方法都有一定的效果,但是同时都在一定程度上降低了支持向量的隶属度,将支持向量视为不重要的样本。本文通过设计一种新的模糊隶属度函数,在提高了支持向量隶属度的同时又降低了噪声样本的隶属度。

### 1 材料与方法

#### 1.1 模糊支持向量机

给定模糊训练样本集:

$$S = \{(\mathbf{x}_1, y_1, s_1), (\mathbf{x}_2, y_2, s_2), \dots, (\mathbf{x}_l, y_l, s_l)\} \quad (1)$$

其中: $\mathbf{x}_i$ 为 $n$ 维实向量, $y_i \in \{-1, 1\}$ 为相应的类别标签, $0 \leq$

收稿日期:2010-09-16;修回日期:2010-11-18。

**作者简介:**孙波(1985-),男,安徽阜阳人,硕士研究生,主要研究方向:真核生物基因识别、机器学习算法、人工智能算法; 李小霞(1976-),女,四川安岳人,副教授,博士,主要研究方向:生物信息学、生物特征识别、生物医学光子学、光谱检测与仪器; 李铖果(1986-),女,安徽亳州人,硕士研究生,主要研究方向:智能交通监控。

$s_i \leq 1$  为模糊隶属度。

假设  $\Phi$  是将原始空间映射到多维特征空间的非线性映射。在特征空间中利用结构风险最小化原理和分类间隔最大化思想,求解最优分类超平面问题可转化为下面的最优化问题:

$$\min \Phi(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l s_i \xi_i \quad (2)$$

$$\text{s.t. } y_i(\mathbf{w} \mathbf{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \dots, l$$

其中:  $\mathbf{w}, b$  分别是分类超平面的权值和偏值;  $\xi_i$  是非负松弛变量;  $C > 0$  是自定义的惩罚系数,保持分类最大间隔与分类误差之间的平衡;模糊隶属度  $s_i$  表示相应点  $\mathbf{x}_i$  的重要性,较小的  $s_i$  可以减小  $\xi_i$  在式中的影响,以致将相应的  $\mathbf{x}_i$  看做不重要的样本。

为求解上述约束最优化问题,引入 Lagrange 函数:

$$L = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l s_i \xi_i - \sum_{i=1}^l a_i [y_i(\mathbf{w} \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^l \beta_i \xi_i \quad (3)$$

其中:  $a_i \geq 0, \beta_i \geq 0$  是 Lagrange 系数。将  $L$  分别对  $\mathbf{w}, b, \xi_i$  求偏微分并令其等于 0:

$$\frac{\partial L(\mathbf{w}, b, \xi, \mathbf{a}, \beta)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^l a_i y_i \mathbf{x}_i = 0 \quad (4)$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \mathbf{a}, \beta)}{\partial b} = - \sum_{i=1}^l a_i y_i = 0 \quad (5)$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \mathbf{a}, \beta)}{\partial \xi_i} = s_i C - a_i - \beta_i = 0 \quad (6)$$

将式(4)~(6)代入(3),把最优化问题转化为等价的对偶规划问题(7):

$$\max Q(\mathbf{a}) = \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l a_i a_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (7)$$

$$\text{s.t. } \sum_{i=1}^l a_i y_i = 0 \\ 0 \leq a_i \leq s_i C, i = 1, 2, \dots, l$$

其中  $K(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j))$  是核函数,并且满足 Mercer 定理。

如果  $a_i > 0$ ,相应的点  $\mathbf{x}_i$  是支持向量,这里有两种支持向量:一种满足  $0 < a_i < s_i C$  的支持向量位于分类面附近,一种满足  $a_i = s_i C$  的支持向量  $\mathbf{x}_i$  为错误分类样本。

相应的最优决策函数为:

$$f(\mathbf{x}) = \text{Sgn}(\mathbf{w}^* \mathbf{x} + b^*) = \text{Sgn} \left( \sum y_i a_i^* K(\mathbf{x}_i, \mathbf{x}) + b^* \right) \quad (8)$$

其中:  $\mathbf{a}^*$  是式(8)的最优解,  $\mathbf{w}^*, b^*$  分别为相应的权值和偏值。

## 1.2 模糊隶属度

由于 SVM 的最优分类面是由靠近类边缘的支持向量所决定的,而噪声样本通常也在这一区域,传统的基于样本到类中心的距离的方法定义的模糊隶属度并不能有效地区分支持向量与噪声点<sup>[8]</sup>,并且这种方法不可避免地降低了支持向量的隶属度,从而降低了 FSVM 算法的分类精度。

在图1中,  $\mathbf{x}$  点到类中心的距离相等。若单纯按照距离计算隶属度,这两点的隶属度应相同,对最终决策的影响也应相同。但是从各样本排列的紧密程度来看,图(a)中的  $\mathbf{x}$  点可能是支持向量,而图(b)中的  $\mathbf{x}$  点更可能为噪声点。因此仅用点到类中心的距离来定义模糊隶属度存在着不足之处。

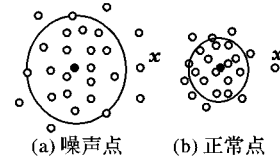


图1 噪声点和正常点紧密程度示意图

本文对基于距离的隶属度函数进行了改进,提高可能成为支持向量的样本的隶属度,同时结合 K 近邻法的思想计算样本点的紧密度,以降低噪声点的隶属度。

设  $S_+, S_-$  分别代表正负训练样本集合:

$$\begin{cases} S_+ = \{\mathbf{x}_i | \mathbf{x}_i \in S, y_i = +1\} \\ S_- = \{\mathbf{x}_i | \mathbf{x}_i \in S, y_i = -1\} \end{cases} \quad (9)$$

定义  $\mathbf{O}_+, \mathbf{O}_-$  分别是正负训练样本的中心:

$$\begin{cases} \mathbf{O}_+ = \frac{1}{n_+} \sum_{\mathbf{x}_i \in S_+} \mathbf{x}_i \\ \mathbf{O}_- = \frac{1}{n_-} \sum_{\mathbf{x}_i \in S_-} \mathbf{x}_i \end{cases} \quad (10)$$

其中:  $n_+$  是正训练样本的个数,  $n_-$  是负训练样本的个数。

定义正样本到正负类样本中心的距离分别为  $d_{i+}^+, d_{i-}^+$ :

$$\begin{cases} d_{i+}^+ = \|\mathbf{x}_i - \mathbf{O}_+\|, \mathbf{x}_i \in S_+ \\ d_{i-}^+ = \|\mathbf{x}_i - \mathbf{O}_-\|, \mathbf{x}_i \in S_+ \end{cases} \quad (11)$$

同样定义负样本到正负类样本中心的距离分别为  $d_{i+}^-, d_{i-}^-$ :

$$\begin{cases} d_{i+}^- = \|\mathbf{x}_i - \mathbf{O}_+\|, \mathbf{x}_i \in S_- \\ d_{i-}^- = \|\mathbf{x}_i - \mathbf{O}_-\|, \mathbf{x}_i \in S_- \end{cases} \quad (12)$$

初步的模糊隶属度按下式构造:

$$s_i^+ = \frac{d_{i-}^+}{d_{i-}^+ + d_{i-}^-}, \quad s_i^- = \frac{d_{i+}^-}{d_{i+}^- + d_{i+}^+} \quad (13)$$

这种方法构造的隶属度使一个离本类中心距离远、离异类中心距离近的样本获得了比靠近类中心的样本更大的隶属度,从而使可能成为支持向量的样本对最优分类面的构造有更大的贡献。但是这种方法使靠近分类面的噪声样本也获得了较大的隶属度,对分类器的性能带来负面影响,为此本文采用基于 K 近邻的方法降低噪声样本的隶属度。

定义正样本数据点  $\mathbf{x}_i$  与其他样本之间的距离为  $d_{ij}^+$ , 并将其按从小到大的顺序排列:

$$\begin{cases} d_{ij}^+ = \|\mathbf{x}_i - \mathbf{x}_j\|; \mathbf{x}_i, \mathbf{x}_j \in S_+, i \neq j \\ d_{i1}^+ \leq d_{i2}^+ \leq \dots \leq d_{in_+}^+ \end{cases} \quad (14)$$

取前  $k$  个最近的距离,设:

$$b_i = 1 / \left( \frac{1}{k} \sum_{j=1}^k d_{ij}^+ \right) \quad (15)$$

于是样本  $\mathbf{x}_i$  的紧密度定义式为:

$$u_i^+ = b_i / B \quad (16)$$

其中  $B = \max(b_1, b_2, \dots, b_{n_+})$ 。

同样的办法可以计算出负样本  $\mathbf{x}_i$  的紧密度  $u_i^-$ 。

这样计算的样本紧密度最大为 1, 由于噪声样本与其他正常样本之间的距离较远, 因此其对应的紧密度比正常样本要小。将样本紧密度式(16)与式(13)计算出的初步的隶属度相乘可以有效地降低噪声样本的隶属度, 削弱其对分类器的影响。最终的样本隶属度为:

$$s_i = \begin{cases} s_i^+ u_i^+, & \mathbf{x}_i \in S_+ \\ s_i^- u_i^-, & \mathbf{x}_i \in S_- \end{cases} \quad (17)$$

## 1.3 数据集

实验材料下载于 ASD (Alternative Splicing Database, <http://www.ebi.ac.uk/asd/altsplice/>) 数据库。这个数据库包含了人类和小鼠的基因剪接模式、可变剪接事件和相关注释

数据。3'和5'选择性剪接事件分别被注释为(II-3P)(II-5P),在剪接事件文件中提取在3'端和在5'端发生选择性剪接,以及在3'和5'两端同时发生选择性剪接的内含子,然后去除非标准内含子及相互重叠的部分,最后选择出了选择性剪接位点正样本集。这些内含子都是以双核苷酸GT开始,以AG结尾,GT和AG都称为剪接位点。图2表示了5'和3'选择性剪接类型。

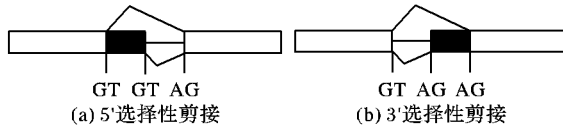


图2 选择性剪接示意图

从标准内含子中去除有选择性剪接位点的内含子,剩余的作为组成性内含子集的正样本。

为了产生负样本集,首先分别计算了5'和3'选择性剪接位点之间的距离,即图2(a)中两个GT选择性位点和图2(b)中两个AG选择性位点之间的距离,图3给出了它们的直方图。

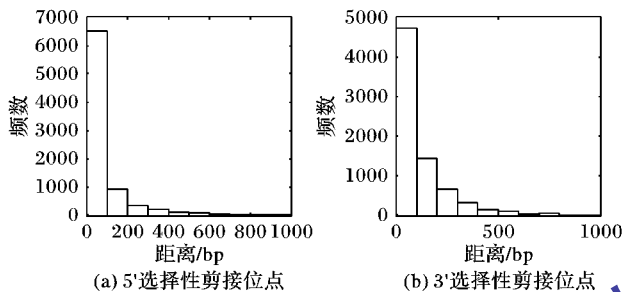


图3 相邻的选择性剪接位点距离直方图

由于DNA是双链互补配对的螺旋结构,在每条链上的一个碱基都有另一条链上的一个碱基与它互补配对,所以在表示DNA序列长度时,无论是单链还是双链,都以碱基对(base pair, bp)作为单位。从图3可以看出,同一个内含子的选择性剪接位点之间的距离大多数都不超过200 bp。3'选择性剪接位点之间的距离有84%小于200 bp,5'选择性剪接位点之间的距离有71%小于200 bp。这表明选择性剪接位点两侧200 bp范围内的序列对选择性剪接事件的发生有重要的影响,确定选择性剪接的保守信息就在此范围之内。

负样本集选自组成性和可变剪接位点的侧翼序列,为了尽可能保持正负样本集的差异,选择组成性和可变剪接位点GT上游(AG下游)200 bp外最近的非剪接位点GT(AG)二联体作为剪接位点预测的负集。最后提取到的正负样本集大小列于表1。

表1 剪接位点正负样本集大小

剪接位点	正样本集	负样本集
选择性3'	4812	4368
选择性5'	4351	4112
组成性3'	16412	15913
组成性5'	16412	15913

对每个剪接位点上下游分别选择长度为L bp的侧翼序列,每条剪接位点序列长度为2L,剪接位点在第L+1和L+2位,如图4。

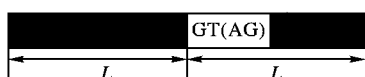


图4 剪接位点样本序列示意图

为了输入支持向量机进行预测,必须将样本序列转化成

数字形式。采取文献[9]的方法,分别对A,T,G,C进行数字编码:A为0001,T为0010,G为0100,C为1000。

#### 1.4 评估指标

采用敏感性(Sensitivity, Sn)、特异性(Specificity, Sp)、总精度(Total Accuracy, TA)和相关系数(Correlation Coefficient, CC)来评估预测结果,这些指标的定义如下:

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

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

$$TA = \frac{TP + TN}{TP + FN + TN + FP} \times 100\%$$

$$CC =$$

$$\frac{TP \times TN - FP \times FN}{((TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP))^{1/2}}$$

其中:TP和FN分别表示被预测为真、假剪接位点的真实剪接位点的数目,FP和TN分别表示被预测为真、假剪接位点的假剪接位点的数目。

## 2 实验结果

令L选取不同的值截取不同的剪接位点序列长度预测了选择性和组成性剪接位点。在图5、6中分别给出了反映选择性和组成性剪接位点预测精确度随L变化的接收机操作特性(Receiver Operating Characteristic, ROC)曲线,以假阳率为横坐标,真阳率为纵坐标。

从图5、6中可以看到,对于3'剪接位点,无论是选择性还是组成性,预测精确性都随着L的增加而增加,当L=60时,精确度达到最大,但是随着L的增加,预测精确性的增加幅度会变小,在图5(a)和图6(a)中,L选择60相对于50的精确度只有很小的改善。对于5'剪接位点,精确度并不是一直随着L的增加而增加,精确度在L=50时达到最大,当L=60时精确度反而会降低,因此对于5'剪接位点L的最优值是50。

在表2中比较了可变剪接位点和组成性剪接位点的预测结果。曲线下面积(Area Under the Curve, AUC)表示ROC曲线所围成的面积大小,总面积为1。

表2 选择性和组成性剪接位点预测结果

剪接位点	Sn/%	Sp/%	TA/%	CC	AUC	L
选择性3'	83.97	79.74	81.77	0.636	0.900	60
选择性5'	92.77	87.89	90.33	0.808	0.961	50
组成性3'	90.49	87.07	88.79	0.776	0.950	40
组成性5'	95.58	93.71	94.65	0.893	0.986	50

从表2可以看出,预测精度最高的是组成性5'剪接位点,总精度达到了94.65%,AUC达到了0.986,敏感性、特异性和相关系数也分别达到了最高;5'剪接位点的预测精确性要高于3'剪接位点。组成性剪接位点的预测精度要高于选择性剪接位点,这是因为选择性剪接位点的信号强度通常要比组成性剪接位点的弱,而且被可变剪接的外显子和内含子也可能会显示一个不同于组成性的组分分布,因此选择性剪接位点要比组成性的更难预测<sup>[10]</sup>。

本文还采用了被广泛用来测试分类器的剪接位点数据集进行实验。该数据集由Kulp和Reese建立([http://www.fruitfly.org/data/seq\\_tools/datasets/Drosophila/splicesitesets/](http://www.fruitfly.org/data/seq_tools/datasets/Drosophila/splicesitesets/)),所有样本都符合GT-AG规则,共包含757个真实的供体剪接位点和757个真实的受体剪接位点,2993个假的供体剪接位



点和2466个假的受体剪接位点。在此数据集上进行了FSVM与代价SVM(Cost SVM, CSVM)对比实验,采用3重交叉验证法,样本侧翼序列长度 $L$ 都选为50,实验ROC曲线及数据见图7和表3。

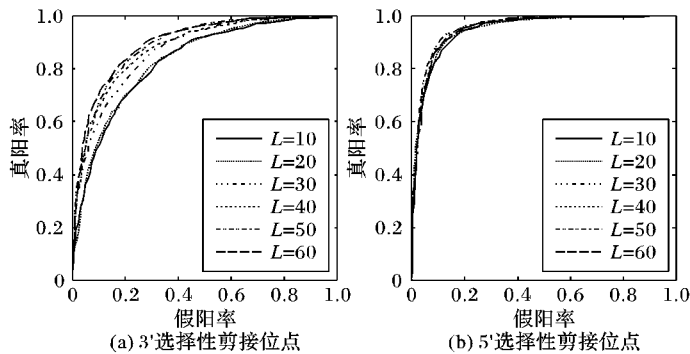


图5 选择性剪接位点预测精度ROC曲线

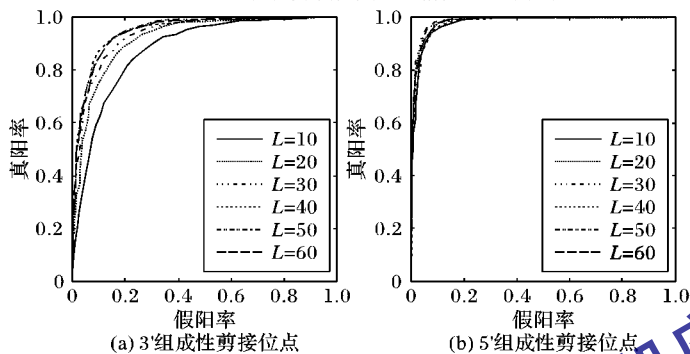


图6 组成性剪接位点预测精度ROC曲线

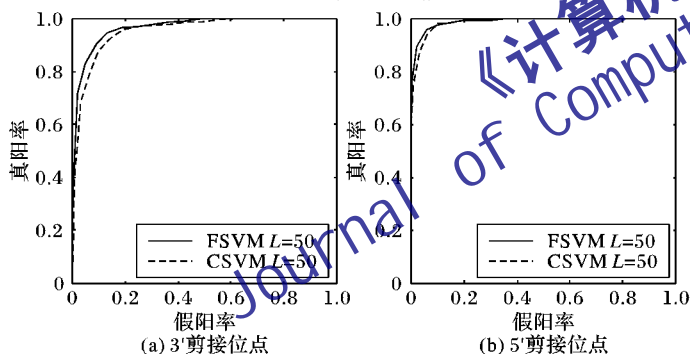


图7 剪接位点ROC曲线对比

表3 剪接位点预测结果

SVM	剪接位点	$Sn/\%$	$Sp/\%$	$TA/\%$	$CC$	$AUC$
FSVM	3'	80.56	96.23	92.55	0.788	0.969
	5'	93.25	97.49	96.44	0.897	0.992
CSVM	3'	72.62	96.11	90.60	0.730	0.957
	5'	87.30	97.19	95.20	0.850	0.988

结果表明,本文方法在3'和5'剪接位点数据集上的总精度都超过了92%,5'剪接位点的总精度达到了96.44%。FSVM在所有评价指标上都超越了CSVM。3'剪接位点的敏感性最多提高了7.94%;相对于3'剪接位点,FSVM有微弱优势, $AUC$ 达到了0.992。在此数据集上同样发现5'剪接位点的识别精度高于3'剪接位点的识别精度。

### 3 结语

本文设计的模糊隶属度函数通过提高支持向量的隶属度,同时降低噪声点的隶属度,提高了支持向量对最优分类面的贡献。结果表明该方法在本文提取的数据集和标准数据库中下载的数据集中识别精度很高,最高达到95.20%。通过对剪接位点侧翼序列长度的选择,我们发现当剪接位点两侧分别选择40~60 bp长度时,可以达到很高的识别精度,这证明了确定剪接位点的保守信息大多在剪接位点两侧60 bp范围内。通过与CSVM对比发现本文方法在敏感性、特异性、识别精度和 $AUC$ 等重要指标上都有一定的优势。在两种数据集的验证也说明了本方法的推广性较好,识别精度高,适合应用在基因识别的研究中。

### 参考文献:

- [1] KAZANK. Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged[J]. Trends Plant Science, 2003, 8(10): 468-471.
- [2] 晋宏岩, 梁江复, 张利绒. 使用估计的反应自由能预测组成性和可变剪接位点[J]. 生物物理学报, 2009, 25(1): 57-58.
- [3] LIN CHUNFU, WANG SHENGDE. Fuzzy support vector machines[J]. IEEE Transactions on Neural Net2 works, 2002, 13(2): 464-471.
- [4] TAKUYA INOUE, SHIGEO ABE. Fuzzy support vector machines for pattern classification[C]. Proceedings of Joint Conference on Neural Networks. New York, USA: IEEE Press, 2001: 1449-1454.
- [5] 唐浩, 廖与禾, 孙峰等. 具有模糊隶属度的模糊支持向量机算法[J]. 西安交通大学学报, 2009, 43(7): 40-43.
- [6] 张翔, 肖小玲, 徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报, 2006, 17(5): 951-958.
- [7] 李雷, 鲁廷玲, 周蒙蒙等. 基于核方法的一种新的模糊支持向量机[J]. 计算机技术与发展, 2010, 20(2): 9-11.
- [8] 张秋余, 竭洋, 李凯. 模糊支持向量机中隶属度确定的新方法[J]. 兰州理工大学学报, 2009, 35(4): 89-93.
- [9] 黄金艳, 李通化, 陈开. 基于知识编码的剪接位点预测[J]. 同济大学学报: 自然科学版, 2007, 35(11): 1548-1551.
- [10] WANG M, MARIN A. Characterization and prediction of alternative splice site[J]. Gene, 2006, 366(2): 219-227.

(上接第1032页)

- [5] TAO WENBING, TIAN JINWEN, LIU JIAN. Image segmentation by three-level thresholding based on maximum fuzzy entropy and genetic algorithm[J]. Pattern Recognition Letters, 2003, 24(11): 3069-3078.
- [6] 吴一全, 潘喆, 吴文怡. 二维直方图区域斜分的最大熵阈值分割算法[J]. 模式识别与人工智能, 2009, 22(1): 162-168.
- [7] AI-HUSSAIN A, EI-ZAART A. Moment-preserving thresholding using Gamma distribution[C]// 2010 International Conference on Computer Engineering and Technology. Chengdu: IEEE, 2010: 323-325.

- [8] 张新明, 刘斌, 李双, 等. 二维直方图斜分最大类间交叉熵的图像分割[J]. 计算机应用, 2010, 30(9): 2453-2457.
- [9] CHEN IPING, LAI MINGYI, WANG LINGLING. An improvement in the moment-preserving thresholding method[J]. International Journal of Imaging Systems and Technology, 2008, 18(5/6): 365-370.
- [10] 于新文. 几种图像分割算法在棉铃虫图像处理中的应用[J]. 中国农业大学学报, 2001, 6(5): 69-75.
- [11] LOPES N V, MOGADOURO do COUTO P A, BUSTINCE H, et al. Automatic histogram threshold using fuzzy measures[J]. IEEE Transactions on Image Processing, 2010, 19(1): 199-204.