文章编号:1001 - 9081(2009)10 - 2748 - 03

基于非线性支持向量机的原核生物基因识别

张继宏,李小霞,孙 波
(西南科技大学 信息工程学院,四川 绵阳 621010)
(hr0217@163.com)

摘 要:应用非线性最小二乘支持向量机对原核生物进行基因识别,通过寻找序列开放阅读框(ORF),并与可靠基因位点文件进行比较产生训练样本集,然后提取样本 GC 含量和 Z 曲线特征,并利用 T 检验方法检验各特征值所包含的信息量,设计出了非线性最小二乘支持向量机分类器识别基因。结果表明非线性最小二乘支持向量机的识别率比 Fisher 判别和线性支持向量机在不同的特征组合下分别提高了 7.09% ~29.97% 和 10.97% ~25.45%,并且在特征值信息量较小的情况下非线性最小二乘支持向量机更能表现其优越性。

关键词:基因识别;非线性最小二乘支持向量机;原核生物;GC 含量;Z 曲线;T 检验
中图分类号: TP302 **文献标志码**:A

Prokaryotes gene identification based on nonlinear SVM

ZHANG Ji-hong, LI Xiao-xia, SUN Bo
(*School of Information Engineering, Southwest University of Science and Technology, Mianyang Sichuan* 621010, *China*)

Abstract: This paper presented a nonlinear least squares support vector machine method to identify the prokaryotes gene. This method generated training sample sets by searching sequence Open Reading Frames (ORF) and comparing ORF sets with reliable gene location document, extracted two sample features of GC content and Z curve, examined information content of these features by T-test, and designed nonlinear least squares support vector machine classifier to recognize gene. The results show that the recognition rates of nonlinear least squares support vector machine are 7.09% ~29.97% and 10.97% ~25.45% higher than Fisher and linear support vector machine respectively under different feature combinations, and the nonlinear support vector machine method performs better when the feature information content is less.

Key words: gene identification; nonlinear least squares support vector machine; prokaryote; GC content; Z curve; T-test

0 引言

开放阅读框(Open Reading Frame,ORF)是一个以起始密码子开始,以同相位的终止密码子结束的一个 DNA 片段。满足 ORF 结构特征的片段不一定编码蛋白质,而编码蛋白质的基因却肯定是 ORF。所以,ORF 的识别是证明一个新的 DNA 序列为特定的部分或全部的蛋白质编码基因的先决条件[1]。文献[2]作者提出的 Z 曲线方法从几何学的角度显示和分析 DNA 序列,其优点在于使用 33 个参数,其中 Z 变换是主要特征,将用于本文的基因识别算法中。

基因识别中常用的模式识别方法有线性判别分析、聚类、神经网络和隐式马尔可夫模型,其中应用最广泛的方法是隐式马尔可夫模型(Hidden Markov Model,HMM)。现有的著名的基因识别程序 GeneMarks[3], Glimmer[4] 和 GeneHacker Plus[5]都是以高阶或隐马尔科夫模型为基础。虽然其识别率已经相当高,但是也存在明显的缺点:使用隐马尔科夫算法需要对已知的基因结构信号进行学习或训练,对那些与学习过的基因结构不大相似的基因,其预测效果不佳[6]。

支持向量机是近几年发展起来的一种新的机器学习方法,它基于结构风险最小化原则,能较好地解决小样本学习问题和非线性问题,因此在模式识别和生物序列分析中得到了广泛的应用。文献[7]讨论了基于最小二乘法的支持向量机分类器。通过解决一组线性方程组而不是经典的二次方程组来达到逼近的效果。文献[8]作者应用一种基于支持向量机的平衡取小法来识别真核生物,能更好地提取剪切位点附近保守序列的统计特征,获得了较好的识别效果。

本文首先介绍了非线性最小二乘支持向量机理论。然后通过寻找开放阅读框产生训练样本集,并对其提取特征。最后设计非线性最小二乘支持向量机分类器,并与 Fisher 判别和线性支持向量机的识别率进行比较得出结论。

1 非线性最小二乘支持向量机

对于非线性样本集 (x_i,y_i),$i = 1,\cdots,n$,$x \in \mathbf{R}^n$,$y \in \{+1,-1\}$,通过一个非线性函数 $\phi(x)$ 将训练集数据 x 映射到一个高维线性特征空间,在这个高维线性空间中构造最优分类超平面,并得到分类器的决策函数。因此,最佳分界面设计问题可以表示为:

$$J(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n}e_i^2;\ i = 1,2,\cdots,n \tag{1}$$

约束条件为:

$$y_i(\langle \boldsymbol{w},\phi(x_i)\rangle - b) = 1 - e_i;\ i = 1,2,\cdots,n \tag{2}$$

定义拉格朗日函数为:

$$L(\boldsymbol{w},b,a) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n}e_i^2 - \sum_{i=1}^{n}a_i\{y_i[\boldsymbol{w}\cdot\phi(x_i) + b] - 1 + e_i\} \tag{3}$$

a_i 为拉格朗日乘子。求拉格朗日函数的极小值可对式(3)求

收稿日期:2009 - 03 - 31;**修回日期**:2009 - 05 - 27。

作者简介:张继宏(1985 -),女,辽宁辽阳人,硕士研究生,主要研究方向:模式识别、智能系统; 李小霞(1976 -),女,四川安岳人,副教授,博士,主要研究方向:模式识别、生物信息学; 孙波(1984 -),男,安徽阜阳人,硕士研究生,主要研究方向:模式识别、智能系统。

偏微分并令它们等于零,如式(4):

$$\begin{aligned}\frac{\partial L}{\partial w} &= w - \sum_{i=1}^{n} a_i y_i \phi(x_i) = 0\\ \frac{\partial L}{\partial b} &= \sum_{i=1}^{n} a_i y_i = 0\\ \frac{\partial L}{\partial e_i} &= a_i - \gamma e_i = 0\\ \frac{\partial L}{\partial a_i} &= y_i[w \cdot \phi(x_i) + b] - 1 + e_i = 0\end{aligned} \tag{4}$$

式(4) 可以改写为求解线性方程式(5):

$$\begin{bmatrix} 0 & -\boldsymbol{Y}^{\mathrm{T}} \\ \boldsymbol{Y} & \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} + \gamma^{-1}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{E} \end{bmatrix} \tag{5}$$

其中:$\boldsymbol{Y} = [y_1, \cdots, y_n]^{\mathrm{T}}$,$\boldsymbol{E} = [1, \cdots, 1]^{\mathrm{T}}$,$\boldsymbol{a} = [a_1, \cdots, a_n]^{\mathrm{T}}$ 都是 $n \times 1$ 维向量,$\boldsymbol{I}$ 是 $n \times n$ 维单位矩阵。$\boldsymbol{Z} = [y_1\varphi(x_1), \cdots, y_n\varphi(x_n)]^{\mathrm{T}}$ 是 $n \times l$ 维矩阵,则 $\boldsymbol{Z}\boldsymbol{Z}_{ij}^{\mathrm{T}} = y_i y_j \phi(x_i) \cdot \phi(x_j) = y_i y_j K(x_i, x_j)$,其中 $K(x_i \cdot y_i) = \phi(x_i) \cdot \phi(y_i)$ 称为核函数。根据泛函的有关理论,只要一种核函数满足 Mercer 条件,它就对应着某一变换空间中的内积。

求解上述问题后,若 a_i^* 和 b^* 为最优解,则得到最优分类函数是:

$$f(x) = \mathrm{sgn}\left[\sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*\right] \tag{6}$$

将大肠杆菌ORF的特征训练样本集代入式(5)求出最优解 a_i^* 和 b^*,设计最优分类器,然后将待识别ORF的样本集代入式(6)进行判别,当判别式大于0时判别此ORF为基因,否则为非基因。这样通过设计非线性最小二乘支持向量机可避免采用非线性映射进行广义线性分类所引起的高维数问题。

2　寻找开放阅读框

由于蛋白质由三联体密码子编码,所以一个双链 DNA 序列就有 6 个潜在的阅读框架,其中一条链的 3 个读框成为“正向”读框,互补链上的 3 个为“反向”读框,ORF 识别包括检测 6 个阅读框架。而 ORF 与基因对应的实际情况却要复杂得多,如由于原核生物的基因组结构比较紧凑,ORF 之间重叠很常见,如果不把这些重叠的 ORF 消除掉,就会大大影响基因识别程序的成绩。本文寻找开放阅读框的步骤如图 1 所示。

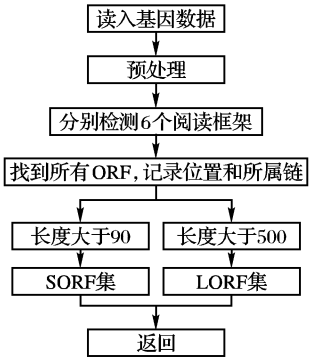

图 1　寻找开放阅读框框图

首先对从美国国立卫生研究院的基因序列数据库(GeneBank)下载的大肠杆菌全基因组序列的 U00096.fna(版本号 168.0)进行预处理,去掉文件中的一些序列附加信息:第一行的注释信息和每行之后的换行,经过预处理后把只包含核苷酸序列的数据保存起来。然后,在双链 DNA 上寻找相互之间不重叠的那些 ORF。最后,选取出所有长度大于 500bp 的 ORF,作为种子 LORF 集,一般来说,ORF 长度越大,编码蛋白质的可能性也越大;并且最短 ORF 的长度设定为 90bp,选出所有长度大于 90bp 的 ORF,作为种子 SORF 集。这样就初步的找到了两个 ORF 集合。

3　训练集产生和特征提取

分别把前面产生的两个 ORF 集与 GeneBank 标注的可靠基因位点文件 ecoli844.dat(http://tubic.tju.edu.cn/GS-Finder/)进行对比:把 LORF 集中与 ecoli844.dat 标注的起始位点和终止位点相同的 ORF 作为训练集的正样本,得到了正样本 71 个;由于 SORF 集中与位点文件 ecoli844.dat 标注的 ORF 相同的有 84 个,所以从与可靠基因位点文件不匹配的 SORF 中任选 84 个作为训练集的负样本。这样得到了训练样本集。

通过分析大肠杆菌全基因组序列和位点信息,提取了训练样本集 GC 含量(即鸟嘌呤(Guanine) 加胞嘧啶(Cytosine)的含量) 和 Z 曲线一共 4 个特征。Z 曲线的变换公式如(7):

$$\begin{cases} X_n = (A_n + G_n) - (C_n + T_n) \\ Y_n = (A_n + C_n) - (G_n + T_n) \\ Z_n = (A_n + T_n) - (G_n + C_n) \end{cases} \tag{7}$$

其中:X_n 表示嘌呤/嘧啶碱基(*R-Y*) 沿序列的分布,Y_n 表示氨基/酮基碱基(*M-K*) 沿序列的分布,Z_n 表示强/弱氢键碱基(*S-W*) 沿序列的分布。通过利用 T 检验方法检验特征值所包含的信息量,实验结果显示 GC 含量的 q 值为 2.1994,Z 曲线三个特征的 q 值分别为 $qX = 0.9093$;$qY = -1.7702$;$qZ = -5.9839$。对于自由度为 $2 \times (71 + 84) - 2$ 和 0.0025 的截尾概率,得到的置信区间为 $[-2, 2]$。因此 Z_n 包含的信息量最大,其次是 *GC* 和 Y_n,最小的是 X_n。在下面的实验中把4个特征中每两个特征进行组合用于分类器的设计。

4　结果和讨论

实验中分别用 Fisher 判别、线性支持向量机(SVM)和非线性最小二乘支持向量机(N-SVM)对训练集进行分类训练,并对训练集的分类效果进行评价。其中,非线性最小二乘支持向量机采用的核函数如式(8):

$$K(x_i \cdot y_i) = \exp\left\{-\frac{|x_i - x_j|^2}{\sigma^2}\right\} \tag{8}$$

其中 $\sigma = 1$。按特征值信息量从大到小排列得到分类器的错分率如表 1 所示。

表 1　三种分类器的错分率统计表　　%

特征值	Fisher	SVM	N-SVM
$Z_n + GC$	18.71	23.23	9.00
$Z_n + Y_n$	20.65	22.58	11.61
$Z_n + X_n$	19.35	23.23	12.26
$GC + Y_n$	41.29	36.13	12.26
$GC + X_n$	41.94	37.42	11.97
$Y_n + X_n$	34.84	30.32	14.19

从表 1 中可以看出:随着特征值信息量的减小,分类器的错分率都有所增加。但是非线性最小二乘支持向量机的识别效果总要优于 Fisher 判别和线性支持向量机的识别效果,并且在利用 *GC* 和 X_n 这对特征时,非线性最小二乘支持向量机的识别率与其他两种分类器的识别率相比提高最多,比 Fisher 判别提高了 30%,比线性支持向量机也提高了 26%。这说明了在特征值信息量较小的情况下非线性最小二乘支持向量机更能表现其优越性,分类效果如图 4 和图 5 所示。

5　结语

本文采用非线性最小二乘支持向量机的方法对大肠杆菌全基因序列数据进行分析。实验结果表明这种方法识别率高,

推广性较强。但是要准确识别出编码蛋白质基因需要进一步的研究。一方面,仅从GC含量和Z曲线特征上识别编码ORF还有一定的局限性,可能还需要借助其他的特征;另一方面,非线性支持向量机的训练速度极大地受到训练集规模的影响。对于超大规模的数据集,如何高效地进行训练和测试也是一个需要研究的重要问题。

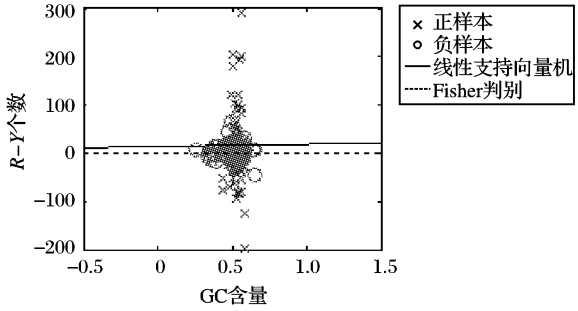

图4 Fisher判别和线性支持向量机

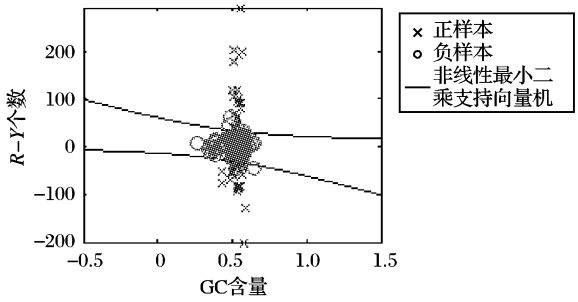

图5 非线性最小二乘支持向量机

参考文献:

[1] 郭峰彪. 原核生物蛋白质编码区识别及基因组序列分析[D]. 天津: 天津大学理学院, 2005.

[2] 张春霆. 人与其他生物基因组若干重要问题的生物信息学研究[J]. 自然科学进展, 2004, 14(12): 1367 - 1374.

[3] BESEMER J, LOMSADZE A, BORODOVSKY M. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions [J]. Nucleic Acids Research, 2001, 29(12): 2607 - 2618.

[4] DELCHER A L, HARMON D, KASIF S, *et al*. Improved microbial gene identification with GLIMMER[J]. Nucleic Acids Research, 1999, 27: 4636 - 4641.

[5] YADA T, TOTOKI Y, TAKAGI T, *et al*. A novel bacterial gene-finding system with improved accuracy in locating start codons[J]. DNA Research, 2001, 8: 97 - 106.

[6] 史良. 国内外基因计算机识别的研究方法及进展[J]. 北京生物医学工程, 2004, 23(1): 73 - 74.

[7] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293 - 300.

[8] 闻芳. 基于支持向量机(SVM)的剪接位点识别[J]. 生物物理学报, 1999, 15(4): 733 - 738.

(上接第2743页)

对比表1和表4可以得出,音频特征向量经过PCA变换后,聚类正确率提高了近8%,说明通过PCA变换有效地去除了类别之间特征数据的相关性,降低了所提取特征中的冗余信息。

表4 未进行PCA变换的四交叉验证聚类正确率 %

实验	第一组	第二组	第三组	第四组	平均
实验一	72	62	72	84	72.5
实验二	74	62	72	84	73.0
实验三	74	72	72	84	75.5
实验四	76	68	72	84	75.0
实验五	68	62	72	84	71.5
平均	72.8	65.2	72.0	84.0	73.5

4 结语

本文提出的基于广播音频聚类算法,利用小波变换和PCA变换对音频片段进行特征提取和处理,采用Mean-Shift算法对音频信号进行初步聚类,然后利用快速近邻法对Mean-Shift的结果进行一次修正,最后合并仅含有单个样本类别的类进行二次修正。该音频算法不需要任何先验条件,而且执行速度快。仿真实验表明,该算法较之单一Mean-Shift算法和未进行PCA处理的特征集的聚类,效果有一定的提高,但对自然的音频流中存在着少数的时间间隔较短的讨论、采访、背景音较大等现象处理得不是很好,因此下一步可以从提高算法的综合性能方面进行改进,以便使算法的应用更加稳定灵活。

参考文献:

[1] (加)韩家炜, (加)坎伯. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001: 223 - 262.

[2] 毛韶阳, 李肯立. 优化K-means初始聚类中心研究[J]. 计算机工程与应用, 2007, 43(22): 179 - 181, 219.

[3] 王元珍, 王健, 李晨阳. 一种改进的模糊聚类算法[J]. 华中科技大学学报: 自然科学版, 2005, 33(2): 92 - 94.

[4] FUKUNAGA K, HOSTETLER L. The estimation of the gradient of a density function, with applications in pattern recognition[J]. IEEE Transactions on Information Theory, 1975, 21(1): 32 - 40.

[5] COMANICIU D, MEER P. Mean shift: A robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603 - 619.

[6] CHENG Y Z. Mean shift, mode seeking, and clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790 - 799.

[7] 郑继明, 俞佳. 基于小波变换和支持向量机的音频分类[J]. 计算机工程与应用, 2009, 45(11): 158 - 161.

[8] PARK C H, PARK H, PARDALOS P. A comparative study of linear and nonlinear feature extraction methods[C]// ICDM: Proceedings of the 4th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2004: 495 - 498.

[9] KIM H C, KIM D, BANG S Y. A PCA mixture model with an efficient model selection method[C]// Proceedings of International Joint Conference on Neural Networks. Washington, DC: IEEE Press, 2001: 430 - 435.

[10] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2003: 84 - 86.