浅谈三种分类方法的优劣

◎俞童慧 王馨敏 杨旷怡 (厦门大学数学科学学院 361005)

【摘要】本文针对 DNA 序列的分类问题,通过使用 SPSS 和 Matlab 软件,运用 Fisher 判别法、BP 人工神经网络模型和支持向量机方法,由已知 AB 类样本分别对未知类型的 20 个 DNA 序列进行分类. 并通过三个统计分类方法的结果同时综合考虑误差及原理分析对三个统计分类方法进行比较,得出三个统计方法的优势及其不全面之处.

【关键词】DNA 序列; Fisher 判别法; BP 人工神经网络模型; 支持向量机方法

2000 年全国大学生数学建模竞赛 A 题的 DNA 人工序列分类要求根据已给出的 20 个分为 A,B 两类的 DNA 序列对另外 20 个未标明类别的人工序列进行分类,我们由这个类型的题目出发,来谈谈与之相关的三个统计分类方法及其优劣.

很常见的一种判别方法就是通过 Fisher 判别法进行特征值提取及分类,用的是投影的思想,定义一个准则函数 J_F,找到其最大的解来得到 Fisher 线性判别式 ω^*,再根据亲疏程度进行判定分类.

在 DNA 的这题中由于序列是由四种碱基构成,A 和 B 两组各有 10 个观察数据,判别分析就是要根据这些数据在适当的准则下,由问题分析中的特征向量 x_i 确定判别函数: $y = u^T x_i = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3$,再确定临界值 y_e,然后进行判别分类.

Fisher 判别法的实现可以借助 SPSS 软件,用上述算法对已知的 1 ~ 20 组进行分类,除了 A 组的第 4 个被错判之外,其余全部分类正确,正确率达到 95%.

对于未知序列 21 ~ 40 组进行分类,得到的结果是:

A 类: 22,23,25,27,29,34,35,36,37.

B 类: 21,24,26,28,30,31,32,33,38,39,40.

另外一种分类的方法是利用 BP 人工神经网络模型,本文考虑两层前传网络,激活函数由函数 $\varphi(x) = \frac{1}{1+e^{-nv}}$ 来决定. 本题输入层包含三个单元 $k = 1,2,3$,分别为 T,G,C 的含量; 中间层取 $j = 1,2,3$; 输出层包含两个单元 $i = 1,2$,记 A 类的理想输出为 $(1,0)$,B 类的理想输出为 $(0,1)$,其后我们将碱基 T,G,C 的含量输入网络,根据输出模式靠近 $(0,1)$ 还是 $(1,0)$ 来判断其归于哪一组. 为了减小误差,我们让实际输出尽可能接近理想输出.

我们将对应于样品 S 的理想输出记作 $\{T_i^S\}$,实际输出记作 $\{Q_i^S\}$,则实际输出与理想输出的差异为 $E(W) = 0.5\sum(T_i^S - Q_i^S)^2$. 由向后传播算法,求得适当的 W,使 $E(W)$ 达到极小值.

为了保证该模型的准确性,我们对数据进行了筛选. 由 Fisher 函数验证结果可以发现,A 类中第 4 组数据在顺序检验和交叉检验中都出现错误,因此我们认为这个样本是错误值,故删去. 使用 Matlab 软件进行神经网络运算,进过 17 次训练达到目标值,得到如下结果.

A 类: 27,25,29,23,35,34,37.

B 类: 21,22,24,26,28,30,31,32,33,36,38,39,40.

最后要提的是一种支持向量机技术,由于四个碱基比例之和为 1,因此我们将 DNA 序列分类问题转化为三个变量的问题,作 DNA 样本散点图时可将其看作是三维立体空间的散点图,寻找超平面 π 把样本空间分割成两部分. 记这些样本 $x_i, y_i, i = 1,2,\cdots,20, y_i \in \{-1,1\}$,定义 A 的输出类别为 1,B 的输出类别为 -1. 假设超平面 π: $\omega x + b = 0$,同时存在两个平行于 π 的超平面 π_1 和 π_2: $\omega x + b = 1$ 和 $\omega x + b = -1$ $(\omega, x \in R^3)$. 使离 π 最近的刚好分别落在 π_1 和 π_2 上,其他样本都将位于 π_1 和 π_2 之外,因此建立规划模型:

$$\min \frac{1}{2}\|\omega\|^2 \quad s.t \quad y_i(\omega x_i + b) \geqslant 1.$$

构造拉格朗日方程,则 ω 求解方程 (λ_i 为拉格朗日系数),$\omega = \sum_{i=1}^{20} \lambda_i y_i x_i$.

输出结果若为 1 则判为 A 类,若为 -1 则判为 B 类. 由 Matlab 运算,我们可得分界面 $15.8315x - 1.9923y + 16.9883z - 8.6174 = 0$ (其中 x,y,z 轴分别为 a,c,g 频率).

下面我们定量地检验该模型的合理性,把 A,B 类共 20 个样本的数据代入,即让 A,B 类共 20 个样本的数据代入上面的分界面方程中,若 $g(x_i,y_i,z_i) = 15.8315x_i - 1.9923y_i + 16.9883z_i - 8.6174 \geqslant 1$,则判为第 i 个序列为 A 类; 而 $g(x_i, y_i, z_i) \leqslant 1$,则判为第 i 个序列为 B 类. 计算结果 20 个样本判别结果与绝大多数实际相符 (B 类第 2,7 个与实际有很小的误差,小于 0.005% 可忽略),说明该模型合理.

支持向量机得到人工序列 21 ~ 40 的分类结果:

A 类: 23, 25, 27,29,34,35.

B 类: 21,22,24,26,28,30,31,32,33,36,37,38,39,40.

三种分类方法各有其优劣. Fisher 线性判别式对确定性和随机性模式的分类都是适用的,但这个模型也存在着一些不足,在本模型中两个样本均值不同,因此可以进行分类,否则无法用此方法. BP 神经网络算法对不确定的问题有自适应和自学习能力,能高精度地逼近连续的非线性函数,很好地协调多种输入信息的关系,从而对未知样本进行分类. 支持向量机方法优势在于,通过分界面能够直观地表现类别间的区分,但通过分类结果与前两种方法的比较,我们可以得知其缺点在于只能用于具有明显差异的小样本间的区分,误差较大.

【参考文献】

[1] 豆丁网,第 19 章神经网络模型,http://www.docin.com/p-392368961.html,2013 年 5 月 20 日.

[2] 道客巴巴,神经网络建模之一,http://www.doc88.com/p-116698859629.html,2013 年 5 月 20 日.

[3] 百度文库,实验 1 Fisher 线性判别实验,http://wenku.baidu.com/view/95a448a9d1f34693daef3e4b.html,2013 年 5 月 20 日.

[4] 道客巴巴,2000 网易杯全国大学生数学建模竞赛题目,http://www.doc88.com/p-713991940876.html,2013 年 5 月 20 日.

[5] 郭显娥,武伟,刘春贵,张景安. 多种 SVM 分类算法的研究. 山西大同大学学报,2010 年 6 月第 26 卷第 3 期.