文章编号:1001 - 9081(2010)11 - 2959 - 03

基于核 Fisher 判别分析的蛋白质氧链糖基化位点的预测

杨雪梅,李世鹏
(咸阳师范学院,数学与信息科学学院,陕西 咸阳 712000)
(xmyang412@gmail.com)

摘　要:以各种窗口长度的蛋白质样本序列为研究对象,实验样本用稀疏编码方式编码,使用核 Fisher 判别分析(KFDA)的方法来预测蛋白质氧链糖基化位点。首先通过非线性映射(由核函数隐含定义)将样本映射到特征空间,然后在特征空间中用 Fisher 判别分析进行分类。进一步,用多数投票策略对各种窗口下的分类器进行组合以综合多个窗口的优势。实验结果表明,使用组合 KFDA 的方法预测的效果优于 FDA 和 PCA 以及单个 KFDA 分类器的预测效果,预测准确率为 86.5%。

关键词:糖基化;蛋白质;核 Fisher 判别分析;特征
中图分类号: TP311.13;TP391.4　　**文献标志码**:A

Prediction of O-glycosylation sites in protein sequence by kernel Fisher discriminant analysis

YANG Xue-mei, LI Shi-peng
(*School of Mathematics and Information Science, Xianyang Normal University, Xianyang Shaanxi* 712000, *China*)

Abstract: To predict the O-glycosylation sites in protein sequence, the method of Kernel Fisher Discriminant Analysis (KFDA) was proposed under various window sizes. Encoded by the sparse coding, the samples were first mapped onto a feature space implicitly defined by a kernel function, and then they were classified into two classes in the feature space by Fisher discriminant analysis. Furthermore, the majority-vote scheme was used to combine all the pre-classifiers to improve the prediction performance. The results indicate that the performance of ensembles of KFDA is better than that of FDA, PCA and pre-classifier. The prediction accuracy is about 86.5%.

Key words: glycosylation; protein; Kernel Fisher Discriminant Analysis (KFDA); feature

《计算机应用》Journal of Computer Applications

0　引言

糖基化是哺乳动物细胞膜合成过程中蛋白质翻译后修饰的重要步骤之一,它具有重要的生物功能。糖基化过程的实现有 4 种形式:1)发生在丝氨酸残基(S)或苏氨酸残基(T)上的氧链糖基化(O-linked);2)发生在天冬酰胺酸残基上的氮链糖基化(N-linked);3)发生在色氨酸残基上的碳链糖基化(C-linked);4)GPI。在本文中只研究第一种。研究表明,并不是所有的 S 或 T 上都会被加上氧链,大约 10% ~30% 的蛋白质不能被氧链糖基化。有很多因素都会影响这个过程,因而氧链糖基化位点的预测问题对于揭示蛋白质的生物学功能就具有重要的意义。

近年来,许多基于人工神经网络(Artificial Neural Network, ANN)、支持向量机(Support Vector Machine, SVM)[1-4]的计算机方法被用来对氧链糖基化位点进行预测,预测准确率达 70%。文献[5]中使用了一种新的蛋白质生物信息处理工具 CKSAAP_OGlySite 来预测氧链糖基化位点,即用基于 *K*-空间氨基酸对组成的编码方式,并借助于 SVM,获得了较好的预测准确率,分别为 81.4%(S)和 83.1%(T)。

文献[6-7]中,使用了主成分分析和马氏距离及混合判别分析的方法来预测氧链糖基化位点,准确率达 82%。

由于核 Fisher 判别分析(Kernel Fisher Discriminant Analysis, KFDA)[8-9]能提取数据的非线性特征,提高分类准确率。本文将使用 KFDA 的方法来预测氧链糖基化位点。样本首先被一个非线性映射映射到特征空间,然后在特征空间中用 Fisher 判别分析进行分类。

1　蛋白质序列与编码

本文用到的蛋白质数据来自糖基化数据库 Uniprot v8.0(见 http://www.ebi.uniprot),包括了 99 种哺乳动物的蛋白质序列共 2000 个,每个序列包含一些丝氨酸和苏氨酸的残基,并对该残基是否糖基化做了标注。每个序列被一个窗口(窗口长度为 w)截成一些以 S 或 T 为中心的子序列。把糖基化的序列叫做正序列(positive),把未糖基化的序列叫做负序列(negative)。

用这些长度为 $w-1$ 的子序列(去掉中心的 S 或 T)来做实验分析。对于蛋白质序列有许多编码的方法,如稀疏编码、5-字母编码、羟基编码和基于物理性质的编码。在本文中用稀疏编码方式。就是把序列中的每个氨基酸残基或空位点用 21 位二进制数 0 或 1 表示,如,氨基酸残基 I 用 100000000000000000000 来表示,氨基酸残基 V 用 010000000000000000000 表示。这样一个子序列的稀疏编码的长度(即样本向量的维数)就是 $(w-1)\times 21$。

各类序列的样本个数总结如表 1。因为负序列的个数比正序列的个数多得多,所以在实验时,随机地从每一类里挑选 100 个样本作为训练样本,50 个作为测试样本。

收稿日期:2010 - 06 - 01;**修回日期**:2010 - 08 - 12。　**基金项目**:陕西省自然科学基金资助项目(2010JQ1013);陕西省教育厅科学研究计划项目(2010JK896;09JK809);咸阳师范学院专项科研基金资助项目(07XSYK107);咸阳师范学院大学生科研训练项目(09101)。
作者简介:杨雪梅(1969 -),女,陕西商洛人,副教授,主要研究方向:模式识别;　李世鹏(1988 -),男,陕西咸阳人,主要研究方向:模式识别。

表 1 实验样本

序列	总数	训练样本数	测试样本数
Positive S	174	100	50
Positive T	292	100	50
Negative S	693	100	50
Negative T	841	100	50

2 核 Fisher 判别分析及预测算法

2.1 Fisher 判别分析(FDA)

FDA 是一种基于统计的分类方法,它能抓住数据的判别信息。它通过线性变换(式(1))使得变换后的样本同时具有最大的类间离散度和最小的类内离散度,从而将各类样品很好地分开。

$$y = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \tag{1}$$

其中 $\boldsymbol{X}$ 是训练样本。$\boldsymbol{W}$ 是如下问题的解:

$$\max J_F(\boldsymbol{W}) = \frac{\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{W}} \tag{2}$$

$\boldsymbol{S}_b$ 是类间离散度矩阵:

$$\boldsymbol{S}_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}} \tag{3}$$

$\boldsymbol{S}_w$ 是总类内离散度矩阵:

$$\boldsymbol{S}_w = \sum_{\boldsymbol{X}\in \boldsymbol{X}_1} (\boldsymbol{X} - \boldsymbol{m}_1)(\boldsymbol{X} - \boldsymbol{m}_1)^{\mathrm{T}} + \sum_{\boldsymbol{X}\in \boldsymbol{X}_2} (\boldsymbol{X} - \boldsymbol{m}_2)(\boldsymbol{X} - \boldsymbol{m}_2)^{\mathrm{T}} \tag{4}$$

$\boldsymbol{S}_b$ 和 $\boldsymbol{S}_w$ 均为对称半正定矩阵。$\boldsymbol{m}_1$、$\boldsymbol{m}_2$ 分别是两类的平均值。

通过求解(式(2))得到最优的变换方向:

$$\boldsymbol{W}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{5}$$

由于 FDA 是一种线性变换,而原始数据往往有非线性的特征,因此 FDA 对于有非线性特征的数据不能有效地区分。

2.2 KFDA

KFDA 是基于核理论的方法,它能有效地解决非线性特征提取问题。它首先将样本通过一个非线性映射映射到特征空间,然后在特征空间中完成 FDA,从而隐含地实现了原输入空间的非线性判别。

设 Φ 是从输入空间到特征空间的非线性映射,$\Phi: \boldsymbol{X} \to F$,在该映射之下,输入空间中的向量集 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N$ 被映射为特征空间中的向量集 $\Phi(\boldsymbol{X}_1), \Phi(\boldsymbol{X}_2), \cdots, \Phi(\boldsymbol{X}_N)$,则在特征空间中可定义两类样本的均值向量为:

$$\boldsymbol{m}_i^{\Phi} = \left(\frac{1}{N_i}\right) \sum_{\boldsymbol{X}\in \boldsymbol{X}_I} \Phi(\boldsymbol{X});\ i = 1,2 \tag{6}$$

样本类间离散度矩阵为:

$$\boldsymbol{S}_b^{\Phi} = (\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_2^{\Phi})(\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_2^{\Phi})^{\mathrm{T}} \tag{7}$$

总类内离散度矩阵为:

$$\boldsymbol{S}_w^{\Phi} = \sum_{i=1,2} \sum_{\boldsymbol{X}\in \boldsymbol{X}_I} (\Phi(\boldsymbol{X}) - \boldsymbol{m}_i^{\Phi})(\Phi(\boldsymbol{X}) - \boldsymbol{m}_i^{\Phi})^{\mathrm{T}} \tag{8}$$

设投影直线的方向为 $\boldsymbol{W}$,则投影后应有:

$$\max J_F(\boldsymbol{W}) = \frac{\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b^{\Phi} \boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w^{\Phi} \boldsymbol{W}} \tag{9}$$

由式(9)解得的最优投影方向为:

$$\boldsymbol{W}^* = (\boldsymbol{S}_w^{\Phi})^{-1}(\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_2^{\Phi}) \tag{10}$$

$\Phi(\boldsymbol{X})$ 在 $\boldsymbol{W}$ 上的投影为:

$$y = \boldsymbol{W}^{*\mathrm{T}} \Phi(\boldsymbol{X}) \tag{11}$$

考虑到 $\boldsymbol{W}$ 可由 $\Phi(\boldsymbol{X}_1), \Phi(\boldsymbol{X}_2), \cdots, \Phi(\boldsymbol{X}_N)$ 线性表示,即有:

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \Phi(\boldsymbol{X}) \tag{12}$$

结合式(6)和(12),有:

$$\bar{y}_i = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{m}_i^{\Phi} = \frac{1}{N_i} \sum_{j=1}^{N} \sum_{k=1}^{N_i} \boldsymbol{\alpha}_j k(\boldsymbol{X}_j, \boldsymbol{X}_k^{\omega_i}) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_i \tag{13}$$

其中:$i = 1,2$;$j = 1,2,\cdots,N$;$k(\cdot,\cdot) = \Phi(\cdot)\cdot\Phi(\cdot)$ 是核函数。$\boldsymbol{M}_i$ 是 $N \times 1$ 矩阵,且:

$$(\boldsymbol{M}_i)_j = \frac{1}{N_i} \sum_{k=1,}^{N_i} k(\boldsymbol{X}_j, \boldsymbol{X}_k^{\omega_i}) \tag{14}$$

结合式(7)和(13),有:

$$\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b^{\Phi} \boldsymbol{W} = \boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_2^{\Phi})(\boldsymbol{m}_1^{\Phi} - \boldsymbol{m}_2^{\Phi})^{\mathrm{T}} \boldsymbol{W} = \boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{M}_1 - \boldsymbol{M}_2)(\boldsymbol{M}_1 - \boldsymbol{M}_2)^{\mathrm{T}} \boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{\alpha} \tag{15}$$

式(15)中定义:

$$\boldsymbol{M} = (\boldsymbol{M}_1 - \boldsymbol{M}_2)(\boldsymbol{M}_1 - \boldsymbol{M}_2)^{\mathrm{T}} \tag{16}$$

结合式(8)和(12),有:

$$\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w^{\Phi} \boldsymbol{W} = \boldsymbol{W}^{\mathrm{T}} \sum_{i=1,2} \sum_{\boldsymbol{X}\in \boldsymbol{X}_i} (\Phi(\boldsymbol{X}) - \boldsymbol{m}_i^{\Phi})(\Phi(\boldsymbol{X}) - \boldsymbol{m}_i^{\Phi})^{\mathrm{T}} \boldsymbol{W} = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{\alpha} \tag{17}$$

其中:

$$\boldsymbol{H} = \sum_{i=1,2} \boldsymbol{K}_i (\boldsymbol{I} - \boldsymbol{L}_i) \boldsymbol{K}_i^{\mathrm{T}} \tag{18}$$

$\boldsymbol{K}_i$ 为 $N \times N_i (i = 1,2)$ 矩阵,并满足:

$$(\boldsymbol{K}_i)_{p,q} = k(\boldsymbol{X}_p, \boldsymbol{X}_q^{\omega_i})$$
$$p = 1,2,\cdots,N; q = 1,2,\cdots,N_i \tag{19}$$

称 $\boldsymbol{K}_i$ 为第 i 类核矩阵,$\boldsymbol{I}$ 为 $N_i \times N_i$ 大小的单位阵;$\boldsymbol{L}_i$ 为 $N_i \times N_i$ 大小的矩阵,其所有元素都为 $\frac{1}{N_i}$。

结合式(16)和(17),式(9)等价为:

$$\max J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{\alpha}} \tag{20}$$

众所周知,$\boldsymbol{\alpha}$ 实质是矩阵 $\boldsymbol{H}^{-1}\boldsymbol{M}$ 的最大特征值对应的特征向量。$\boldsymbol{\alpha}$ 可以直接如下求得:

$$\boldsymbol{\alpha} = \boldsymbol{H}^{-1}(\boldsymbol{M}_1 - \boldsymbol{M}_2) \tag{21}$$

特征空间中 $\Phi(\boldsymbol{X})$ 在 $\boldsymbol{W}$ 上的投影变换为 $k(\cdot,\boldsymbol{X})$ 在 $\boldsymbol{\alpha}$ 上的投影,即:

$$y = \boldsymbol{W}^{\mathrm{T}} \Phi(\boldsymbol{X}) = \sum_{j=1}^{N} \boldsymbol{\alpha}_j k(\boldsymbol{X}_j, \boldsymbol{X}) \tag{22}$$

设 $\tilde{\boldsymbol{m}}_i^{\Phi}$ 为投影后的各类别的平均值,则:

$$\tilde{\boldsymbol{m}}_i^{\Phi} = \frac{1}{N_i} \sum_{y\in \omega_i} y \tag{23}$$

分界阈值点 y_0 可选为:

$$y_0 = \frac{N_1 \tilde{\boldsymbol{m}}_1^{\Phi} + N_2 \tilde{\boldsymbol{m}}_2^{\Phi}}{N_1 + N_2} \tag{24}$$

将待测样品 $\boldsymbol{X}$ 进行投影得到 y,若 $y > y_0$,则 $\boldsymbol{X} \in \boldsymbol{\omega}_1$,若 $y < y_0$,则 $\boldsymbol{X} \in \boldsymbol{\omega}_2$。

2.3 实现步骤

1) $(\boldsymbol{M}_i)_j = \frac{1}{N_i} \sum_{k=1,}^{N_i} k(\boldsymbol{X}_j, \boldsymbol{X}_k^{\omega_i})(i = 1,2; j = 1,2,\cdots,N)$,

选用如下的核函数:

a) 多项式核函数:

$k(\boldsymbol{X}_i, \boldsymbol{X}) = [\langle \boldsymbol{X}, \boldsymbol{X}_i \rangle + c]^d$

b) 高斯径向基核函数(Radial Basis Function, RBF):

$$k(\boldsymbol{X}_i, \boldsymbol{X}) = \exp\left(-\frac{\| \boldsymbol{X}_i - \boldsymbol{X} \|^2}{2\sigma^2}\right)$$

2) 计算 $\boldsymbol{H} = \sum_{i=1,2} \boldsymbol{K}_i (\boldsymbol{I} - \boldsymbol{L}_i) \boldsymbol{K}_i^{\mathrm{T}}$。

3) 计算 $\boldsymbol{\alpha} = \boldsymbol{H}^{-1}(\boldsymbol{M}_1 - \boldsymbol{M}_2)$。

4) 求训练集内各类样品的投影:

$$y_j = \boldsymbol{W}^{\mathrm{T}} \Phi(\boldsymbol{X}_j) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i k(\boldsymbol{X}_i, \boldsymbol{X}_j); j = 1,2,\cdots,N$$

《计算机应用》 Journal of Computer Applications

5) 求均值 $\tilde{m}_i^{\Phi} = \frac{1}{N_i}\sum_{y_j \in \omega_i} y_j$。

6) 求阈值点 y_0。

7) 计算待定样品 X 的投影点 y。

8) 根据决策规则分类。

3　预测及结果分析

用各种窗口长度 (w = 5,7,9,11,21,31,41,51) 的蛋白质序列做实验,则原始数据的维数为 $21(w-1) = 84,126,168,210,420,630,840,1050$。训练样本的总数为 $M = 400$,则核空间的维数是 400。分别用了主成分分析(Principal Component Analysis, PCA)、FDA 和 KFDA 的方法进行预测分类,核函数取多项式核函数和 RBF。测试了 200 个样本,结果见表 2 ~4 和图 1。

表 2　FDA 和 PCA 预测准确率比较　　%

窗口长度	FDA	PCA	窗口长度	FDA	PCA
5	73.0	76.5	21	67.5	79.5
7	72.5	77.5	31	72.5	75.0
9	77.5	80.5	41	70.5	76.5
11	78.0	82.0	51	76.0	77.5

表 3　预测准确率(方法: KFDA, 核函数: RBF)　　%

窗口长度	参数(σ)	negative	positive	平均
5	2.6	76	88	82.0
7	1.6	81	84	82.5
9	1.8	82	85	83.5
11	2.0	85	84	84.5
21	3.2	86	83	84.5
31	3.5	91	75	83.0
41	4.5	90	72	81.0
51	4.4	90	71	80.5

表 4　预测准确率(方法: KFDA, 核函数: 多项式)　　%

窗口长度	参数(c,d)	negative	positive	平均
5	$c = 4, d = 7$	75	85	80.0
7	$c = 4, d = 5$	83	82	82.5
9	$c = 8, d = 4$	82	85	83.5
11	$c = 7, d = 4$	85	81	83.0
21	$c = 4, d = 2$	86	83	84.5
31	$c = 5, d = 2$	87	76	81.5
41	$c = 6, d = 2$	90	72	81.0
51	$c = 3, d = 2$	91	71	81.0

由实验结果可以看出:1)使用 KFDA 方法的预测准确率高于使用 PCA 和 FDA 方法的预测准确率,这归功于 KFDA 的非线性特征的提取能力。2)当窗口长度为 9,11,21 时,预测准确率较高。这是因为,若窗口长度过小,比如 5,7,由于在实验时去掉了中心位置的 S 或 T,这时一些正序列和一些负序列是相同的向量,从而造成错误的判断;若窗口长度过大,比如 41,51,这时输入向量的维数较高,使得特征过于复杂而不易提取,而相对来说,400 个样本也显得过少。因此若用单个 KFDA 分类器,则在选用样本时,样本的窗口长度不宜过小,也不宜过大。3)当使用多项式核函数时,对于不同的窗口长度,最合适的阶数 d 也不同,当窗口长度较小时,d 较大;当窗口长度较大时,d 较小。这是因为小窗口的样本需要用较高阶的多项式核函数来提取它的特征。4) 图 1 显示了窗口长度为 11 时,用 RBF 作为核函数的 KFDA 方法的预测准确率随参数 σ 的变化情况,当 $\sigma = 2$ 时获得了较好的预测准确率,因此 2 是最合适的参数;其他窗口长度下 σ 对结果的影响情况类似。由表 3 看出,在不同的窗口长度下,最合适的 σ 也不同,但总的趋势是,随着窗口长度的增加,参数 σ 也增大,因为 σ 是一个宽度参数。5)在表 3 ~4 中,分别列出了两类 positive 和 negative 各自的预测准确率,可以看出,无论是用多项式核函数,还是 RBF,随着窗口长度的增大,positive 的预测准确率都呈现出减小的趋势,而 negative 则正好相反,这说明小窗口对 positive 的识别率较高,而大窗口对 negative 的识别率较高。为了充分利用这些信息,综合多个窗口的优势,考虑设计一个集成分类器,对 8 个不同窗口长度的子分类器采用多数投票策略进行组合,以提高预测准确率。RBF 和多项式预测结果的准确率均为 86.5%。

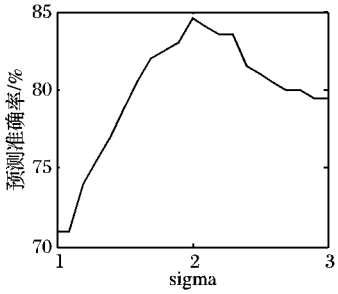

图 1　σ 对结果的影响(窗口长度为 11)

4　结语

用 KFDA 的方法实现了蛋白质氧链糖基化位点的预测。用各种窗口长度(w = 5,7,9,11,21,31,41,51)的蛋白质序列做实验。首先通过一个由核函数隐含定义的非线性映射将样本映射到特征空间,然后在核特征空间中用 FDA 进行分类(预测)。最后用多数投票策略对多个窗口长度下的分类结果进行综合判断。实验结果表明,多窗口下组合 KFDA 方法的预测结果优于 FDA、PCA 方法及单窗口下 KFDA 的预测结果,预测准确率为 86.5%。

《计算机应用研究》 Journal of Computer Applications

参考文献:

[1] NISHIKAWA I, SAKAMOTO H, NOUNO I, *et al*. Prediction of the O-glycosylation sites in protein by layered neural networks and support vector machines [M]. Berlin: Springer, 2006: 953 -960.

[2] SASAKI K, NAGAMINE N, SAKAKIBARA Y. Support vector machines prediction of N- and O-glycosylation sites using whole sequence information and subcellular localization[J]. IPSJ Transactions on Bioinformatics, 2009(2): 25 -35.

[3] JULENIUS K, MOLGAARD A, GUPTA R, *et al*. Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites[J]. Glycobiology, 2004(15): 153 -164.

[4] LI S. Predicting O-glycosylation sites in mammalian proteins by using SVMs [J]. Computational Biology and Chemistry, 2006(30): 203 -208.

[5] CHEN YONG-ZI. Prediction of mucin-type O-glycosylation sites in mammaliam protein using the composition of k-spaced amino acid pairs[J]. BMC Bioinformatics, 2008(2): 101 -107.

[6] YANG XUE-MEI, CHEN YEN-WEI, ITO M, *et al*. Principal component analysis of O-linked glycosylation sites in protein sequence [C]// Third International Conference on IIHMSP. Washington, DC: IEEE: 2007: 121 -126.

[7] YANG XUE-MEI. Prediction of O-linked glycolsylation sites in protein sequence by PCA-LDA [C]// Proceedings of the 2009 Ninth International Conference on Hybrid Intelligent Systems. Washington, DC: IEEE Computer Society, 2009: 158 -161.

[8] 杨淑莹. 模式识别与智能计算: Matlab 技术实现[M]. 北京: 电子工业出版社, 2008: 111 -120.

[9] SHAWE-TAYLOR J, CRISTIANINI N. Kernel methods for pattern analysis[M]. 北京: 中国机械工业出版社, 2005: 137.