中国农业大学学报 2005,10(2):58-64
Journal of China Agricultural University

基于支持向量机的DNA序列分类系统的设计与实现

蔡 春[1,2] 万潇楠[2] 逯燕玲[2]
(1. 中国农业大学 经济管理学院,北京 100083; 2. 北京联合大学 应用文理学院,北京 100083)

摘 要 针对传统统计方法进行DNA序列分类时要求DNA序列样本的概率分布函数已知,但多数情况下概率分布函数未知这一问题,采用支持向量机这一新的机器学习方法对DNA序列进行分类;以VB和Matlab为主要工具开发了基于支持向量机的DNA序列分类系统。结果表明:该系统能够动态选择DNA训练样本、待测试样本,以及支持向量机模型中的参数,并根据用户的指定条件动态输出计算结果;对于预测一批已知正确分类答案的DNA序列,系统能够自动统计识别率,以观察参数变化对于算法执行结果的影响。支持向量机能够在概率分布函数未知的条件下对DNA序列进行分类。

关键词 DNA序列分类; 支持向量机; 动态化输入; 动态化输出
中图分类号 TP 182 **文章编号** 1007-4333(2005)02-0058-07 **文献标识码** A

Design and realization of a DNA sequence classification system based on support vector machines

Cai Chun[1,2], Wan Xiaonan[2], Lu Yanling[2]
(1. College of Economic and Management, China Agricultural University, Beijing 100083, China;
2. College of Arts and Sciences of Beijing Union University, Beijing 100083, China)

Abstract The distribution of DNA sequence samples must be known when classifying by the traditional statistical methods, unfortunately, it is unknown in most application cases. This paper mainly developed a DNA sequence classification system based on support vector machines (SVM) by VB and Matlab and proposed an new approach to express the DNA sequence data. The test results showed that the system had the merits of dynamically selecting DNA training samples and the samples to be tested, as well as supporting the parameters in SVM model. The system can also dynamically output the calculating results on demand of users, automatically make a statistics of reorganization rate to investigate the effects of parameters variations on the the computing results for a prediction process of a set of correctly classified DNA sequences.

Key words DNA sequence classification; support vector machines; dynamic input; dynamic output

传统的统计学在解决机器学习问题中起重要作用;但其研究的主要是渐近理论,即当样本趋向于无穷多时的统计性质,而在现实问题中样本数目通常是有限的,所以传统的统计学经常表现出较差的推广能力。20世纪70年代末国外出现了研究小样本情况下机器学习性质和规律的"统计学习理论",90年代在这一理论下发展出了新的模式分类器——支持向量机(support vector machines,SVM),广泛应用于手写体识别、人脸检测、文本分类等领域,与传统方法相比取得了相当或更好的结果,从而推动了它在其他模式识别领域的应用[1-5]。

DNA序列由4种核苷酸a、c、g、t排列组成,序列分类是基因研究的基础。DNA序列的长度不一,传统的统计方法要求DNA序列样本的概率分布函数已知,但在多数情况下,DNA序列样本的概率分布函数是未知的。支持向量机是数据挖掘的一项新

收稿日期:2004-09-06
基金项目:国家自然科学基金资助项目(10371131)
作者简介:蔡春,博士研究生,主要从事运筹学、模式识别的研究,E-mail:caichun@ygi.edu.cn

技术,关于用支持向量机方法研究 DNA 序列分类的文献很有限[6-8],且大多针对蛋白质分子这一特殊的 DNA 序列的亚细胞定位预测进行研究。笔者提出基于支持向量机的 DNA 序列分类系统的总体设计,旨在解决 DNA 序列分类问题。

1　支持向量机分类算法简介

给定训练集$\{(x_i, y_i) \mid i = 1,2,\cdots,m\}$,其中 $x_i \in \mathbf{R}^n$,标签 $y_i \in \{+1,-1\}$ 是点 x_i 的类别(这里只讨论 2 类问题)。根据已知样本的信息提炼出分类的本质,以及对未知类别的数据进行很好的分类预测(模式识别)[10],传统的统计方法如贝叶斯方法、马尔可夫链、主成分分析方法可以求解此类问题(前提是样本的概率分布函数 $P(x^{\mathrm{T}}, y)$ 已知),但随着机器学习和数据挖掘的迅速发展,以及 Vapnik V 等提出的支持向量机方法赋予分类(模式识别)问题新的意义[1],分类问题再次引起了研究人员的热切关注。

在未知总体分布 $P(x^{\mathrm{T}}, y)$ 的前提下,独立地抽取未知分布的训练集$\{(x_i, y_i) \mid i = 1, \cdots, m\}$,支持向量机方法进行学习的主要任务是找到 1 个机器即"函数",$f: \mathbf{R}^n \to \{-1, +1\}$,不但能"更好"地拟合训练集,而且具有"更好"的推广能力[2]。所谓推广能力即期望风险,$R_{P(x^{\mathrm{T}},y)}(f) = \int_{\mathbf{R}^n \times \{+1,-1\}} I_{\{f(x^{\mathrm{T}}) \neq y\}}(x^{\mathrm{T}}, y)\,\mathrm{d}P(x^{\mathrm{T}}, y)$,其中 $I_A(x)$ 称为 0-1 损失函数,A 为 $f(x^{\mathrm{T}}) \neq y$,其定义为:当 $z \in A$,$I_A(z) = 1$;当 $z \notin A$,$I_A(z) = 0$。由于样本分布函数 $P(x^{\mathrm{T}}, y)$ 未知,因此期望风险无法计算,如果线性模式分类器 $f(x^{\mathrm{T}}) = \mathrm{sgn}(\langle w, x^{\mathrm{T}}\rangle + b)$ 能够对训练集进行正确分类,其中的 $\langle w, x^{\mathrm{T}}\rangle + b$ 为规范形式,且 $\| w \| \leq \sqrt{2m - 128\zeta^2}/(8\sqrt{2}\zeta)$,文献[9]给出了函数 f 期望风险的 1 个上界:

$$R(f) \leq \frac{2}{m}\left[128\zeta^2(\| w \|^2 + 1)\log_2\left(\frac{em}{16\zeta^2(\| w \|^2 + 1)}\right)\log_2 32m + \log_2\left(\frac{2m}{\delta}\right)\right]$$

为了使上界达到最小值,则要使 $\| w \|^2$ 最小,即在经验风险为 0 的情况下最优化分类器的推广能力;如果线性模式分类器不能对训练集进行正确分类,则引入惩罚系数 C 平衡推广能力和经验风险,得到下面的最优化问题[9]:

$$\left.\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{m}\xi_i \\ &y_i(x_i w + b) + \xi_i \geq 1 \quad i = 1,2,\cdots,m \\ &\xi_i \geq 0 \quad i = 1,2,\cdots,m \end{aligned}\right\} \tag{1}$$

应用 KKT 条件[11,12],可以得到式(1)的对偶问题:

$$\left.\begin{aligned} &\max_{\alpha_i} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\langle x_i, x_j\rangle \\ &0 \leq \alpha \leq C \quad i = 1,2,\cdots,m \\ &\sum_{i=1}^{m}\alpha_i y_i = 0 \end{aligned}\right\} \tag{2}$$

从而得到问题(2)的最优解 $\alpha_i\,(i = 1,2,\cdots,m)$,线性分类器参数 $w = \sum_{i=1}^{m} y_i\alpha_i x_i$,相应的线性分类器 $f(x^{\mathrm{T}}) = \mathrm{sgn}\left(\sum_{i=1}^{m} y_i\alpha_i(x, x_i) + b\right)$,$\alpha_i > 0$ 所对应向量 x_i 为支持向量[2]。

可以看出样本在最优化问题(2)中仅仅以向量的内积形式出现,正是这一重要特点,使支持向量机线性分类器可以推广到非线性情况,这是 Wolfe 对偶问题带来的一个最好的副产品。用线性分类器学习效果不好时,设法用非线性分类器进行学习,即引入映射 $\Phi: \mathrm{R}^n$,把 $x_i \in \mathrm{R}^n$ 映射到特征空间 F 中(一般特征空间的维数比原空间高),在特征空间 F 中采用线性分类器,由于分类器的参数总以样本的内积形式出现,为此引入核函数 $K(x_i, x_j) = \Phi(x_i) \times \Phi(x_j)$,相应的问题(2)变形为:

$$\left.\begin{aligned} &\max_{\alpha_i} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K\langle x_i, x_j\rangle \\ &0 \leq \alpha_i \leq C \quad i = 1,2,\cdots,m \\ &\sum_{i=1}^{m}\alpha_i y_i = 0 \end{aligned}\right\} \tag{3}$$

得到相应的分类器 $f(x^{\mathrm{T}}) = \mathrm{sgn}\left(\sum_{i=1}^{m} y_i\alpha_i K(x, x_i) + b\right)$,这样在 R^n 空间中就实现了用非线性分类器进行学习的目的。常用的核函数有多项式核函数(poly)和径向基核函数(rbf)。

2　基于支持向量机的 DNA 序列分类系统

2.1　数字化 DNA 序列的方法构想

由于每个 DNA 序列都是由 a、c、g、t 4 个字符(核苷酸)随机组成的字符串(如 aggcacggaaaaacgggaatt),数字化 DNA 序列的方法很多,不同的方法对于分类的影响也很大;由于支持向量机

算法对样本点具有同样维数的要求,可以分别统计每个 DNA 序列中 a、c、g、t 出现的个数,这样就可以用 4 维向量标记每个 DNA 序列,但这种提取特征的方法缺点很明显:一方面没有生物意义;另一方面很可能出现不同类别的 DNA,但标记它们的 4 维向量却是相同的。生物学中"密码子"的概念就是决定氨基酸的 3 个相邻的核苷酸,即 DNA 序列中由 4 个字符组成的 64 种不同的 3 字符串序列片段,其中大多数用于编码构成蛋白质的 20 种氨基酸[14],为此本研究采用统计每个 DNA 序列相邻 3 个字符 aaa、aac、aag、aat、…、ttt 出现的个数的方法,这样每个 DNA 序列可由 1 个 64 维的向量进行表示(4 ×4 × 4 = 64)。

2.2 系统开发工具的选择

由于支持向量机分类算法涉及到大规模的矩阵运算,处理的主要任务又为较复杂的二次规划问题,所以 SVM 分类算法采用 Matlab 语言编写[15]。另外,本系统是 SVM 应用系统,需要设计用户端的界面、实现一些较复杂的功能,Visual Basic 6.0 具有使用便捷、执行速度快等优点[16],本系统中采用 Visual Basic 6.0 设计应用系统的主界面,编写主程序,同时利用接口技术实现 VB 和 Matlab 的数据传递。后台数据库管理系统采用 MS SQL-SERVER。

2.3 数据准备

训练样本取自 2000 年全国大学生数学建模竞赛题目[13]提供的数据文件 Art-model-data 中的前 40 号样本,其中 110 号 DNA 序列为 a 类,11～20 号为 b 类,21～40 号样本的类别未知。现根据已知类别的 DNA 样本将 21～40 号未知类别的 DNA 进行分类。将已知类别的 DNA 序列和对应的类别存入数据表 OriginalData 中;由于第 21～40 号 DNA 序列和另一待测数据文件 Nat-model-data 中的所有 DNA 序列都是未知类别的(但已经获得了它们的正确分类答案),所以它们作为系统可能的待测试样本,与对应的待测类别答案分别被存入 2 个数据表中,分别命名为 TestData1 和 TestData2。

2.4 系统总体设计

系统宏观运行流程见图 1,总体功能见图 2。

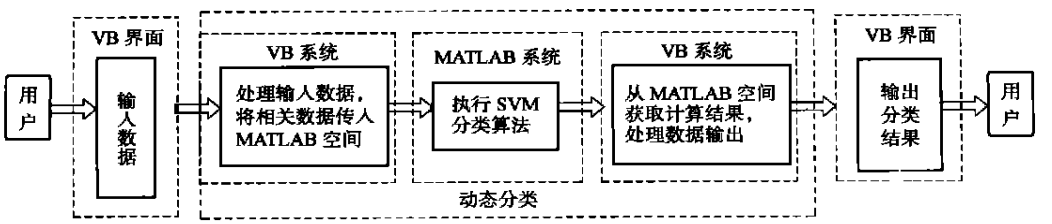

图 1 系统宏观运行流程

Fig. 1 Macro running process of the system

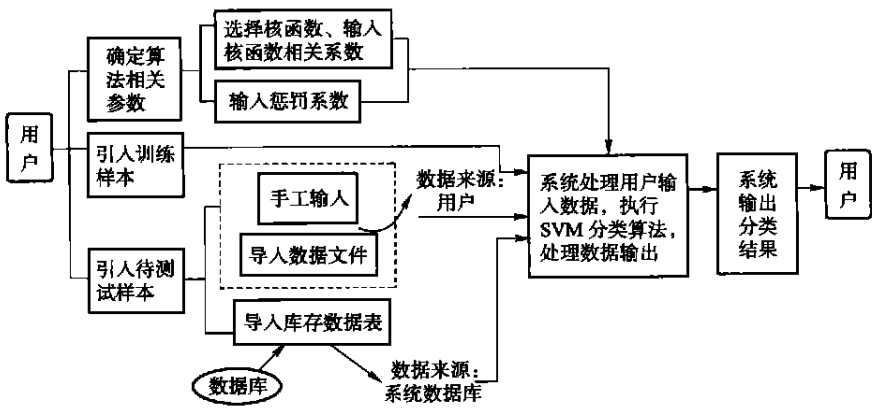

图 2 系统总体功能

Fig. 2 Overall function of the system

1)算法相关参数输入的动态化。用户可任意选择核函数、输入核函数相关系数和惩罚系数,通过变换参数值观察算法执行结果的变化。

2)用于算法调用的训练样本的动态化输入。备

选训练样本由数据表 OriginalData 引入,用户可在其中选择用于算法调用的训练样本,不但数量可以变化,而且可以任意选择训练样本,这样一方面可以观察小样本情况下算法执行的效果,另一方面可以观察当样本数量确定时选择不同训练样本对算法执行结果的影响,并分析产生原因。

3) 待测试样本的动态化。可由用户随意指定待测试样本的引入方式:用户手工输入、非库存数据文件导入和库存数据表导入。具体描述如下:a. 手工输入情况。待测试样本个数由用户指定,根据指定系统动态加载文本框控件用于用户输入 DNA 序列。b. 导入非库存数据文件情况。样本数 > 15 时,用户自制包含任意条 DNA 序列的文本文件,可保存在磁盘上的任何位置,通过系统指定文件位置。c. 导入库存数据表情况,即导入数据表 TestData1 或 TestData2 中的数据,其意义在于对指定样本进行分类。由于库存了这 2 批样本的类别答案,所以可调用算法进行分类后,将计算结果和库存答案进行对比,统计正确识别率情况;可以通过观察正确识别率的变化来体验训练样本和算法相关参数的变化对于算法执行结果的影响,并分析原因。

4) 分类结果输出的动态化。根据用户输入数据执行 SVM 分类算法,输出相应的计算结果和提示信息。

2.5　系统主要数据处理

1) 采用 Matlab 设计的主函数为 function [mytstY] = mysvc(myX, myY, mytstX, myker, myC),其中 myX 表示数字化后的若干 DNA 训练样本组成的矩阵,a 类记为 1,b 类记为 - 1;myY 表示 DNA 训练样本相应类别的向量;mytstX 表示用户输入的若干待测试 DNA 样本数字化后组成的矩阵;myker 表示核函数;myC 表示惩罚系数。Matlab 获得这些参数和核函数相关系数之后,调用该函数便可输出 mytstY,即相应的待测试样本的类别(1 或 - 1)。

2) 数字化 DNA 序列的操作要经常使用,本研究中由子过程 str_convert (ByRef strDNA() As String, count As Integer) 实现。strDNA() 表示 1 个字符串数组,其中的每个元素存储 1 个字符串,在本问题中就是 1 个 DNA 序列;count 表示字符串数组包含的元素个数,在本问题中就是待处理的 DNA 序列的个数。str_convert() 的作用是生成相应的数字化后的若干 DNA 序列组成的二维数组 n,如处理 10 个序列,最后就生成 10 行 64 列的二维数组。函数实参根据用户指定动态确定。

3) VB 和 Matlab 传递数据功能由子过程 call_matlab() 实现。根据在 VB 中控制 Matlab 对象的语法,利用 PutFullMatrix 命令将 VB 变量传递到 Matlab 空间的基本形式为: Call matlab.PutFullMatrix ("Matlab 变量名", "base", Matlab 变量名, VB 变量名),Matlab 变量名是向 Matlab 工作空间传递的数组名称,对应前面所述的 myX, myY, mytstX, myker, myC 以及核函数相关系数 myp1 和 myp2。在 VB 中设置相应的变量为 X, Y, tstX, ker, C, p1, p2,以及对应的 Xi, Yi 等,它们是向 Matlab 工作空间中传递的数组的实部和虚部。利用 matlab.Execute (strexe) 命令,告知 Matlab 需执行的命令,这里 strexe 即 mysvc.m 中的主函数语句。利用 GetFullMatrix 在 VB 空间中提取 Matlab 的计算结果: Call matlab.GetFullMatrix("mytstY", "base", tstY, tstYi),mytstY 为 VB 提取的 Matlab 输出的数组,tstY 和 tstYi 为其实部和虚部,在 VB 中通过判断 tstY 值输出预测类别。

4) 输出分类结果功能由 Output (ByRef strDNA() As String, count As Integer) 子过程实现。strDNA() 存储的或为用户手工输入的各个文本框中的 DNA 序列字符串,或为导入的文本文件中的 DNA 序列字符串,或为库存数据表中每个 DNA 序列字符串;count 对应为用户欲处理的 DNA 序列个数,即动态加载的文本框的个数或本文件中的 DNA 序列个数或库存数据表中 DNA 序列记录的个数。调用 str_convert() 将当前待测试 DNA 样本进行数字化处理,生成 tstX,之后调用 call_matlab() 进行计算,通过判断 tstY 值输出预测的类别。针对来源为用户的数据,搜索整个数据库,如果用户输入的待测试样本恰为库存已知类别答案的 DNA 序列,则输出该类别答案,并和 SVM 计算结果做比较,输出 SVM 计算结果,正误情况已知;对于非库存数据,则告知用户答案未知、正误情况未知。这样,针对用户数据,将计算结果的相关信息尽可能地告知用户,在一定程度上弥补了系统仅作为分类工具输出分类结果而对于类别估计是否正确用户完全不知的不足;针对来源为系统库存数据表的数据,则只需将计算结果和当前数据表中已存答案相比较,输出计算结果、答案和正误情况,并统计输出测试样本数、识别数和识别率。

5) 以上几个子过程都为动态分类过程 cmdclas-

sify.click()提供调用,动态分类功能由分类程序实现,通过操作界面中的"进行分类"按钮完成。数据处理步骤如下。

训练样本输入步骤:

```
If CheckListBox2.ListCount = 0 Then
  re = MsgBox("您没有在备选训练样本中选择,是否默认用于算法调用的为全部备选训练样本?", vbYesNo)
  If re = vbNo Then
    MsgBox"请选择用于算法调用的训练样本"
    Exit Sub
  Else
    ReDim stroriginalX(rsoriginal.RecordCount - 1) As String
    ReDim stroriginalY(rsoriginal.RecordCount - 1) As String
    rsoriginal.MoveFirst
    For i = 0 To rsoriginal.RecordCount - 1
    stroriginalX(i) = rsoriginal.Fields(1)
    stroriginalY(i) = rsoriginal.Fields(2)
    rsoriginal.MoveNext
    Next i

    ReDim Y(rsoriginal.RecordCount - 1) As Double
    For i = 0 To rsoriginal.RecordCount - 1
    If stroriginalY(i) = "a" Then
    Y(i) = 1
    Else
    Y(i) = - 1
    End If
    Next i
      str.convert stroriginalX(), rsoriginal.RecordCount
    ReDim X(rsoriginal.RecordCount - 1, 63) As DoubleElse (用户选择了训练样本)
  End If
Else
  ma = 0
  mb = 0
  For i = 0 To CheckListBox2.ListCount - 1
   If CheckListBox2.ItemForeColor(i) = &H808000 Then
    ma = ma + 1
    Else
    mb = mb + 1
    End If
    Next i
    If ma = CheckListBox2.ListCount Then
    MsgBox"请选择至少一个 b 类训练样本"
    Exit Sub
    End If
    If mb = CheckListBox2.ListCount Then
    MsgBox"请选择至少一个 a 类训练样本"
    Exit Sub
    End If
  End If
```

核函数输入步骤:当用户未选择核函数并默认用于算法调用的核函数为线性核时,重新为储存核函数字符串的动态数组 ker 申请 6 bit ("linear"长度为 6B),将字符串"linear"的每个字符所对应的 ASCII 码存储到 ker。如果用户选择了核函数,提取用户所选核函数字符串长度 lenker,重新为 ker 申请 lenker 位 bit,将用户所选核函数字符串的每个字符所对应的 ASCII 码存储到 ker(由于 VB 中没有数据类型 char,Matlab 中没有 string 类型,而统一数据类型才能够实现二者传递数据,因此设计为在 VB 中将核函数字符串转换为 ASCII 码后,即 ker 定义为 double 型,再传递给 Matlab)。

核函数参数输入步骤:若核函数为 poly,需要输入 degree 参数;若核函数为 rbf,则需要输入 sigma 参数。

惩罚系数输入步骤:由于在 VB 中没有"无穷大"的内定变量,故做以下处理:当用户未输入惩罚系数并默认用于算法调用的惩罚系数为无穷大时,将 strexe 赋值为"[mytstY] = mysvc(myX,myY,mytstX,myker,Inf)",即指定 Matlab 执行的命令字符串,直接代入惩罚系数值为"Inf"。如果用户输入了惩罚系数且输入的为数字,提取用户输入的惩罚系数值赋给 *C*,将 strexe 赋值为"[mytstY] = mysvc(myX,myY,mytstX,myker,myC)";否则判断输入的是否为"Inf",如果是,strexe ="[mytstY] = mysvc(myX,myY,mytstX,myker,myC)",如果否,Output"惩罚系数必须是 1 个数字!"。

3　实验结果及分析

待测试样本引入方式选择“导入库存数据表”,选择“批处理数据 1”(TestData1 中数据),利用本系统对这批 DNA 序列预测类别,参数随机选择,实验记录与分析如下。

1) 核函数对算法执行结果的影响。默认用于算法调用的为全部备选训练样本, C = Inf,选择几个常用核函数,执行结果见表 1。

表 1　核函数对识别数的影响

Table 1　Influence of kernel function on identity number

核函数	核函数系数	测试样本数	识别数
linear		20	14
poly	degree = 2	20	14
rbf	sigma = 2	20	15
sigmiod	scale = 2,offset = 2	20	13
spline		20	18

2) 惩罚系数 C 对于算法执行结果的影响。以 spline 核为例,默认用于算法调用的为全部备选训练样本,执行结果见表 2。

表 2　惩罚系数 C 对识别数的影响

Table 2　Influence of penalty fparameter on identity number

惩罚系数 C	测试样本数	识别数
1	20	17
100	20	17
200	20	17
Inf	20	18

3) 核函数参数对于算法执行结果的影响。以 poly 核为例,默认用于算法调用的为全部备选训练样本, C = Inf,执行结果见表 3。

表 3　核函数参数对识别数的影响

Table 3　Influence of parameters in kernel function on identity number

阶数	测试样本数	识别数
2	20	14
5	20	17
15	20	16
30	20	13

可以看到,当多项式核参数的阶数由小变大时,正确识别数开始也在变大,但大到一定程度时又随之下降,这正体现了多项式核函数阶数增大时数值结果的不稳定性。

4) 小样本对于算法执行结果的影响。以 poly 核为例, degree = 5, C = Inf,训练样本中选择 2 个(a、b 类各选择 1 个),执行结果见表 4。

表 4　小样本对识别数的影响

Table 4　Influence of small sample on identity number

序号	选择的训练样本	测试样本数	识别数
1	第 8 和 17 号	20	11
2	第 1 和 15 号	20	15
3	第 2 和 20 号	20	17
4	第 1 和 4 号	20	18
5	第 3 和 13 号	20	19

可以看出,小样本情况下,如利用第 3 和 13 号样本作为训练样本,得出了比利用所有备选训练样本(20 个)好的结果(19∶17)。这说明了传统上认为的“原始样本越多估计效果越好”在支持向量机中并不一定适用,这也从一方面表明在小样本情形下 SVM 仍然具有较强的推广能力。其中 3 号样本为: cgggcggatttaggccgacggggacccgggattcgggacccgaggaaattcccggattaaggtttagcttcccgggatttagggcccggatggctgggaccc,a 类。13 号样本为:cagttagctgaatcgtttagccatttgacgtaaacatgattttacgtacgtaaattttagccctgacgtttagctaggaatttatgctgacgtagcgatcgactttagcac,b 类。

由表 4 可以看出:序号 1 试验的识别数比序号 5 的少 8 个,在 Matlab 中观察数字化后训练样本组成的矩阵,产生该问题的原因可比较直观地反映出来。如前所述,矩阵的每 1 列为 1 种密码子在各个 DNA 序列中出现的次数,可以看出有些密码子在 a 类样本中的含量明显比 b 类的多,反之亦然,将这样的列称作 a 类或 b 类样本的特征列。观察第 3 号和第 13 号训练样本,同一列上的 2 个数字相差明显较大的有 9 列,而第 8 号和第 17 号训练样本这样的列只有 3 个。用户正好选择了 2 个虽然分别为 a 类和 b 类却没有很好体现出 a 类和 b 类特征的序列,所以自然得不到很好的结果。

4　结　论

所开发的 DNA 序列分类系统可根据已知类别的 DNA 序列提取信息,不需要知道 DNA 序列的分布函数,不仅克服了用传统统计方法对 DNA 序列

分类的局限性,而且能对未知类别的 DNA 序列进行很好的分类预测。该系统有如下特点:

1)操作方便,界面友好。用户可随意指定条件,输入、输出数据,不仅能测试 DNA 序列类别,而且在导入库存数据表情况下可观察参数变化对识别率的影响,并分析产生原因。

2)可移植性强。由于系统的所有数据均采用动态化处理,且程序具备模块化特征,若处理其他关于支持向量机模式识别的问题,只需将数字化 DNA 的子过程改写为其他问题提取实际样本特征的程序,并将数据库中的数据替换即可。

3)直观性强。很容易观察出不同核函数、同一核函数不同惩罚系数 C,对分类结果的影响。

为更加便于分析实验数据,对于每次实验,如果要记录当前实验详细情况(实验结果和实验用到的相关数据等),应该考虑进一步开发 1 个数据库表进行存储、并打印统计报表;另外惩罚参数 C 的优化值得进一步研究。

本研究是在博士生导师邓乃扬教授的指导下完成的,谨致谢意。

参 考 文 献

[1] Vapnik V. 统计学习理论的本质[M]. 张学工译. 北京:清华大学出版社,2000. 1 - 118

[2] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel based learning methods[M]. Cambridge :Cambridge University Press, 2000. 1 - 161

[3] 阎辉,张学工,李衍达. 基于核函数的最大间隔聚类算法[J]. 清华大学学报(自然科学版), 2002,42(1): 132 - 134

[4] Braddley P S, Mangasarian O L. Massive data discrimination via linear support vector machines[J]. Optimization Method and Software, 2000, 13(1) :1 - 10

[5] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2) :121 - 167

[6] Hua Sunjun, Sun Zhirong. Support vector machine approach for protein subcellular location prediction [J]. Bioinformatics, 2001,8(17) :721 - 728

[7] Emanuelsson O, Nielsen H, Brunak S. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence[J]. MolBiol, 2000,300(1) :1005 - 1016

[8] Nakai K. Protein sorting signals and prediction of subcellular localization[J]. Adv Protein Chem, 2000,54(1): 277 - 344

[9] 邓乃扬,田英杰. 数据挖掘中的新方法:支持向量机[M]. 北京:科学出版社,2004. 1 - 272

[10] 边肇祺,张学工. 模式识别(第2版)[M]. 北京:清华大学出版社,2000. 1 - 303

[11] 邓乃扬,诸梅芳. 最优化方法[M]. 沈阳:辽宁教育出版社,1987. 133 - 159

[12] 张春华. 支持向量机最优化问题的研究[D]. 北京:中国农业大学,2004

[13] 赫孝良,戴永红,周义仓. 数学建模竞赛赛题简析与论文点评[M]. 西安:西安交通大学出版社,2002. 91 - 108

[14] 北京医学院. 生物化学[M]. 北京:人民卫生出版社, 1984. 40 - 45

[15] 刘志俭,潘献飞,连军想. Matlab 外部程序接口(6. x)[M]. 北京:科学出版社,2002. 43 - 146

[16] 彭珲,罗强. Visual Basic 程序设计教程[M]. 北京:清华大学出版社,2004. 56 - 99