基于模糊核匹配追寻的特征模式识别

李 青[1),2)] 焦李成[2)] 周伟达[2)]

[1)](南京电子技术研究所 南京 210013)

[2)](西安电子科技大学智能信息处理研究所 西安 710071)

摘 要 核匹配追寻算法是近年来新兴的模式识别方法,在处理非线性及高维模式识别问题中表现出了突出的优点.传统的核匹配追寻在处理模式识别的问题中平等地对待所有样本,最终的判决函数是针对所有样本的一个平等综合考虑,要求总识别误差尽可能小,并不能对某一类指定样本进行针对性识别,然而实际应用中经常会碰到这样的情况:要求对某一类样本的识别精度很高,尤其是对于非平衡样本中或者对于具有时间属性的样本序列,由于标准核匹配追寻学习机自身的局限性,使其不能有效地处理这些问题.文中针对这些问题,提出了模糊核匹配追寻学习机,预先根据分类的要求对每个样本做出了不同的重要性定义,学习机根据重要性不同,对样本进行程度不同的学习,最终得到基于问题的判决——对重要样本保持很高的分类精度;最后通过实际的仿真实验证明了模糊匹配追寻的有效性及可行性.

关键词 机器学习;核匹配追寻;模糊核匹配追寻;时间序列; 特征目标识别

中图法分类号 TP18 **DOI 号**: 10.3724/SP.J.1016.2009.01687

Pattern Recognition Based on the Fuzzy Kernel Matching Pursuit

LI Qing[1),2)] JIAO Li-Cheng[2)] ZHOU Wei-Da[2)]

[1)](*Nanjing Institute of Electronic Technology*, *Nangjing* 210013)

[2)](*Institute of Intelligent Information Processing*, *Xidian University*, *Xi'an* 710071)

Abstract Kernel Matching Pursuit (KMP), a novel method of the pattern recognition, presents excellent performance in solving the problems with small sample, nonlinear and local minima. KMP has been proposed to provide a good generalization performance for both classes, yet the classification precision of some important data can't be classified precisely. Because the decision function found by KMP is the synthetic consideration results of all the data, it has greatly limited its use in many practical problems, such as time series identification and unbalanced data classification. In this paper, an fuzzy kernel matching pursuit machine is (FKMP) proposed, which can classify the appointed important samples much more precisely according to the predefined importance of the data. Lots of experiments have been given in the paper to prove the feasibility and validation of the fuzzy kernel matching pursuit machine.

Keywords machine learning; kernel matching pursuit; fuzzy kernel matching pursuit; time series identification; unbalanced data classification

1 引 言

核匹配追寻(Kernel Matching Pursuit,KMP)是近年来新提出的一种模式识别方法,它首先通过核映射将训练样本映射成为一组基原子字典,通过贪婪算法在基函数字典中寻找一组基原子的线性组合来最小化损失函数,该线性组合即为所要求解的

收稿日期:2006-05-08;最终修改稿收到日期:2009-06-22.**李 青**,男,1979 年生,博士,工程师,研究方向为机器学习、模式识别及统计学习理论. E-mail: kingdomyangfan@hotmail.com. **焦李成**,男,1959 年生,博士,教授,博士生导师,研究领域为非线性理论、神经网络、数据挖掘、进化算法与子波理论. **周伟达**,男,1974 年生,博士,副教授,主要研究方向包括机器学习、模式识别等.

判别函数.核匹配追寻分类器的分类性能几乎可以达到支撑矢量机的分类性能,同时与其他经典的核机器算法相比,具有更为稀疏的解[1].

然而在实际问题中,存在这样几种情况:(1)对指定类别的识别精度有特殊性要求——在识别问题中,一类样本(或某些样本)比另一类样本(或其余样本)更为重要,要求对这些重要样本的识别精度要高(例如对癌细胞的检测、非法入侵的检测);(2)所获得的样本是具有特征时间属性的,也就是说,在某些特定的问题中,某一时间段内的样本相比其它样本具有更为重要的意义,这就需要对处于这一时间段内的样本给予特殊的对待,使得这些样本对最终的判决起到更为重要的作用;(3)非平衡样本的识别,在很多实际问题中,两类样本的个数是不平衡的,尤其是当所采得的特征样本(或弱势样本)相对于另一类样本很少时,对弱势样本的识别就变得非常困难,由于传统核匹配追寻的最终决策是针对整个样本集做出的综合考虑,这就使得学习机弱势样本识别很难.

虽然核匹配追寻已经成功地应用于许多领域,如人脸识别、手写体识别、笔记身份鉴定、数据挖掘等[2-3],然而,传统的核匹配追寻在处理模式识别的问题中平等地对待所有的样本,最终的求解是对错分误差和分类间隔进行折中的结果,它可以对两类样本做出平等综合的考虑,要求总识别误差尽可能小,并不能对某一类或某一些指定的样本进行针对性的识别,这就限制了核匹配追寻在这些有特殊要求问题中的应用.

本文认真分析了核匹配追寻的原理,提出了模糊核匹配追寻,根据样本之间的重要性,对每个样本分别设定不同的模糊因子,使得学习机训练出针对目标样本的决策,进一步扩展了核匹配追寻的应用范围.最后,通过实际的实验证明了模糊核匹配追寻的可行性及有效性.

2 核匹配追寻

2.1 基本匹配追寻算法

给定 l 个观测点$\{x_1,\cdots,x_l\}$,相应的观测值为$\{y_1,\cdots,y_l\}$.匹配追寻的基本思想是:在一个高度冗余的字典(dictionary)空间 D 中将观测值$\{y_1,\cdots,y_l\}$分解为一组基函数的线性组合,其中字典 D 是定义在希尔伯特空间中的一组基函数[2,4].假定字典包含 M 个基函数:

$$D=\{g_m\},\ m=1,2,\cdots,M \tag{1}$$

同时,定义损失函数(亦称为重构误差):

$$\|\boldsymbol{R}_N\|^2=\|\boldsymbol{y}-\boldsymbol{f}_N\|^2 \tag{2}$$

其中,$\boldsymbol{R}_N$ 称为残差,$f_{N,j}=\sum_{i=1}^{N}\alpha_i\boldsymbol{g}_i(\boldsymbol{x}_j)$ 是对 $j(j=1\sim l)$ 个观测点的观测值 y_j 的逼近.匹配追寻算法在每一步的迭代中从字典中寻找一个基函数 $\boldsymbol{g}_{N+1}$ 及其相应的系数 α_{N+1},使得当 $f_{N+1}=f_N+\alpha_{N+1}\boldsymbol{g}_{N+1}$ 时,当前的残差能量 $\|\boldsymbol{R}_{N+1}\|^2$ 最小,即

$$(\alpha_{N+1},\boldsymbol{g}_{N+1})=\arg\min_{\alpha\in R,\boldsymbol{g}\in D}\|\boldsymbol{R}_{N+1}\|^2=\arg\min_{\alpha\in R,\vec{g}\in D}\|\boldsymbol{R}_N-\alpha\boldsymbol{g}\|^2 \tag{3}$$

由匹配追寻算法[5],

$$\boldsymbol{g}_{N+1}=\arg\min_{\boldsymbol{g}\in D}\left(\frac{\langle\boldsymbol{g},\boldsymbol{R}_N\rangle}{\|\boldsymbol{g}\|}\right)^2 \tag{4}$$

$$\alpha_{N+1}=\left(\frac{\langle\boldsymbol{g}_{N+1},\boldsymbol{R}_N\rangle}{\|\boldsymbol{g}_{N+1}\|^2}\right) \tag{5}$$

其中,$\langle\cdot,\cdot\rangle$表示两个向量的点积,$\|\cdot\|$表示向量的二范数.

由上可知,匹配追寻实际上采用了贪婪算法,每次迭代都是从字典中查找与当前残差相关系数最大的基函数分量,随着分解次数的增加,式(5)右端基函数向量的线性组合理论上可以任意地逼近原始观测值,但是通常在满足某种精度条件时就终止了,如残差能量低于某一阈值,或者当基函数的个数大于预先设定的值.

2.2 后拟合匹配追寻算法

基本匹配追寻算法在每一步的优化迭代中,针对当前残差寻找与之相关系数最大的基函数 $\boldsymbol{g}_{m_N}$ 及其系数 α_N,这样,观测值在第 N 代的逼近为

$$\boldsymbol{f}_N=\sum_{k=1}^{N-1}\alpha_k\boldsymbol{g}_{m_k}+\alpha_N\boldsymbol{g}_{m_N} \tag{6}$$

然而,当增加 $\alpha_N\boldsymbol{g}_{m_N}$ 后,匹配追寻在第 N 代对观测值的逼近并不一定是最优的;可以通过后拟合的方法修正 $\boldsymbol{f}_N$,使其进一步逼近观测值[3].所谓后拟合,就是增加 $\alpha_N\boldsymbol{g}_{m_N}$ 项后,重新调整系数 $\alpha_1,\alpha_2,\cdots,\alpha_N$,使得当前的残差能量最小,即

$$\begin{aligned}\alpha_1,\cdots,\alpha_N&=\arg\min_{\alpha_1,\cdots,\alpha_N}\|\boldsymbol{f}_N-\boldsymbol{y}\|^2\\&=\arg\min_{\alpha_1,\cdots,\alpha_N}\left\|\sum_{k=1}^{N}\alpha_k\boldsymbol{g}_k-\boldsymbol{y}\right\|^2\end{aligned} \tag{7}$$

上式的优化过程是一个非常耗时的计算,通常采用折中的方法:匹配追寻算法在迭代运算数步后进行一次后拟合[2].

2.3　核匹配追寻

核匹配追寻实际上是将匹配追寻应用于机器学习问题中的一个非常简单的思想:采用核方法生成函数字典[1].

给定核函数 $K:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$,利用观测点$\{x_1,\cdots,x_l\}$处的核函数值生成函数字典:$D=\{g_i=K(\cdot,x_i)\mid i=1,\cdots,l\}$.

核方法的应用受启发于机器学习方法中的支撑矢量机;在支撑矢量机中,应用的核函数要满足Mercer条件[6-7],然而在匹配追寻中,核函数不必满足次条件,并且,可以在生成函数字典时同时采用多个核函数.通常采用的核函数有[8-9]:

(1) 多项式核. $K(x,x_i)=[(x,x_i)+1]^d$;

(2) 径向基核. $K(x,x_i)=\exp(-\|x-x_i\|/2p)$;

(3) Sigmoid 核. $K(x,x_i)=S(v(x,x_i)+c)$.

2.4　损失函数的拓展

基本的匹配追踪算法采用的损失函数是能量损失函数(即平方损失函数),可以通过梯度下降法将匹配追踪的损失函数进行拓展,使学习机能够对任意给定的损失函数进行学习.

假设损失函数 $L(y_i,f_n(\boldsymbol{x}_i))$,当观测值为 y_i时计算预测值 $f_n(\boldsymbol{x}_i)$的残差 $\widetilde{\boldsymbol{R}}_n$定义如下[1]:

$$\widetilde{\boldsymbol{R}}_n=\left(-\frac{\partial L(y_1,f_n(\boldsymbol{x}_1))}{\partial f_n(\boldsymbol{x}_1)},\cdots,-\frac{\partial L(y_l,f_n(\boldsymbol{x}_l))}{\partial f_n(\boldsymbol{x}_l)}\right) \tag{8}$$

那么,由匹配追踪算法,在每一次迭代中所要寻求的最优基函数为

$$\boldsymbol{g}_{i+1}=\arg\max_{\boldsymbol{g}\in D}\left|\frac{\langle\boldsymbol{g}_{i+1},\widetilde{\boldsymbol{R}}_i\rangle}{\|\boldsymbol{g}_{i+1}\|}\right| \tag{9}$$

对应此最优基函数的系数 α_{i+1}为

$$\alpha_{i+1}=\arg\min_{\alpha\in R}\sum_{k=1}^{l}L(y_k,\boldsymbol{f}_i(\boldsymbol{x}_k)+\alpha\boldsymbol{g}_{i+1}(\boldsymbol{x}_k)) \tag{10}$$

此时,后拟合即是进行如下的优化过程:

$$\alpha_{1,\cdots,i+1}^{(i+1)}=\arg\min_{(\alpha_1,\cdots,i+1)\in R^{i+1}}\sum_{k=1}^{l}L\left(y_k,\sum_{m=1}^{i+1}\alpha_m g_m(\boldsymbol{x}_k)\right) \tag{11}$$

通常在神经网络中所采用的损失函数均可以应用于核匹配追踪学习机中,例如:

(1) 平方损失.

$$L(y,f_n(\boldsymbol{x}))=(\hat{f}(\boldsymbol{x})-y)^2 \tag{12}$$

(2) 修正双曲正切损失.

$$L(y,f_n(\boldsymbol{x}))=(\tanh\hat{f}(\boldsymbol{x})-0.65y)^2 \tag{13}$$

由于在分类问题中,观测值 $y\in\{-1,+1\}$,故而,将核匹配追踪方法应用于分类领域中可以采用间隔损失函数,假定分类器输出为 $f(\boldsymbol{x})$,则间隔损失函数为

(1) 平方间隔损失.

$$(f(\boldsymbol{x})-y)^2=(1-m)^2 \tag{14}$$

(2) 修正双曲正切间隔损失.

$$(\tanh f(\boldsymbol{x})-0.65y)^2=(0.65-\tanh(m))^2 \tag{15}$$

其中,$m=yf(\boldsymbol{x})$,称为分类间隔.

最终,由核匹配追踪学习机训练所得到的判决超平面为

$$f_N(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i g_i(\boldsymbol{x})=\sum_{i\in\{sp\}}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i) \tag{16}$$

其中$\{sp\}$①表示由模糊核匹配追寻算法得到的支撑模式.

3　模糊核匹配追寻(Fuzzy Kernel Matching Pursuit,即 Fuzzy KMP 或 FKMP)

3.1　基于平方损失函数的模糊学习机

核匹配追寻在模式识别领域中表现出了显著的优势,然而,由于算法本身的特点,限制了对一些特定问题的应用(如时间序列样本识别等).下面,我们将详细建立模糊核匹配追寻,对传统的核匹配追寻进行扩展.

由式(2),传统的核匹配追寻采用平方损失,并令残差$\boldsymbol{R}_N=\boldsymbol{y}-\boldsymbol{f}_N$,通过该残差函数,任意一点的残差均等于目标值 y_i与该点的逼近值 $f(\boldsymbol{x}_i)$的差值.这样,所有样本的残差定义并没有区别,从而使得标准 KMP 算法平等地对待所有的样本,最终做出的判决也是对所有样本的一个平等综合考虑,要求总识别误差尽可能小,并不能对某一类指定的样本进行针对性的识别.然而在实际的问题中,样本之间的重要性是不同的(如癌变细胞与正常细胞),问题的核心就是对这些重要的样本做出尽可能准确的判别.本文提出了模糊核匹配追寻,根据每个样本的重要性对其赋予不同的权重 s_i,$i=1\sim l$(称之为模糊因子),并根据模糊因子重新定义其残差,使得学习机对每一个样本的学习程度不同,对应较大的 s_i要求学习机对其充分学习,尽可能地保证识别正确,而

① KMP 的训练类似于支撑矢量机(SVM),即每一个训练样本均对应一个系数 α_i,而决策超平面仅取决于那些对应系数 α_i不为零的样本,为了区分于支撑矢量机中关于支撑矢量(SV)的定义,我们将核匹配追寻中对应系数 α_i不为零的样本称为支撑模式(Support Pattern,SP).

对于较小的 s_i 则要求学习机仅对其进行粗略的学习.这样,学习机最终得到的判别函数就是考虑了不同权重样本的结果,能够对权重高的样本做出尽可能精确的识别.

首先,我们给出如下定义.

定义 1(⊙运算). 对于两个向量 $\boldsymbol{x}=(x_1,\cdots,x_m)$,$\boldsymbol{y}=(y_1,\cdots,y_m)$,向量之间的⊙运算定义为

$$\boldsymbol{x}\odot\boldsymbol{y}=(x_1\cdot y_1,\cdots,x_m\cdot y_m) \tag{17}$$

同时,

$$\|\boldsymbol{x}\odot\boldsymbol{y}\|^2=\sum_{i=1}^{m}(x_i\cdot y_i)^2 \tag{18}$$

下面,我们详细地建立基于平方损失函数的模糊核匹配追寻.

给定样本$\{(\boldsymbol{x}_1,y_1,s_i),\cdots,(\boldsymbol{x}_l,y_l,s_l)\}$,其中$\boldsymbol{x}\in R^N$为其特征,$y\in R$ 为观测值,$s\in R$ 为其相应的权重因子(模糊因子),采用核函数 $K:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$,利用观测点$\{x_1,\cdots,x_l\}$处的核函数值生成函数字典:$D=\{g_i=K(\cdot,x_i)\,|\,i=1,\cdots,l\}$.

重新定义残差

$$\boldsymbol{r}_N=S\odot(\boldsymbol{y}-\boldsymbol{f}_N)=\begin{bmatrix}s_1(y_1-f_N(\boldsymbol{x}_1))\\ \cdots\\ s_l(y_l-f_N(\boldsymbol{x}_l))\end{bmatrix} \tag{19}$$

其中,$f_N(\boldsymbol{x}_i)=\sum_{j=1}^{N}\alpha_j\boldsymbol{g}_j(\boldsymbol{x}_i)$是第 $i(i=1\sim l)$点的估计值 $\hat{y}_i$,则其重构误差为

$$\|\boldsymbol{r}_N\|^2=\|S\odot(\boldsymbol{y}-\boldsymbol{f}_N)\|^2=\sum_{i=1}^{l}(s_i(y_i-f_N(x_i)))^2 \tag{20}$$

由匹配追寻算法,

$$\begin{aligned}\|\boldsymbol{r}_{N+1}\|^2&=\|S\odot(y-(\boldsymbol{f}_N+\alpha_{N+1}\boldsymbol{g}_{N+1}))\|^2\\&=\|S\odot(y-\boldsymbol{f}_N)-S\odot(\alpha_{N+1}\boldsymbol{g}_{N+1})\|^2\\&=\|\boldsymbol{r}_N-S\odot(\alpha_{N+1}\boldsymbol{g}_{N+1})\|^2\\&\triangleq\|\boldsymbol{r}_N-S\odot(\alpha\boldsymbol{g})\|^2\end{aligned} \tag{21}$$

则

$$\|\boldsymbol{r}_{N+1}\|^2=\|\boldsymbol{r}_N\|^2-2\alpha\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle+\alpha^2\|S\odot\boldsymbol{g}\|^2 \tag{22}$$

寻找相应的 $\alpha\in R,\boldsymbol{g}\in D$,使得重构误差 $\|\boldsymbol{r}_{N+1}\|^2$ 最小,令$\frac{\partial\|\boldsymbol{r}_{N+1}\|^2}{\partial\alpha}=0$,可得

$$-2\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle+2\alpha\|S\odot\boldsymbol{g}\|^2=0 \tag{23}$$

故

$$\alpha=\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|^2} \tag{24}$$

将 $\alpha=\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|^2}$代入式(21),得

$$\begin{aligned}\|\boldsymbol{r}_{N+1}\|^2&=\|\boldsymbol{r}_N\|^2-2\cdot\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|^2}\cdot\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle+\\&\left(\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|^2}\right)^2\|S\odot\boldsymbol{g}\|^2\\&=\|\boldsymbol{r}_N\|^2-\left(\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|}\right)^2\end{aligned} \tag{25}$$

由上,模糊核匹配追寻即是在由核函数生成的字典 D 中,寻找基函数$\boldsymbol{g}$,使得$\|\boldsymbol{r}_{N+1}\|^2$最小,即

$$\boldsymbol{g}_{N+1}=\arg\min_{\boldsymbol{g}\in D}\left(\|\boldsymbol{r}_N\|^2-\left(\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|}\right)^2\right) \tag{26}$$

式(27)等价为

$$\boldsymbol{g}_{N+1}=\arg\max_{\boldsymbol{g}\in D}\left|\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}\rangle}{\|S\odot\boldsymbol{g}\|}\right| \tag{27}$$

相应的

$$\alpha_{N+1}=\frac{\langle\boldsymbol{r}_N,S\odot\boldsymbol{g}_{N+1}\rangle}{\|S\odot\boldsymbol{g}_{N+1}\|^2} \tag{28}$$

采用同标准匹配追寻相似的方法,每 $fitN$ 步进行一次后拟合来修正系数 $\alpha_1,\alpha_2,\cdots\alpha_i$,使$\boldsymbol{f}_i$进一步逼近观测值,即

$$\begin{aligned}\alpha_1,\cdots,\alpha_i&=\arg\min_{\alpha_1,\cdots,\alpha_i}\|S\odot(\boldsymbol{f}_i-\boldsymbol{y})\|^2\\&=\arg\min_{\alpha_1,\cdots,\alpha_i}\left\|S\odot\left(\sum_{k=1}^{i}\alpha_k\boldsymbol{g}_k-\boldsymbol{y}\right)\right\|^2\end{aligned} \tag{29}$$

最终得到判决函数

$$f_N(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i g_i(\boldsymbol{x})=\sum_{i\in\{sp\}}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i) \tag{30}$$

其中$\{sp\}$表示由模糊核匹配追寻算法得到的支撑模式.

3.2 基于任意损失函数的模糊核匹配追踪学习机

类似于核匹配追踪学习机向非平方损失函数地拓展策略,采用梯度下降法将模糊核匹配追踪学习机拓展到任意的非平方损失函数.

给定某损失函数 $L(y_i,f_N(\boldsymbol{x}_i))$,结合模糊因子我们重新建立基于损失函数 $L(y_i,f_N(\boldsymbol{x}_i))$的自适应残差为 $s_i\cdot L(y_i,f_N(\boldsymbol{x}_i))$,即

$$\tilde{\boldsymbol{r}}_N=\left(-s_1\frac{\partial L(y_1,f_N(\boldsymbol{x}_1))}{\partial f_N(\boldsymbol{x}_1)},\cdots,-s_l\frac{\partial L(y_l,f_N(\boldsymbol{x}_l))}{\partial f_N(\boldsymbol{x}_l)}\right) \tag{31}$$

利用贪婪算法,在第 $N+1$ 步迭代中,最优基原子和相应的系数为

$$\boldsymbol{g}_{N+1}=\arg\max_{\vec{g}\in D}\left|\frac{\langle\boldsymbol{g},\tilde{\boldsymbol{r}}_N\rangle}{\|\boldsymbol{g}\|}\right| \tag{32}$$

$$\alpha_{N+1}=\arg\min\sum_{i=1}^{l}(s_i\cdot L(y_i,\boldsymbol{f}_N(\boldsymbol{x}_i)+\alpha_{N+1}\boldsymbol{g}_{N+1}(\boldsymbol{x}_i))) \tag{33}$$

当增加 $\alpha_{N+1}\boldsymbol{g}_{N+1}$后,匹配追踪在第 i 代对观测值的

逼近并不一定是最优的;仍然通过后拟合的方法修正 $\boldsymbol{f}_i$,使其进一步逼近观测值,即重新调整系数 α_1, α_2,…,α_{N+1},使得当前的自适应残差能量最小:

$$\alpha_1,\cdots,\alpha_{N+1}=\underset{\alpha_k\in R(k=1\sim N+1)}{\arg\min}\sum_{i=1}^{l}s_i\cdot L\left(y_i,\sum_{j=1}^{N+1}\alpha_j g_j(\boldsymbol{x}_i)\right) \tag{34}$$

最后得到的模糊核匹配追踪学习机的判决超平面为

$$f_N(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i g_i(\boldsymbol{x})=\sum_{i\in\{sp\}}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i) \tag{35}$$

其中$\{sp\}$表示由模糊匹配追踪算法得到的支撑模式.

3.3　模糊参数的选取

模糊核匹配追寻根据每个样本的重要性对其赋予不同的权重 s_i,$i=1\sim l$(称之为模糊因子),使得学习机对每一个样本的学习程度不同,对应较大的 s_i要求学习机对其充分学习,而对于较小的 s_i则要求学习机仅对其进行粗略的学习,从而使得不同的样本对最终的判决函数做出相应的贡献.

3.3.1　阶跃参数

在实际的应用中,经常碰到这样一种问题:要求对其中一类样本的识别精度很高,甚至只考虑对指定类别样本的识别(如对癌症细胞的识别),这就使得指定类别样本比其余样本更为重要,要求对这类样本的识别精度很高,此时模糊因子选取如下:

$$s_i=\begin{cases}1+D, & y_i\text{为指定类别}\\ 1-D, & y_i\text{为非指定类别}\end{cases} \tag{36}$$

这里,$D\in(0,1)$是折中因子,对两类样本的识别精度取折中(在这一类问题中,即使我们对非指定类别产生较大的错分误差,它带来的风险仍然比较小,错识指定类别造成的风险低,所以,允许学习机对非指定类别有一个较低的识别率,而对指定类别必须具有较高的精度识别):D 越大,对指定类别的样本学习程度越充分,识别精度越高,同时非指定类别样本的识别精度损失也越大.

3.3.2　时间参数

在某些特定的工程应用(如经济预测、气象预报等)中,样本是随着时间逐次到达的,并且由先验信息已知某时间段或晚到的样本具有相对重要的意义.因而,设计模糊函数 $S_i=f(t_i)$,$i=1\sim l$ 是对时间的函数①,可以采用如下的表达形式[10]

$$S_i=1-1\Big/\left(1+\exp\left(2a\left(\frac{i}{l}-b\right)\right)\right) \tag{37}$$

这里,i 代表第 i 个到达的样本,共采集到 l 个样本,$a>0$ 是衰减因子,$b\in[0,1]$为遗忘因子,通过图 1 和图 2 可以清晰地看出 c_i随时间及衰减因子 a、遗忘因子 b 的变化.

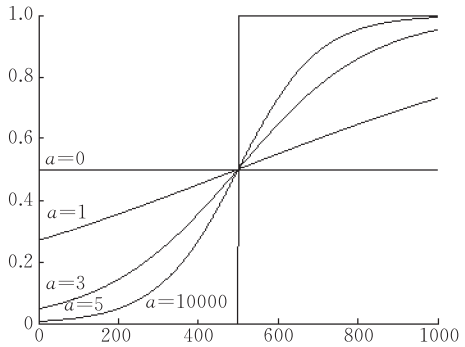

图 1　时间参数图 $M=1$,$b=0.5$(x 轴为样本序列;当 $a=0$ 时,$f(t_i)$均为 0.5,随着 a 的增大,函数左半部分下降,右半部分上升,至 $a=10000$ 时,$f(t_i)$相当于阶跃函数)

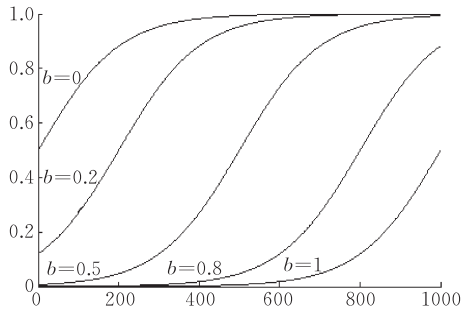

图 2　时间参数图 $M=1$,$a=5$(x 轴为样本序列;当 $b=1$ 时,大部分先前的样本被遗忘,随着 b 的减小,被遗忘的样本数量下降)

4　仿真实验

4.1　指定样本高精度识别

产生两类交错的同心圆样本 $\begin{cases}x=\rho\cdot\cos\theta\\ y=\rho\cdot\sin\theta\end{cases}$,$\theta\in U[0,2\pi]$,其中第一类样本的半径为均匀分布 $U[0,6]$,第二类样本的半径为均匀分布 $U[3,10]$,两类样本各 50 个作为训练样本,采用 RBF 核 $K(\boldsymbol{x},\boldsymbol{y})=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2p^2)$, $p=6$,核匹配追寻②参数 $maxN=30$,$fitN=4$ 对样本进行实验,采用折中因子选取 $D=0.3$.分别用“+”和“◇”表示两类样本,要求对样本“◇”(即中心区域样本)的识

① 在时间参数中,本文给出的时间参数是基于“晚到的样本具有相对重要的意义”这一情况,对于“特定时间段内的重要样本”,其模糊因子的设定可采用 3.3.1 节中阶跃参数的形式.

② 本文中,KMP 均采用了早停策略(即预设贪婪算法的最大迭代次数,用 $maxN$ 表示);$fitN$ 表示每经过 $fitN$ 步进行一次后拟合,参见文献[1].

别精度尽可能高.图 3、图 4 分别给出了用标准核匹配追寻和模糊核匹配追寻识别的结果,从图中清晰地看出,模糊核匹配追寻能够很好地满足我们的要求,对"◇"样本达到 100%的识别,而标准核匹配追寻则不能.

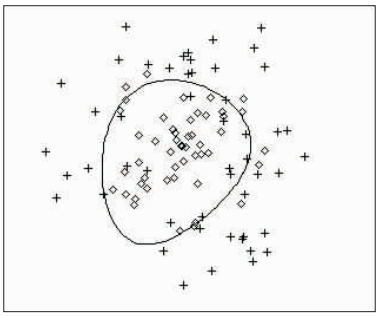

图 3 标准核匹配追寻对同心圆样本的识别

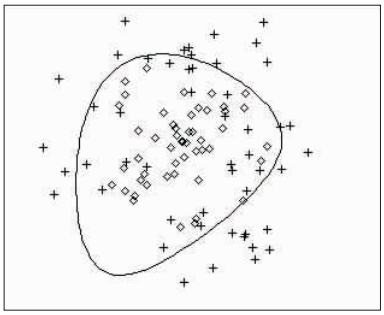

图 4 模糊核匹配追寻对同心圆样本的识别(要求对"◇"样本的识别精度)

4.2 时间序列样本识别

产生两类交错分布的同心圆样本各 26 个;用数字记录该样本的位置及到达时刻,用阴影数字两类样本;分别用标准核匹配追寻和本文提出的模糊匹配追寻对两类样本进行了识别,要求能够对新颖样本的识别率尽可能高.实验采用时间学习因子选取 $a=8, b=1$,RBF 核参数 $p=6$,核匹配追寻参数 $maxN=30, fitN=4$.图 5 是标准核匹配追寻给出的结果,图 6 是模糊核匹配追寻给出的分类结果.由图可知:模糊核匹配追寻对最后采得的 20 个样本作出了精确的分类而传统的核匹配追寻则不然.

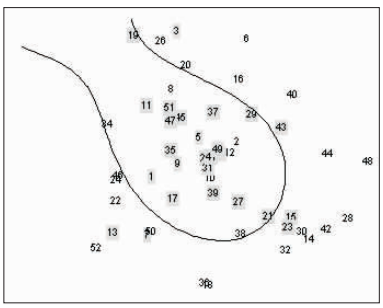

图 5 标准核匹配追寻对时间序列样本的识别

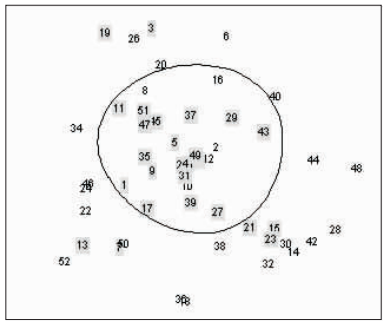

图 6 模糊核匹配追寻对时间序列样本的识别

4.3 FKMP 有效性测试

选取 UCI① 数据库中的 Heart Disease 数据,Heart Disease 数据由 13 个含噪特征属性和一个类别属性构成,是一个 2 类问题,共 270 个样本,选取 170 个样本进行训练(74 个正类样本),其余 100 个样本中的 44 个正类样本作测试.实验中模糊核匹配追寻选取阶越参数,图 7 给出了不同折中因子 D 取值下对正类样本和负类样本的测试误差.其中,核匹配追寻参数选取:$maxN=80, fitN=8$,RBF 核参数 $p=1.0$,模糊因子在[0.01,0.5]上等间采样 50 次.由图可知:随着 D 的增大,目标样本的识别误差随之下降.

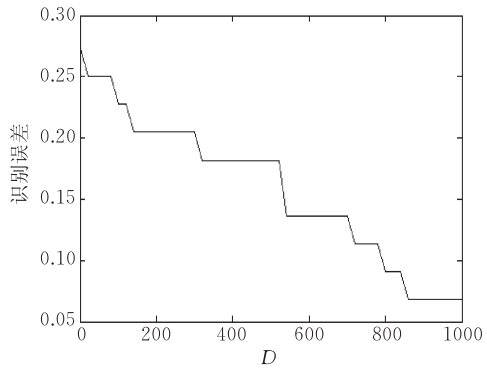

图 7 阶越学习因子 D 对目标样本识别影响

4.4 对实际数据的测试

选取 UCI 数据库中的 Breast Cancer、Diabetis、Heart Disease 及 Thyroid 数据对本文提出的模糊核匹配追寻算法进行测试.其中,Breast Cancer 数据由 9 个含噪特征属性和一个类别属性构成,是一个 2 类问题,共 277 个样本,选取 200 个作为检验样本,其余 77 个样本中的 23 个正类样本作测试;Pima Indians Diabetes 数据由 8 个含噪特征属性和一个类别属性构成,是一个 2 类问题,共 768 个样本,选取 256 个样本进行训练,其余 512 个样本中的

① http://www.ics.uci.edu/~mlearn/MLRepository.html

174 个正类样本作测试; Thyroid 数据由 5 个含噪特征属性和一个类别属性构成,是一个 2 类问题,共 215 个样本,选取 140 个样本进行训练,其余 75 个样本中的 26 个正类样本作测试.

在本实验中,我们更为关注对于正类样本的分类性能,这是因为正类样本均刻画了检测呈阳性的病理状态,学习机的任务就是要对这一类样本尽可能地精确识别. 实验参数:采用 RBF 核 $K(\boldsymbol{x},\boldsymbol{y})=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2p^2)$,对于 Breast Cancer 数据 $maxN=60$, $fitN=5$, $p=0.8$,模糊因子 $D=0.6$;对于 Pima Indians 数据 $maxN=100$, $fitN=8$, $p=6.0$,模糊因子 $D=0.8$,对于 Thyroid 数据 $maxN=50$, $fitN=3$, $p=0.14$,模糊因子 $D=0.1$,分别用标准 KMP 和 FKMP 对病理类别特征样本进行识别测试. 我们是在 matlab 环境下,P4 3.2GHz、2GB 内存的微机上独立进行 30 次实验取平均的结果,表 1 给出了具体的实验结果. 由于采用的样本为非平衡样本(即两类样本个数相差较大),传统的核匹配追寻不能对弱势样本(数量小的一类样本)进行有效的识别,甚至失去了识别能力(识别率<50%),而采用模糊核匹配追寻,就可以有效地解决这一问题,仿真的实验结果中,当标准 KMP 对弱势样本的识别率<50%(34.78%)时,利用模糊 KMP,仍然可以使识别精度可以达到 99%以上.

表 1　对 UCI 非平衡数据的测试

数据	训练样本	检验样本	损失函数	算法	支撑模式	识别率/%
Breast Cancer	+1 类:58 −1 类:142	+1 类:23	*Loss-mse*①	Fuzzy KMP	4	99.92
				KMP	55	34.78
			*Loss-tanh*②	Fuzzy KMP	6	99.87
				KMP	64	33.56
Pima Indians Diabetes	+1 类:94 −1 类:162	+1 类:174	*Loss-mse*	Fuzzy KMP	87	99.14
				KMP	100	56.32
			Loss-tanh	Fuzzy KMP	69	99.25
				KMP	102	57.22
Thyroid	+1 类:39 −1 类:101	+1 类:26	*Loss-mse*	Fuzzy KMP	50	100
				KMP	50	84.62
			Loss-tanh	Fuzzy KMP	48	100
				KMP	50	83.69

5　总　结

核匹配追寻具有很强的推广能力、强大的非线性处理能力和高维处理能力,同时较其它核机器相比,其稀疏性更优. 然而在实际问题中经常遇到这样几种情况:(1) 所获得的样本是具有时间属性的;(2) 要求其中一类样本的识别精度;(3) 非平衡样本的识别. 由于传统的核匹配追寻在处理模式识别的问题上平等对待所有的样本,它要求总识别误差尽可能地小,但是并不能对某一类或某一些指定的样本进行针对性的识别,这就限制了核匹配追寻在这些实际问题中的应用.

针对这些问题,本文提出了模糊核匹配追寻,根据问题的要求对每个样本作出重要性定义,学习机可以根据样本的重要性定义进行程度不同的学习,对次要的样本粗略学习,而对重要的样本进行充分学习,使学习机的最终判决对指定的重要样本达到较高的识别精度. 本文进行了大量的仿真实验,结合分类图例证实了模糊核匹配追寻可行性及有效性;在对 UCI 数据的性能测试中可以得出:当传统的核匹配追寻已不能对弱势样本进行识别(识别率小于 50%)时,模糊核匹配追寻仍然对弱势样本保持了较高的识别精度.

参　考　文　献

[1] Vincent Pascal, Bengio Yoshua. Kernel matching pursuit. Machine Learning, 2002, 48: 165-187

[2] Davis G, Mallat S, Zhang Z. Adaptive time-frequency decompositions. Optical Engineering, 1994, 33(7): 2183-2191

[3] Pati Y, Rezaiifar R, Krishnaprasad P. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition//Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers, CA, USA, 1993: 40-44

[4] Mallat S, Zhang Z. Matching pursuit with time-frequency dictionaries. IEEE Transactions on Signal Processing, 1993, 41(12): 3397-3415

[5] Mallat S. A theory for nuliresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11: 674-693

① *Loss-mse* 表示核匹配追寻采用平方损失函数.
② *Loss-tanh* 表示核匹配追寻采用修正双曲正切损失函数.

[6] Vapnik V N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999

[7] Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 1-47

[8] Schölkopf B, Smola A. Learning with Kernels. Cambridge, MA: MIT Press, 1999

[9] Burges C J. Geometry and invariance in kernel based method//Advance in Kernel Method-Support Vector Learning. Cambridge, MA: MIT Press, 1999: 86-116

[10] Cao L J, Francis E H. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on Neural Networks, 2003, 14(6): 1506-1518

LI Qing, born in 1979, Ph. D., engineer. His current research interests include machine learning, pattern recognition and statistic learning theory.

JIAO Li-Cheng, born in 1959, Ph. D., professor, Ph. D. supervisor. His current research interests include nonlinear theory, neural network, data mining, evolutionary computation and wavelet theory.

ZHOU Wei-Da, born in 1974, Ph. D.. His current research interests include intelligent information processing, machine learning, statistic learning theory and data mining.

Background

The authors have made researches on many fields of the support vector machine, such as Linear programming support vector machine, kernel matching pursuit classifier ensemble, support vector regression based on unconstrained convex quadratic programming and so on, and have applied these methods to the field of SAR image processing, recognition of plane HRRP and many other fields. This paper belongs to the part of novel method of machine learning and focuses on proposing an fuzzy kernel matching pursuit machine (FKMP), which can classify the appointed important samples much more precisely according to the predefined importance of the data, so as to develop the practical applications of the KMP.