文章编号: 1001-0920 (2016) 11-1960-07 **DOI:** 10.13195/j.kzyjc.2015.1465

基于弱贪婪策略的快速直觉模糊核匹配追踪方法

樊 雷[1a], 雷英杰[1b], 赵 敏[1b], 段宏燕[2]

(1. 空军工程大学 a. 训练部，b. 防空反导学院，西安 710051；2. 中国人民解放军 94259 部队，山东 蓬莱 265600)

摘 要: 针对现有直觉模糊核匹配追踪算法采用贪婪算法搜索最优基函数而导致学习时间过长的局限性, 基于弱贪婪策略, 提出一种随机直觉模糊核匹配追踪算法. 该算法不需要保证每次迭代过程都能搜索到当前最优基函数, 仅需要在原搜索空间随机抽取一个较小的核字典子集进行搜索来获得近似最优基函数, 从而有效减少一次迭代过程的搜索空间, 大大降低了算法的训练时间. 仿真结果表明, 所提出方法在保持识别精度相当的情况下, 有效缩短了一次匹配追踪时间, 计算效率明显提高, 且所得模型具有稀疏性好、泛化能力高等优点.

关键词: 直觉模糊集；核匹配追踪；弱贪婪算法；目标识别

中图分类号: TP182；TP39 **文献标志码:** A

Fast intuitionistic fuzzy kernel matching pursuit-based week greedy algorithm

FAN Lei[1a], *LEI Ying-jie*[1b], *ZHAO Min*[1b], *YIN Hong-yan*[2]

(1a. Department of Training，1b. School of Air and Missile Defense，Air Force Engineering University，Xi'an 710051, China；2. PLA 94259 Troop，Penglai 265600，China．Correspondent：FAN Lei，E-mail：fanlei1771@163.com)

Abstract: In order to overcome the long learning time caused by searching optimal basic function data based on the greedy strategy from a redundant basis function dictionary for the intuitionistic fuzzy kernel matching pursuit(IFKMP), the random intuitionistic fuzzy kernel matching pursuit algorithm based on the weak greedy strategy is proposed. Rather than getting the present optimal basic function in each search, the approximate optimal basic function can be obtained by searching a random kernel dictionary subset of the original searching space, so that the searching space of matching pursuit can be reduced, and the training time can be decreased greatly. Simulation results show that, compared with the conventional approaches, the proposed algorithm can decrease training time and improve calculation efficiency obviously leaving the classification accuracy almost unchanged, while the model has better sparsity and generalization.

Keywords: intuitionistic fuzzy set；kernel matching pursuit；week greedy algorithm；target recognition

0 引 言

核匹配追踪[1](KMP) 是 2002 年提出的一种新型核机器学习方法, 其主要思想源于信号处理中的匹配追踪算法和支持向量机中的核方法[2-3]. 核匹配追踪学习机的性能与支持向量机相当, 但有更为稀疏的解[4]. 目前, KMP 理论已成功应用于图像识别、目标分类、人脸识别、特征模式识别等多个领域[5-12].

虽然核匹配追踪理论已经在目标识别领域取得了成功应用, 但在实际应用情况中却存在一种特殊情况: 某一类目标的重要程度 (或威胁程度) 比其余目标更高, 因此需要对重要类别目标进行更高精度的识别, 而对其余目标则可以适当降低识别精度要求, 例如未来反导拦截作战, 对诱饵、碎片等虚假目标的识别精度远远低于对真弹头的识别精度要求. 传统 KMP 算法的判决函数是针对所有学习样本的一个综合考虑, 因此无法针对某一类指定样本进行有效识别, 导致 KMP 理论在很多特殊性领域的使用受到限制. 针对该问题, 文献 [13] 提出了模糊核匹配追踪学习机 (FKMP), 通过对每个样本类别设定不同模糊因子, 使学习机作出针对指定类别样本的判决. 但该方法根据人工经验设定模糊因子, 会对训练过程带来一定风险进而导致识别信息的损失. 文献 [14] 将核匹配

收稿日期: 2015-11-27；**修回日期:** 2016-04-25.

基金项目: 国家自然科学基金项目(61272011, 61309022)；陕西省自然科学青年基金项目(2013JQ8031).

作者简介: 樊雷 (1981–), 男, 讲师, 博士, 从事智能信息处理与信息融合的研究；雷英杰 (1956–), 男, 教授, 博士生导师, 从事智能信息处理与智能决策等研究.

追踪算法拓展到直觉模糊理论领域, 提出了直觉模糊核匹配追踪(IFKMP)学习机, 解决了对特殊重要样本进行高精度识别这一难题.

随着云时代的来临, 大数据(BD)也吸引了越来越多的关注[15]. 未来是一个大数据的时代, 实际应用中也必将面临越来越多的大数据问题. 然而, 直觉模糊核匹配追踪学习机在本质上仍是采用贪婪策略搜索最优基函数的线性组合, 每次匹配过程所需的大量的计算资源和运算时间限制了其在大数据问题上的应用. 鉴于此, 本文对直觉模糊核匹配追踪学习机进行改良, 基于弱贪婪策略[16-17]提出一种随机直觉模糊核匹配追踪(SIFKMP)的快速识别算法, 使其能够以可接受范围内的资源代价处理大数据问题. 实验阶段选取3组不同的数据集进行仿真实验, 并将其与传统算法进行比较, 实验结果充分表明了SIFKMP算法的优越性和有效性.

1　直觉模糊核匹配追踪算法

定义1 (e运算)[13]　对于两个向量$x=\{x_1,x_2,\cdots,x_m\}$和$y=\{y_1,y_2,\cdots,y_m\}$, 向量之间的e运算定义为

$$\boldsymbol{x}e\boldsymbol{y}=(x_1gy_1,x_2gy_2,\cdots,x_mgy_m). \tag{1}$$

同时有

$$\|\boldsymbol{x}e\boldsymbol{y}\|^2=\sum_{i=1}^{m}(x_igy_i)^2. \tag{2}$$

输入训练样本集为$\{(x_1,y_1,\omega(y_1)),\cdots,(x_l,y_l,\omega(y_l))\}$. 其中: ω_i为直觉模糊参数, $\boldsymbol{x}_i\in\boldsymbol{R}^N$和$y_i\in\boldsymbol{R}$分别为训练样本的特征值和观测值. 令核函数为$K:\boldsymbol{R}^d\times\boldsymbol{R}^d\to\boldsymbol{R}$, 利用训练样本$\{x_1,x_2,\cdots,x_l\}$处的核函数值生成函数字典库$D=\{g_i=k(\cdot,x_i)|i=1,2,\cdots,l\}$.

定义残差为

$$\boldsymbol{r}_N=\boldsymbol{\omega}e(\boldsymbol{y}-\boldsymbol{f}_N)=\begin{bmatrix}\omega(y_1)(y_1-f_N(\boldsymbol{x}_1))\\ \vdots\\ \omega(y_l)(y_l-f_N(\boldsymbol{x}_l))\end{bmatrix}, \tag{3}$$

其中$f_N(\boldsymbol{x}_i)=\sum_{j=1}^{N}\alpha_jg_j(\boldsymbol{x}_i)$为$\boldsymbol{x}_i$的观测估计值$\hat{y}_i$. 重构误差为

$$\|\boldsymbol{r}_N\|^2=\|\boldsymbol{\omega}e(\boldsymbol{y}-\boldsymbol{f}_N)\|^2=\sum_{i=1}^{l}(\omega(y_i)(y_i-f_N(\boldsymbol{x}_i)))^2. \tag{4}$$

由匹配追踪算法可得

$$\begin{aligned}&\|\boldsymbol{r}_{N+1}\|^2=\\&\|\boldsymbol{\omega}e(\boldsymbol{y}-(\boldsymbol{f}_N+\alpha_{N+1}\boldsymbol{g}_{N+1}))\|^2=\\&\|\boldsymbol{r}_N\|^2-2\alpha\langle\boldsymbol{r}_N,\boldsymbol{\omega}e\boldsymbol{g}\rangle+a^2\|\boldsymbol{\omega}e\boldsymbol{g}\|^2,\end{aligned} \tag{5}$$

搜索相应的$\alpha\in\boldsymbol{R}$, $\boldsymbol{g}\in D$使重构误差最小, 令$\partial\|\boldsymbol{r}_N\|^2/\partial\alpha=0$, 求得

$$\alpha=\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}e\boldsymbol{g}\rangle}{\|\boldsymbol{\omega}e\boldsymbol{g}\|^2}. \tag{6}$$

将式(6)代入(5), 得到

$$\|\boldsymbol{r}_{N+1}\|^2=\|\boldsymbol{r}_N\|^2-\Big(\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}e\boldsymbol{g}\rangle}{\|\boldsymbol{\omega}e\boldsymbol{g}\|}\Big)^2. \tag{7}$$

根据最小化重构误差准则, 可得

$$\boldsymbol{g}_N=\arg\max_{g\in D}\Big|\Big(\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}e\boldsymbol{g}\rangle}{\|\boldsymbol{\omega}e\boldsymbol{g}\|}\Big)\Big|. \tag{8}$$

对应地, 有

$$\alpha_N=\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}e\boldsymbol{g}_N\rangle}{\|\boldsymbol{\omega}e\boldsymbol{g}_N\|^2}. \tag{9}$$

最终得到判决函数为

$$f_N(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i\boldsymbol{g}_i(\boldsymbol{x})=\sum_{i\in\{sv\}}\alpha_ik_i(\boldsymbol{x},\boldsymbol{x}_i), \tag{10}$$

其中$\{sv\}$为学习机训练所得的支持向量集.

2　基于弱贪婪策略的直觉模糊核匹配追踪算法

通过对直觉模糊核匹配追踪算法的步骤进行观察可以发现, 该算法的计算"瓶颈"在于每一次匹配过程均采用贪婪策略搜索最优基函数, 大大增加了计算量. 本文对直觉模糊核匹配追踪学习机进行改良, 并基于弱贪婪策略提出一种SIFKMP算法, 使其能够以可接受范围内的资源代价处理大数据问题.

2.1　弱贪婪策略

文献[16]提出了一种能够近似匹配过程的弱贪婪算法(WGA). 文献[17]分析了WGA几种不同的构想, 并证明了其在不同条件下的收敛性. 与经典贪婪算法不同, WGA采用下式逼近目标函数:

$$\begin{aligned}&\tilde{f}_{N+1}=\tilde{f}_N+t_{N+1}\alpha_{N+1}\boldsymbol{g}_{N+1}=\\&\tilde{f}_N+\tilde{\alpha}_{N+1}\boldsymbol{g}_{N+1},\ t_{N+1}\in[0,1],\end{aligned} \tag{11}$$

其中$\tilde{\alpha}_{N+1}=t_{N+1}\alpha_{N+1}$. 当$t_{N+1}=1$时, WGA算法退化为经典贪婪算法. 序列$\tau=\{t_1,t_2,\cdots,t_N\}$称为衰减序列, 若衰减序列$\tau$满足$\exists\tilde{t}>0$, 则当$\forall N\geqslant1$, $t_N>\tilde{t}$时, WGA算法能够保证其收敛性[16]. 这意味着在匹配追踪过程中, 不需要保证每次迭代过程都能搜索到当前最优值, 只需要搜索到一个近似最优值, 在连续无穷次迭代后也能保证算法收敛到目标函数. 当然, 与经典贪婪算法相比, 每一次搜索到的近似最优值越接近实际最优值, 弱贪婪算法的性能越接近经典贪婪算法[18-19].

2.2　随机直觉模糊核匹配追踪学习机

由弱贪婪策略原理可知, 每次匹配过程中得到一个近似最优基函数也能保证学习机最后所得的判决函数以给定精度收敛于目标函数. 可以通过从原核字

典库中随机抽取一个较小的核字典子集进行搜索的方式获得近似最优基函数, 这样便达到了对直觉模糊核匹配追踪学习机进行改良并减少其训练时间的目的. 下面从概率统计的角度分析如何确定核字典子集的规模.

设 $X_1, X_2, \cdots, X_n$ 为 n 个独立同分布的随机变量, 分布函数为 $F(x)$, 令 $\zeta = \max\{X_1, X_2, \cdots, X_n\}$, 表示 n 个随机变量的极大值, ζ 的分布函数为

$$\Pr(\zeta \leqslant x) = \Pr(X_1 \leqslant x, \cdots, X_n \leqslant x) = \Pr(X_1 \leqslant x) \cdots \Pr(X_n \leqslant x) = [F(x)]^n. \tag{12}$$

若 $X_1, X_2, \cdots, X_n$ 为 (0,1) 上的均匀分布, 则其分布函数为

$$F(x) = \begin{cases} 0, & x < 0; \\ x, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1. \end{cases} \tag{13}$$

极值 ζ 的分布函数为

$$[F(\zeta)^n] = \begin{cases} 0, & x < 0; \\ \zeta^n, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1. \end{cases} \tag{14}$$

分位数是大数据集和数据流上经常使用的一种统计方法, 通过分位数查询能够获取的统计信息, 以便为决策提供数据支持. 若 $x_i (i = 1, 2, \cdots, N)$ 是一组递增数列, 则 x_1 为最小的观测值, x_N 为最大的观测值. 每个观测值 x_i 与一个 τ_i $(0 < \tau_i < 1)$ 匹配, 指出大约 $100 \times \tau_i\%$ 的样本小于或等于 x_i, x_i 是相应于 $\tau_i\%$ 的分位数.

定义 2 (分位数)[20] 假设实随机变量 X 的分布函数为 $F(x) = P(X \leqslant x)$, 对于任意 $0 < \tau < 1$, 定义满足 $F(x) \geqslant \tau$ 的最小值为 x_τ 的 τ 分位数, 表示为

$$x_\tau = \inf\{x : F(x) \geqslant \tau\}. \tag{15}$$

若最小值满足 $x_\tau = \varepsilon$, 则极值 ζ 的 τ 分位数表示为

$$F^{-1}(\varepsilon) = \tau^{1/n} = \varepsilon. \tag{16}$$

经典直觉模糊核匹配追踪学习机在每次迭代过程中, 需要搜索 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 中的极大值, 从而获取对应的最优基函数. 基于弱贪婪思想, 只需在每次迭代过程中搜索到 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 中的近似极大值, 但近似极大值只是一个模糊概念, 实际操作过程中需要根据分位数的概念对其进行精确描述. 为了简化计算, 大胆假设 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的值服从 (0,1) 上的均匀分布 (在后续的实验部分将对该假设进行验证分析). 若需要取到的 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的近似极大值为 ε, 且满足 ε 为 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的 τ 分位数, 则有

$$\tau^{1/n} = \varepsilon \Rightarrow \frac{1}{n} = \log_\tau \varepsilon \Rightarrow n = \ln\tau / \ln\varepsilon. \tag{17}$$

假设 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 满足服从 (0,1) 上的均匀分布, 其极大值为 1. 若原核字典集中包含 10 000 个样本, 需要取到的 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的近似极大值为 $\varepsilon = 0.95$ (分位数 τ0.05), 则仅需从 D 中随机选取一个包含 $[\ln 0.05/\ln 0.95] = 59$ 个元素的核字典子集代入运算即可满足需求, 而算法一次匹配过程的计算量也降低为原来的 0.59%. 令 M_t 为第 t 代的核函数子集, M_t 中的样本从原始核字典集 D 中随机选取, 随机直觉模糊核匹配追踪算法的具体步骤如下所示.

算法 1 随机直觉模糊核匹配追踪算法.

输入: 样本数据集 $X = \{(x_1, y_1), \cdots, (x_l, y_l)\}$, 直觉模糊参数 $\omega(y_i)$, 后拟合参数 fit$N$, 核参数 σ, 迭代停止阈值 η, 最大迭代次数 $T_{\max}$, $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的近似极大值 ε 和分位数 τ;

输出: 判决函数 f_N.

Step 1: 生成核函数字典库 $D = \{g_i = K(\cdot, x_i) | i = 1, 2, \cdots, l\}$.

Step 2: 初始化残差 $r_1 = \omega \odot y$, 根据式 (17) 计算子集规模 n 的大小, 并令计数器 $t = 1$.

Step 3: 随机从核字典库中选取一个规模为 n 的子集 M_t, 从中搜索当前的近似最优基函数 $g_t = \arg\max\limits_{g \in M_t} |(\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||)|$.

Step 4: 根据式 (9) 计算基函数 g_t 对应的系数 α_t.

Step 5: 更新残差 $r_{t+1} = \omega \odot (r_t - \alpha_t g_t)$.

Step 6: 判断是否达到终止条件, 若达到, 则停止迭代, 输出判决函数 f_N; 否则, 令 $t = t+1$, 转至 Step 3.

由弱贪婪策略的收敛性可知, 算法 1 每次匹配追踪过程中, 只需要搜索到一个近似最优值, 在连续经历足够多的迭代次数后, 也能保证该算法收敛到指定精度. 此外, 通过调节 $|\langle \boldsymbol{r}, \boldsymbol{\omega e g}\rangle|/||\boldsymbol{\omega e g}||$ 的近似极大值 ε 和分位数 τ 的值, 可以对算法进行控制. 近似极大值 ε 和分位数 τ 决定了核字典子集的规模 n, n 越小, 算法一次匹配时间越短, 但达到预定精度所需的支持向量数目越多; n 越大, 算法越接近原始 IFKMP 算法.

该方法还涉及到核参数 σ 的选取, 如何对核参数 σ 进行取值, 目前尚缺乏理论支持, 更多的是依靠经验取值, 通常的解决方法是用一组专门的验证数据集确定核参数 σ. 而且, 随机直觉模糊核匹配追踪学习机本质上仍是一种二分类器, 对于多类分类问题, 通常有两种解决方案: 第 1 种方法将 N 类分类问题转化为 N 个二类分类问题, 这种方法需要 N 个分类器; 第 2 种方法将 N 类样本两两分类, 方法需要 $N(N-1)/2$ 个分类器.

2.3 SIFKMP 算法复杂度分析

直觉模糊核匹配追踪算法为了寻找最优权系数和对应的基函数, 每一次都要采用贪婪算法进行全

局搜索, 因此, 本文主要基于算法一次匹配过程的计算次数衡量算法的时间复杂度. 若字典集规模为 N, 算法迭代次数为 L, 则 IFKMP 算法一次匹配需执行的计算次数为 $N \cdot L$, 时间复杂度为 $O(N \cdot L)$. 若设定 $|\langle \boldsymbol{r}, \boldsymbol{\omega eg} \rangle|/||\boldsymbol{\omega eg}||$ 的近似极大值为 ε, 分位数为 τ, 则 SIFKMP 算法中一次匹配过程的核字典子集规模为 $n = \ln\tau/\ln\varepsilon$, SIFKMP 算法一次匹配需要运行的计算次数为 $n \cdot L$, 算法的时间复杂度为 $O(n \cdot L)$. SIFKMP 算法核字典子集规模 n 仅与近似极大值 ε 和分位数 τ 有关, 与原字典集规模 N 无关. 因此, 从算法的时间复杂度看, 当字典规模 N 较小时, SIFKMP 算法相对 IFKMP 算法的计算优势并不明显; 当字典规模 N 较大时, SIFKMP 算法的计算量远小于 IFKMP 算法.

3　实验分析

实验仿真环境如下: 操作系统 Window XP, 编译软件 Matlab7.6, Pentium(R) Dual-Core CPU E5500 @2.8GHz, 内存 2 GB. 为了验证所提出算法的优越性, 实验将标准核匹配追踪 (KMP)算法[1]、直觉模糊核匹配追踪算法 (IFKMP)[11]、基于粒子群优化的直觉模糊核匹配追踪算法 (PS-IFKMP)[21]和随机直觉模糊核匹配追踪算法 (SIFKMP) 进行对比. 实验过程中, 选取高斯核 $K(x,y) = \exp(-\|x-y\|^2/2\sigma^2)$ 作为核函数, 并选择不同的样本集合进行测试, 为了避免随机误差, 每次实验分别进行 50 次蒙特卡洛仿真.

3.1　Shuttle 数据识别

实验首先选取 UCI 数据集中数据量较大的 Shuttle 数据集进行验证. Shuttle 是一个七类数据集, 包含 43 500 个训练样本和 14 500 个测试样本, 其中约 80% 的样本为 1 类样本, 其余 2~7 类样本约占总样本的 20%. 为了方便计算, 将 1 类样本标记为正类样本, 其余 2 ~ 7 类样本标记为负类样本 (即异常样本).

为了验证算法在不同训练规模下的性能, 分别从原始训练样本中随机选取 2 500、5 000、7 500 和 10 000 个样本作为训练集进行实验. 实际应用中, 需要对异常样本进行检测, 因此对该类样本的识别应该比正类样本具备更高的识别精度. 本实验主要检测算法对异常样本的检测概率, 因此从 14 500 个测试样本中选取其中全部的负类样本 (共 3 022 个) 作为测试集进行测试. 同时, 为了研究算法在不同核字典子集规模条件下的识别率和运行时间, 分别设置 SIFKMP 算法的核字典子集规模为 59 ($\varepsilon = 0.95$, $\tau = 0.05$) 和 228 ($\varepsilon = 0.99$, $\tau = 0.02$). 参数设置如下: 通过实验验证得到核参数 $\sigma = 1$, 令正类样本 y_1 的直觉模糊参数 $\omega(y_1) = 0.9$, 负类样本 y_2 的直觉模糊参数 $\omega(y_2) = 1.2$. 粒子种群规模分别为 200、400、600、800, 最大迭代次数 $L = 1\,000$, 迭代误差阈值 $\eta = 0.02$, 实验结果如表 1 ~ 表 4 所示.

为了更好地对 3 种算法的效果进行对比, 将实验

表 1　训练规模为 2 500 条件下的识别结果

算法	训练规模	测试规模	支持向量	一次匹配时间/s	训练时间/s	识别率/%	偏差/%
KMP			283	0.085 6	61.099	94.48	2.23
IFKMP			527	0.180 8	223.45	**98.76**	**1.03**
PS-IFKMP	2 500	−:3 22	**45**	0.012 9	**49.672**	95.49	1.54
SIFKMP(59)			546	**0.006 2**	131.393	98.65	1.57
SIFKMP(228)			542	0.018	140.088	98.57	1.32

表 2　训练规模为 5 000 条件下的识别结果

算法	训练规模	测试规模	支持向量	一次匹配时间/s	训练时间/s	识别率/%	偏差/%
KMP			353	0.264 2	261.97	94.65	1.64
IFKMP			976	0.556 9	702.95	**98.97**	1.13
PS-IFKMP	5 000	−:3 022	**72**	0.042 8	**187.49**	96.36	1.21
SIFKMP(59)			1 000	**0.017**	422.21	98.91	1.48
SIFKMP(228)			999	0.039 4	475.43	98.94	**1.05**

表 3　训练规模为 7 500 条件下的识别结果

算法	训练规模	测试规模	支持向量	一次匹配时间/s	训练时间/s	识别率/%	偏差/%
KMP			394	0.554	662.05	94.49	1.67
IFKMP			998	1.011 1	1 328.6	**99.02**	**0.93**
PS-IFKMP	7 000	−:3 022	**139**	0.114 4	**431.52**	97.73	1.14
SIFKMP(59)			1 000	**0.032 3**	779.21	98.98	1.27
SIFKMP(228)			1 000	0.063 9	824.97	99.01	1.12

表 4　训练规模为 10 000 条件下的识别结果

算法	训练规模	测试规模	支持向量	一次匹配时间/s	训练时间/s	识别率/%	偏差/%
KMP			394	0.881	1 015.2	94.22	1.49
IFKMP			998	1.500 5	2 088.7	98.69	**0.84**
PS-IFKMP	10 000	−:3 022	**186**	0.210 2	**850.80**	97.92	1.37
SIFKMP(59)			1 000	**0.058 2**	1 223.3	98.78	1.33
SIFKMP(228)			1 000	0.094 7	1 264.8	**98.79**	1.14

结果以直方图的形式进行展示, 3 种算法的一次匹配时间、支持向量个数、识别率和训练时间随训练规模的变化情况如图 1 所示. 图 1 中, 四柱形的训练规模依次为 2 500、5 000、7 500、10 000.

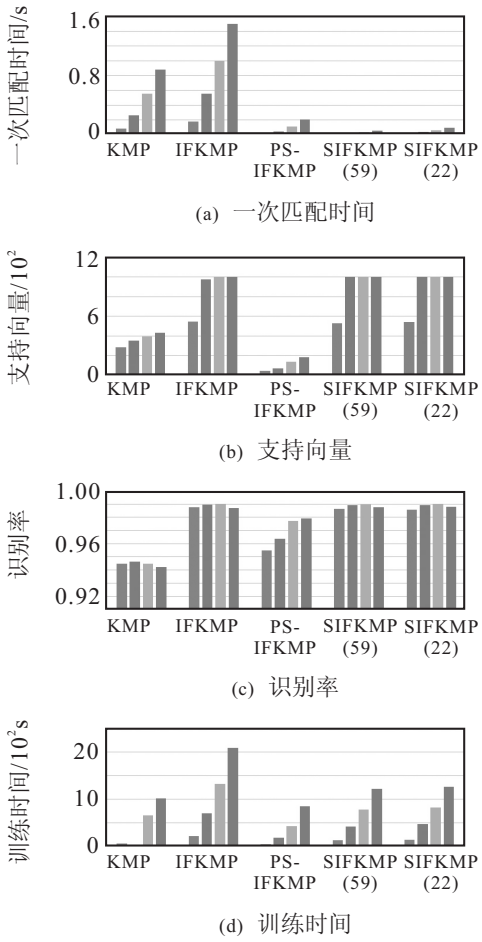

图 1 各算法性能随训练规模变化情况

由图 1(a) 可见, 对 Shuttle 数据集而言, IFKMP 算法的一次匹配时间最长, KMP 算法次之, PS-IFKMP 算法居中, SIFKMP 算法的一次匹配时间最短, 且远小于 KMP 算法和 IFKMP 算法. 同时, 两种核字典子集规模条件下, SIFKMP(228) 算法的一次匹配时间略大于 SIFKMP(59) 算法的一次匹配时间. 这是因为 SIFKMP 算法在每次迭代过程中仅从原始核字典集 D 中随机选取一个核字典子集进行搜索, 其一次匹配时间远小于 IFKMP 算法的一次匹配时间, 且核字典子集规模越大, 算法的一次匹配时间越长.

由图 1(b) 可见, PS-IFKMP 算法所需的支持向量数量最少, 表明 PS-IFKMP 算法具有较快的全局搜索能力, KMP 算法次之, IFKMP 算法和 SIFKMP 算法所需的支持向量数量较多且基本相当. 从理论上分析, SIFKMP 在每次匹配过程中只搜索到了近似最优基函数, 因此, 在相同的迭代误差阈值下, SIFKMP 算法所需的支持向量个数应略大于 IFKMP 算法. 同时注意到, 除训练规模为 2 500 外, 其他 3 种情况, SIFKMP 算法的支持向量个数与 IFKMP 算法基本相等, 与理论设想情况不符. 这是由于在这 3 种训练规模情况下, 算法迭代误差还没达到阈值就已达到停机条件 (最大迭代次数 $L = 1\,000$) 所致.

由图 1(c) 可见, KMP 算法的识别率最差, PS-IFKMP 算法次之, IFKMP 算法的识别效果最好, SIFKMP 算法的总体识别率只稍逊于 IFKMP 算法, 两者基本相当. 这是由于传统 KMP 算法在面对非平衡训练样本集 (即两类样本数目差距较大) 时, 对其中重要样本 (弱势样本) 识别效果不佳. IFKMP 有效地解决了这个问题, 通过对弱势样本的充分学习, 识别率可以达到 98% 以上. SIFKMP 算法虽然在每次匹配过程中只搜索到了近似最优基函数, 但理论上只要有足够多的迭代次数, 其识别精度应与 IFKMP 算法基本相当.

由图 1(d) 可见, IFKMP 算法训练时间最长, PS-IFKMP 算法训练时间最短, KMP 和 SIFKMP 算法的训练时间基本相等. 这是由于 IFKMP 算法的一次匹配时间和所需的支持向量个数均大于传统 KMP 算法, 训练时间远大于其他 KMP 算法. PS-IFKMP 算法所需的一次迭代时间大于 SIFKMP 算法, 但由于其局部搜索能力较强, 所需的支持向量个数远小于 SIFKMP 算法, 这意味着 PS-IFKMP 算法完成训练所需的迭代次数也远小于 SIFKMP 算法, 总的训练时间相对最短. SIFKMP 算法因其大大缩短了一次匹配时间, 总体训练时间也相对较短. 同时, 训练规模越大, SIFKMP 算法在训练时间上的优化效果也体现得越明显. 另外, 从实验偏差看, SIFKMP 算法在稳定性上并不逊色于 IFKMP 算法, 可见, SIFKMP 算法在性能与 IFKMP 算法相当的情况下, 大大降低了一次匹配时间, 总的训练时间也明显小于 IFKMP 算法.

3.2 弹道目标 HRRP 数据集识别

在未来反导作战中, 如何将弹头从诱饵中识别出来一个是技术难题, 因此本实验将弹头目标设为重

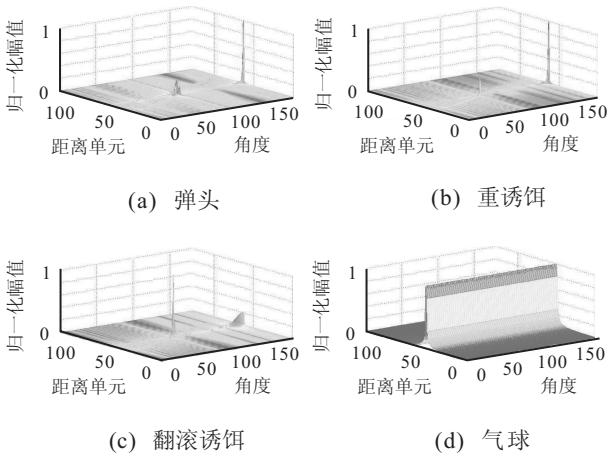

图 2 4 种目标在 0° ~ 180° 范围内的 HRRP 数据

要识别目标类别, 对SIFKMP算法进行测试. 如图2所示, 实验数据来源于电磁仿真软件生成的弹头、重诱饵、翻滚诱饵和气球4类目标HRRP数据[22]. HRRP数据角度范围为0°～180°, 间隔为0.1°, 频点为128个. 实验数据为样本数目为7 204、维数为128维的HRRP数据集, 4类样本的数目均为1 801.

考虑到直觉模糊核匹配追踪学习机为二分类器, 令弹头HRRP数据为正类, 其余3类目标为负类样本. 根据验证, 令核参数 $\sigma = 0.1$, 类别 y_1 的直觉模糊参数 $\omega(y_1) = 1.2$, 类别 y_2 的直觉模糊参数 $\omega(y_2) = 0.7$, 设置SIFKMP算法的核字典子集规模为59 ($\varepsilon = 0.95$, $\tau = 0.05$). 令粒子种群规模为150, 最大迭代次数 $L = 500$, 迭代误差阈值 $\eta = 0.02$, 50次蒙特卡洛仿真实验结果如表5所示.

表 5　弹道中段目标HRRP数据集的识别结果

算法	训练规模	测试规模	支持向量	一次匹配时间/s	训练时间/s	识别率/%	偏差/%
KMP			139	0.09	107.25	89.76	2.21
IFKMP			99	0.096	112.93	**95.60**	1.23
PS-IFKMP	3 000	+:500	**39**	0.014	**53.19**	94.10	1.06
SIFKMP(59)			219	**0.007**	62.29	95.25	**1.01**

综合以上3组实验结果可以发现, 传统KMP算法无论是面对非平衡UCI数据集、人工含噪数据集或是弹道中段目标HRRP数据集, 均无法对其中重要样本进行高精度识别. IFKMP算法将传统的KMP算法拓展到直觉模糊领域, 通过对重要样本进行充分学习, 使识别率达到令人满意的效果. 但IFKMP算法仍采用贪婪策略搜索最优基函数的线性组合, 从对Shuttle数据集的实验结果看, 当训练样本达到10 000时, IFKMP算法的一次训练时间长达34.8 min, 因此训练时间过长的缺陷将限制算法在大数据量问题上的应用. SIFKMP算法在每次迭代过程中仅从原始核字典集 D 中随机选取一个核字典子集进行搜索, 其一次匹配时间远小于IFKMP算法的一次匹配时间, 虽然每次匹配过程中, SIFKMP算法仅能以搜索到当前近似最优基函数, 导致所需的支持向量个数多于IFKMP算法, 但总体训练时间仍明显优于IFKMP算法, 且训练规模越大, 迭代次数越多, 优化效果体现得越明显. 从识别效果看, SIFKMP算法的识别效果与IFKMP算法基本相当, 都略优于KMP算法和PS-IFKMP算法. 这是因为从整个训练过程看, 在算法迭代次数足够多的前提下, SIFKMP算法有可能得到与IFKMP算法收敛精度相当的判决函数. 因此SIFKMP算法能够较好地应用于需要兼顾识别率和实效性的弹道目标识别领域.

虽然所提出的SIFKMP算法大大降低了算法的一次匹配时间, 进而有效降低了算法的总体训练时间, 但总体训练时间下降比例远达不到一次匹配时间所下降的比例. 从实验结果中也可以看出, 在Shuttle数据识别实验中, 当训练样本达到10 000时, SIFKMP(59)的一次匹配时间与IFKMP算法相比下降了96.12%, 而总体训练时间却只下降了41.4%. 这是由于SIFKMP算法仍需要采用核方法生成函数字典库, 这是一个相当费时的计算过程, 总体训练时间的优化程度远不如一次匹配时间明显.

注意到, 本节算法的提出基于 $|\langle \boldsymbol{r}, \boldsymbol{\omega} e \boldsymbol{g} \rangle|/||\boldsymbol{\omega} e \boldsymbol{g}||$ 的值服从(0,1)上的均匀分布这一假设. 但在实际实验过程中, 通过经验观察到 $|\langle \boldsymbol{r}, \boldsymbol{\omega} e \boldsymbol{g} \rangle|/||\boldsymbol{\omega} e \boldsymbol{g}||$ 的值近似服从指数分布, 而非(0,1)上的均匀分布. 将指数分布函数代入式(14), 可得

$$[F(\zeta)^n] = (1 - \exp(-\zeta))^n. \tag{18}$$

因此, 最小值满足 $x_\tau = \varepsilon$, 极值 ζ 的 τ 分位数可表示为

$$F^{-1}(\tau) = -\ln(1 - \tau^{1/n}), \tag{19}$$

其中 n 的取值为

$$-\ln(1 - \tau^{1/n}) = \varepsilon \Rightarrow 1 - \tau^{1/n} = \exp(-\varepsilon) \Rightarrow$$
$$n = 1/\log_\tau(1 - \exp(-\varepsilon)). \tag{20}$$

与第2.2节假设中提出的分位数 $\varepsilon^{1/n}$ 相比, $-\ln(1 - \tau^{1/n})$ 是一个更为激进的分位数(例如, 在 $\varepsilon = 0.95$, $\tau = 0.05$ 的条件下, n 的取值仅为7), 可见, 前文提出 $|\langle \boldsymbol{r}, \boldsymbol{\omega} e \boldsymbol{g} \rangle|/||\boldsymbol{\omega} e \boldsymbol{g}||$ 的值服从(0, 1)上的均匀分布这一假设实际上是一保守假设(或悲观假设), 这也与实验中发现SIFKMP的识别效果比预计的更好这一情况相吻合.

4　结　论

本文基于弱贪婪策略提出了一种随机直觉模糊核匹配追踪快速识别算法, 通过在原始核字典集 D 中随机选取一个核字典子集进行搜索, 较好地克服了原有算法计算量大的缺点. 实验结果表明, 与传统方法相比, 所提出方法在保持识别精度相当的情况下有效缩短了算法的计算时间, 且字典集规模越大, 优化效果体现得越明显, 能够较好地应用于大数据问题. 但是, 所提出算法仍有需要完善的地方, 如何进一步降低计算核函数字典所带来计算量和核参数的确定方法, 均是下一步亟待解决的问题.

参考文献(References)

[1] Pascal V, Bengio Y. Kernel matching pursuit[J]. Machine Learning, 2002, 48(1/2/3): 165-187.

[2] Cevher V, Krause A. Greedy dictionary selection for sparse representation[J]. IEEE J of Selected Topics Signal Processing, 2011, 5(5): 979-988.

[3] Sun P, Yao X. Sparse approximation through boosting for learning large scale kernel machines[J]. IEEE Trans on Neural Networks, 2010, 21(6): 883-894.

[4] Cho J, Lee M, Chang H J, et al. Robust action recognition using local motion and group sparsity[J]. Pattern Recognition, 2014, 47(5): 1813-1825.

[5] Amit B, Guy W, Amir A. Cover-based bounds on the numerical rank of Gaussian kernels[J]. Applied and Computational Harmonic Analysis, 2014, 36(2): 302-315.

[6] Zhang X H, Saha A, Vishwanathan S V N. Accelerated training of max-margin markov networks with kernels[J]. Theoretical Computer Science, 2014, 519: 88-102.

[7] Saitoh S. Theory of reproducing kernels and its applications[M]. Harlow: Longman Scientific & Technical, 1988: 16-21.

[8] Popovici V, Bengio S, Thiran J P. Kernel matching pursuit for large datasets[J]. Pattern Recognition, 2005, 38(12): 2385-2390.

[9] 缑水平, 焦李成, 张向荣. 基于免疫克隆的核匹配追踪集成图象识别算法[J]. 模式识别与人工智能, 2009, 22(1): 79-85.
(Gou S P, Jiao L C, Zhang X R. Image recognition with kernel matching pursuit classifier ensemble based on immune clone[J]. Pattern Recognition and Artificial Intelligence, 2009, 22(1): 79-85.)

[10] 付丽华, 李宏伟, 张猛. 基于更贪心策略的快速正交核匹配追踪算法[J]. 电子学报, 2013, 41(8): 1580-1585.
(Fu L H, Li H W, Zhang M. Fast orthogonal kernel matching pursuit based on greedier strategy[J]. Acta Electronica Sinica, 2013, 41(8): 1580-1585.)

[11] 雷阳, 孔韦韦, 雷英杰. 基于直觉模糊c均值聚类核匹配追踪的弹道中段目标识别方法[J]. 通信学报, 2012, 33(11): 136-143.
(Lei Y, Kong W W, Lei Y J. Technique for target recognition based on intuitionistic fuzzy c-means clustering and kernel matching pursuit[J]. J on Communications, 2012, 33(11): 136-143.)

[12] Li J W, Lu Y. Refining kernel matching pursuit[J]. Lecture Notes in Computer Science, 2010, 6064(1): 25-32.

[13] 李青, 焦李成, 周伟达. 基于模糊核匹配追寻的特征模式识别[J]. 计算机学报, 2009, 32(8): 1687-1694.
(Li Q, Jiao L C, Zhou W D. Pattern recognition based on the fuzzy kernel matching pursuit[J]. Chinese J of Computers, 2009, 32(8): 1687-1694.)

[14] 雷阳, 雷英杰, 周创明. 基于直觉模糊核匹配追踪的目标识别方法[J]. 电子学报, 2011, 39(6): 1441-1446.
(Lei Y, Lei Y J, Zhou C M. Techniques for target recognition based on intuitionistic fuzzy kernel matching pursuit[J]. Acta Electronica Sinica, 2011, 39(6): 1441-1446.)

[15] Schonberger V M. Big data: A revolution that will[M]. England: Hodder Export, 2013: 31-35.

[16] Temlyakov V. Weak greedy algorithms[J]. Advances in Computational Mathematics, 2000, 12(2/3): 213-227.

[17] Sil'nichenko A V. On the convergence of order-preserving weak greedy algorithms[J]. Mathematical Notes, 2008, 84(5/6): 741-747.

[18] Guebbai H, Grammont L. A new degenerate kernel method for a weakly singular integral equation[J]. Applied Mathematics and Computation, 2014, 230: 414-427.

[19] Livshits E D. On n-term approximation with positive coefficients[J]. Mathematical Notes, 2007, 82(3/4): 332-340.

[20] 徐志科, 平根建. 非参数方法估计分位数模型的研究综述[J]. 数学的实践与认识, 2014, 44(1): 151-156.
(Xu Z K, Ping G J. The summary of nonparametric quantile estimation[J]. Mathematics in Practice and Theory, 2014, 44(1): 151-156.)

[21] 余晓东, 雷英杰, 岳韶华. 基于粒子群优化的直觉模糊核匹配追踪算法[J]. 电子学报, 2015, 43(7): 1309-1314.
(Yu X D, Lei Y J, Yue S H. Research of pso-based intuitionistic fuzzy kernel matching pursuit algorithm[J]. Acta Electronica Sinica, 2015, 43(7): 1309-1314.)

[22] 冯德军, 王博, 王伟. 弹道中段目标雷达识别研究进展综述[J]. 中国电子科学研究院学报, 2013, 8(2): 142-147.
(Feng D J, Wang B, Wang W. Overview of progress in midcourse radar target recognition[J]. J of CAEIT, 2013, 8(2): 142-147.)

(责任编辑：郑晓蕾)