基于 SVM 和 KNN 的蛋白质耐热性分类

丁彦蕊[1,2],蔡宇杰[2,3],孙 俊[1],须文波[1]
DING Yan-rui[1,2],CAI Yu-jie[2,3],SUN Jun[1],XU Wen-bo[1]

1.江南大学 信息工程学院,江苏 无锡 214122
2.江南大学 工业生物技术教育部重点实验室,江苏 无锡 214036
3.江南大学 生物工程学院,江苏 无锡 214036
1.School of Information Technology,Southern Yangtz University,Wuxi,Jiangsu 214122,China
2.Key Laboratory of Industrial Biotechnology,Wuxi,Jiangsu 214036,China
3.School of biotechnology,Wuxi,Jiangsu 214036,China
E-mail:yanrui_ding76@yahoo.com.cn

DING Yan-rui,CAI Yu-jie,SUN Jun,et al.Classification of protein thermostability using Support Vector Machines and K-Nearest Neighbors.Computer Engineering and Applications,2007,43(16):228-230

Abstract: Regarding amino acid composition as eigenvector,protein thermostability is classified using Support Vector Machines and K-Nearest Neighbors.It is found that the result of using support vector machines is better than K-Nearest Neighbors.The local accuracy and global accuracy are 82.4% and 83.4% respectively.But the local accuracy and global accuracy are 77.6% and 79.9% respectively using K-Nearest Neighbors.The prediction accuracy of two kinds of methods can both prove that the amino acid composition is the main factor that influences the protein thermostability.
Key words: amino acid composition;SVM;KNN;protein thermostability

摘 要:以氨基酸含量为特征向量,研究了 SVM 和 KNN 预测蛋白质耐热性的准确度。结果表明,基于 SVM 的分类效果较好,其局部预测率和全局预测率分别为 82.4%和 83.4%;而基于 KNN 方法的局部预测率和全局预测率分别为 77.6%和 79.9%。两种方法的预测率均表明氨基酸含量是影响蛋白质耐热性的主要因素。
关键词:氨基酸含量;SVM KNN;蛋白质耐热性
文章编号:1002-8331(2007)16-0228-03 **文献标识码**:A **中图分类号**:TP183

1 引言

尽管常温菌和嗜热菌的蛋白质都是由 20 种相同的氨基酸组成,然而它们的耐热性却有很大差别。多数常温蛋白质在 60℃以上会很快失去活性,耐热蛋白质则可以耐受 60℃~120℃甚至更高的温度。长期以来,人们认为氨基酸含量是影响蛋白质耐热性的主要因素[1]。

支持向量机(SVM)是根据统计学习理论提出的一种机器学习方法,而 K-最近邻规则(KNN)是一种典型的监督学习技术,两种方法在分类方面有着广泛的应用。

本文以氨基酸含量为特征向量分别研究了 SVM 和 KNN 预测蛋白质耐热性的准确度,这样不仅可以选择适合于蛋白质耐热性分类的方法,又可以判断氨基酸含量是否是影响蛋白质耐热性的主要因素。

2 材料与方法

2.1 数据库

本文以将微生物进行详细分类的美国国立生物技术信息中心的蛋白质直系同源簇(NCBI Cluster of Orthologous Groups of Proteins COG)数据库[2-4],以及与之对应的 PDB 结构数据库[5],作为数据的来源。

首先从免费的 ftp 服务器上下载(ftp://ftp.ncbi.nlm.nih.gov/pub/COG/)所有全基因组微生物的蛋白质序列。所下载的序列数据以 Fasta 格式存在,仅提供了序列及其在 NCBI 中的 ID 号、COG 号,关于蛋白质序列的信息量很少,因此根据 COG 中提供的 ID 号从 NCBI 中以 GenPept 格式将详细的序列信息下载至本地。根据这些信息最终筛选到 101 868 个常温蛋白质,3 974 个高温蛋白质以及 15 187 个超高温蛋白质序列。

2.2 支持向量机

SVM 是 Vapnik 等[6,7]根据统计学习理论提出的一种机器学习方法。它具有:基于结构风险最小化原则,保证学习机器具有良好的泛化能力;巧妙地解决了算法复杂度与输入向量维数密切相关的问题;应用核技术,将输入空间中的非线性问题,通过非线性函数映射到高维特征空间中,在高维空间中构造线性判别函数;专门针对小样本情况,它的最优解基于已有样本信息,

作者简介:丁彦蕊(1976-),女,讲师,博士,主要从事生物信息学,人工智能方面的研究;蔡宇杰(1973-),男,副教授,博士,主要从事色谱分离,发酵工程方面的研究;孙俊(1971-),男,讲师,副教授,主要从事人工智能优化的研究;须文波(1946-),男,教授,博士生导师,主要从事生物信息学,人工智能及系统控制方面的研究。

而不是样本数趋于无穷大时的最优解;算法最终转化为一个凸优化问题,保证了算法的全局最优性等优点。因此,SVM 有很广泛的应用,如多光谱显微细胞图像分割[8],文本兼类标注[9],遥感影像目标检测中样本的选取[10],预测真核生物 RNA 剪切位点[11],蛋白质折叠识别[12],蛋白质相互作用预测[13]。

对于分类问题,SVM 可以简述为将输入空间中的样本通过某种非线性函数关系映射到一个特征空间中(维数可能较高),使两类样本(可以推广到多类样本)在此特征空间中线性可分,并寻找样本在此特征空间中的最优线性分类超平面。

假定一个训练集有 m 个样本或者是输入向量 $\boldsymbol{X}_i \in R^d (i=1,\cdots,m)$,输出向量为 $\boldsymbol{y}_i \in \{-1,+1\}(i=1,\cdots,m)$,这里,1 代表一类样本的输出值,-1 代表另一类样本的输出值。SVM 就是要寻找一个超平面,使其到两类样本的距离最大。其判别函数为:

$$f(X)=\mathrm{sgn}\left(\sum_{i=1}^{m} y_i\alpha_i\cdot(\Phi(X)\cdot\Phi(X_i))+b\right)=$$

$$\mathrm{sgn}\left(\sum_{i=1}^{m} y_i\alpha_i\cdot(K(X,X_i)+b\right) \tag{1}$$

公式中的 α_i 可以通过解下面的凸优化问题求得。

$$\text{最大化:}W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(X_i,X_j) \tag{2}$$

$$\text{条件:}0\leqslant\alpha_i\leqslant\frac{1}{m},\ \sum_{i=1}^{m}\alpha_i y_i=0 \text{ 和 } \sum_{i=1}^{m}\alpha_i\geqslant v \tag{3}$$

公式(3)中的 $K(X_i,X_j)$ 是核函数,该函数决定了一个使输入向量从 Euclidean 空间 R^d 到高维的 Hilbert 空间的非线性平面。核函数的选取应使其为特征空间的一个点积,即存在函数 ϕ,使 $\phi(X_i)\times\phi(X_j)=K(X_i,X_j)$。常用的核函数有:

线性核函数(Linear Function)

$$K(X_i,X_j)=X_i^{\mathrm{T}}X_j \tag{4}$$

多项式核函数(Polynomial Function)

$$K(X_i,X_j)=(X_i\times X_j+1)^d \tag{5}$$

径向基核函数(Radial Basis Function,RBF)

$$K(X_i,X_j)=\exp(-\gamma\parallel X_i-X_j\parallel^2) \tag{6}$$

Sigmoid 核函数(Sigmoid Function)

$$K(X_i,X_j)=\tanh[b(X_i\times X_j)+c] \tag{7}$$

由于线性核函数是 RBF 核函数的一种特殊形式,RBF 不仅可以处理线性可分问题还可以处理线性不可分问题,本文采用了 RBF 核函数来预测蛋白质耐热性。

公式(3)中利用参数 v 来控制支持向量的数目和误差。参数 v 的范围为 $v\in(0,1]$,上限限制训练误差,下限限制支持向量,这不同于传统的 CSVM[14]。很明显 $v\in[0,1]$比 $C\in[0,\infty]$更容易调节和控制。本章利用 v-SVM 预测蛋白质耐热性,对于研究的数据集,只有核函数 γ 和调节参数 v 是需要选择的。

本文利用 OSU_SVM matlab 工具箱(http://www.ece.osu.edu/~maj/osu_svm/)预测蛋白质耐热性问题,该软件支持 one-against-one 多分类问题。

2.3　K-最近邻规则

KNN 是一种典型的监督学习技术,已经在病毒检测[15]、文本分类[16]等方面有了广泛的应用。要了解 K-NN 规则的工作原理,首先要从最近邻规则介绍[17]。

最近邻规则　假定有 ρ 个类别 $\omega_1,\omega_2,\cdots,\omega_\rho$ 的模式识别问题,每类有表明类别的样本(即训练样本)N_i 个,$i=1,2,\cdots,\rho$。规定 ω_i 类的判别函数为:

$$g_i(x)=\min_i \parallel x-x_i^k \parallel,\ k=1,2,\cdots,N_i \tag{8}$$

其中 x_i^k 的角标 i 表示 ω_i 类,k 表示 ω_i 类 N_i 个样本中的第 k 个。根据公式(8),决策规则可以写为:

若 $g_j(x)=\min_i g_i(x), i=1,2,\cdots,\rho$,则决策 $x\in\omega_j$。

这一决策方法称为最近邻法。其直观解释是相当简单的,就是说对未知样本 x,只要比较 x 与 $N=\sum_{i=1}^{\rho}N_i$ 个已知类别的样本之间的欧氏距离,并决策 x 与离它最近的样本同类。

所谓 KNN 就是取未知样本 x 的 k 个近邻,看这 k 个近邻中多数属于哪一类,就把 x 归于哪一类。具体说就是在 N 个已知样本中,找出 x 的 k 个近邻。设这 N 个样本中,来自 ω_1 类的样本有 N_1 个,来自 ω_2 类的样本有 N_2 个,来自类的样本有个,若 $k_1,k_2,\cdots,k_c$ 分别是 k 个近邻中属于 $\omega_1,\omega_2,\cdots,\omega_\rho$,类的样本数,则可以定义判别函数为:

$$g_i(x)=k_i,\ i=1,2,\cdots,\rho \tag{9}$$

决策规则为:若 $g_j(x)=\max_i k_i$,则决策 $x\in\omega_j$。

以上就是 KNN 法的基本规则。根据以上的讨论可以看出,最近邻法是 K-NN 法当 $k=1$ 时的一个特例

2.4　特征向量提取

氨基酸含量是影响蛋白质耐热性的主要因素,因此蛋白质序列可以表示为如下的特征向量:$\boldsymbol{X}_j^\rho=[x_{j,1}^\rho,x_{j,2}^\rho,\cdots,x_{j,20}^\rho]^{\mathrm{T}}$,式中 ρ 表示蛋白质的类别,j 表示每类蛋白质的样本个数,$x_{j,i}^\rho$ $(i=1,2,\cdots,20)$表示 ρ 类蛋白质第 j 个蛋白质序列中 i 种氨基酸的含量。

2.5　分类方法检测

本文通过 10-fold-cross-validation 对分类方法的预测精度进行检验。所谓10-fold-cross-validation 就是将训练集中的样本近似随机平分为 10 份,依次将每一份取出作为测试集,另外 9 份作为训练集,这样测试集和训练集经过了 10 次轮换,每一次都得到一个预测精度,将 10 次预测精度平均后就得到了当前算法参数下的 cross-validation 预测精度,显然 cross-validation 预测精度最高时的参数就是要选取的最优参数。本文根据 Keun-Joon Park 采用的检验方法[18],通过总预测率和局部预测率来选择参数和衡量分类方法的优劣。

总预测率定义为:

$$TA=\frac{\sum_{i=1}^{\rho}T_i}{N} \tag{10}$$

局部预测率定义为:

$$LA=\frac{\sum_{i=1}^{\rho}P_i}{\rho},\text{其中 } P_i=\frac{T_i}{n_i} \tag{11}$$

公式中的 N 为数据集中所有蛋白质的个数,ρ 为蛋白质的种类(超高温蛋白质、高温蛋白质和常温蛋白质),n_i 为第 i 类蛋白质中蛋白质的个数,T_i 为第 i 类蛋白质中成功预测的蛋白质的个数。

2.6　训练和预测过程

样本两种类的数据严重不平衡时会影响预测结果。由于数

据库中超高温蛋白质,高温蛋白质和常温蛋白质的数量相差比较悬殊,直接预测会降低预测率,考查训练集中三类蛋白质的比例与预测率的关系是十分必要的。将三类蛋白质以一定的比例从数据集中随机选取作为训练集, 对于训练集利用 10 fold cross validation 来检验不平衡数据对预测精度的影响。通过 10 fold cross validation 也可以判断两种分类方法的优劣,从而选择适合预测蛋白质耐热性的机器学习方法及算法参数。

调整参数不仅影响分类机器的复杂度而且影响训练速度。为了更好地解决分类问题,选取最优的调整参数无疑是最重要的。对于 SVM,最优参数 γ 和 v 通过 grid search(网格搜索)法选取[19]。对于 KNN,k 是唯一需要确定的函数,因此参数搜索空间为一维,10-fold-cross-validation 精度最高时的 k 值为最优参数。

最优参数确定后, 对整个训练集进行训练得到分类器,再对数据集中剩余的数据(测试集)进行预测,所得到的预测率就是对三类蛋白质的分类准确度。训练和预测过程见图 1(以 SVM 为例)。

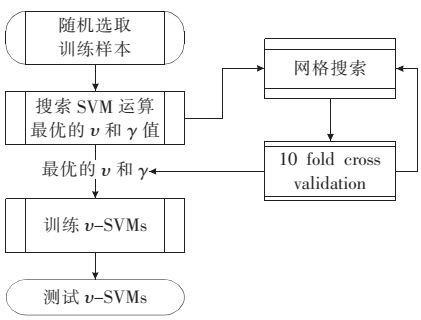

图 1 SVM 训练和预测过程流程图

3 结果与讨论

“原核生物蛋白质序列数据集”中各种蛋白质之间的比例为 25:1:4,显然三类蛋白质数量不平衡,考虑到高温蛋白质只有 3 974 个,训练集中三类蛋白质的比例分别取 1 000:1 000:1 000;2 000:2 000:2 000;3 000:3 000:3 000;4 000:3 000:4 000;5 000:3 000:5 000;6 000:3 000:6 000;7 000:3 000:7 000;8 000:3 000:8 000;9 000:3 000:9 000。把这些训练集分别随机均分为 10 份,按照 10-fold-cross-validation 的方法计算了预测率,并选取了最优参数。

表 1 SVM 预测蛋白质耐热性

训练样本比例	超高温蛋白质预测率	高温蛋白质预测率	常温蛋白质预测率	局部预测率(LA)	全局预测率(TA)	γ	v
1 000:1 000:1 000	80.16	77.37	80.91	79.48	80.74	220	0.5
2 000:2 000:2 000	80.89	78.17	83.60	80.89	83.22	200	0.5
3 000:3 000:3 000	**82.40**	**81.31**	**83.58**	**82.43**	**83.44**	**220**	**0.5**
4 000:3 000:4 000	84.31	75.36	85.41	81.70	85.22	280	0.5
5 000:3 000:5 000	85.75	69.61	86.69	80.68	86.46	140	0.5
6 000:3 000:6 000	87.70	66.84	87.04	80.53	86.92	240	0.5
7 000:3 000:7 000	89.12	65.30	87.34	80.58	87.28	180	0.5
8 000:3 000:8 000	88.85	61.29	88.16	79.44	87.96	180	0.5
9 000:3 000:9 000	89.57	61.40	88.31	79.76	88.14	200	0.5

SVM 和 KNN 对三类蛋白质的分类结果,以及全局预测率和局部预测率分别见表 1 和表 2。

从表 1 可以看出, 当训练样本比例为 3 000:3 000:3 000 时高温蛋白质的预测率最高,同时局部预测率也最高,而超高温蛋白质的预测率也达到了 82.40%, 常温蛋白质的预测率 83.58%。这说明 SVM 从氨基酸含量预测蛋白质耐热性时,三类蛋白质的比例为 3 000:3 000:3 000 时, 既可以保证学习的信息量足够大,也可以消除数据不平衡的影响作用,因此选择训练样本的比例为 3 000:3 000:3 000。与之对应的 γ 和 v 的值分别为 220 和 0.5。

表 2 KNN 预测蛋白质耐热性

训练样本比例	超高温蛋白质预测率	高温蛋白质预测率	常温蛋白质预测率	局部预测率(LA)	全局预测率(TA)	k
1 000:1 000:1 000	73.46	74.38	73.20	73.68	73.26	5
2 000:2 000:2 000	77.92	74.77	75.74	76.14	75.96	9
3 000:3 000:3 000	78.83	77.41	76.57	77.61	76.81	14
4 000:3 000:4 000	80.98	71.97	79.94	77.63	79.97	12
5 000:3 000:5 000	83.61	66.63	81.46	77.23	81.52	9
6 000:3 000:6 000	85.03	64.17	82.06	77.09	82.15	8
7 000:3 000:7 000	87.32	59.75	82.86	76.64	82.98	13
8 000:3 000:8 000	87.92	56.67	83.46	76.02	83.52	9
9 000:3 000:9 000	88.77	55.13	84.25	76.05	84.25	15

从表 2 中同样可以看出高温蛋白质的预测率在 3 000:3 000:3 000 时最高,其值为 77.41%,但局部预测率在 4 000:3 000:4 000 时最高,充分考虑超高温蛋白质,高温蛋白质和常温蛋白质的预测率,以及局部和全局预测率,认为4 000:3 000:4 000 是比较好的训练样本比例,尽管其预测率与3 000:3 000:3 000 时相差不大,此时的最优 k 值为 12。

通过表 1 和表 2 可以看出无论是何种预测率,不论哪种训练样本比例,SVM 的预测率都高于 KNN, 这说明 SVM 是更适合于蛋白质耐热性预测的机器学习方法。

4 结束语

不管是何种蛋白质只要是耐热的,它们在某一方面就有相同的特性,共性的特性使它们都具有耐高温的特点。从预测率可以看出蛋白质一级结构对蛋白质耐热性的影响最大。

尽管耐热微生物是获得热稳定蛋白质的主要来源,但对于那些不能在嗜热微生物中发现的酶类,单点突变和基因敲除可以将常温酶改造为耐热酶, 在实验之前可以利用 SVM 进行预测,既可以缩短试验成本又可以节省时间。

(收稿日期:2006 年 9 月)

参考文献:

[1] Vieille C,Zeikus G J.Hyperthermophilic enzymes:sources,uses,and molecular mechanisms for thermostability[J].Microbiol Mol Biol Rev,2001,65(1):1-43.

[2] Tatusov R L,Koonin E V,Lipman D J.A genomic perspective on protein families[J].Science,1997,278(5338):631-637.

[3] Tatusov R L,Galperin M Y,Natale D A,et al.The COG database:a tool for genome-scale analysis of protein functions and evolution[J].Nucleic Acids Res,2000,28(1):33-36.

[4] Tatusov R L,Natale D A,Garkavtsev I V,et al.The COG database:new developments in phylogenetic classification of proteins from complete genomes[J].Nucleic Acids Res,2001,29(1):22-28.

[5] Berman H M,Westbrook J,Feng Z,et al.The protein data bank[J].Nucleic Acids Res,2000,28(1):235-242.

[6] Vapnik V.The nature of statistical learning theory[M].New York:Springer,1995.

(下转 237 页)