doi:10.11959/j.issn.1000-436x.2015260

基于直觉模糊核匹配追踪集成的目标识别方法

余晓东[1]，雷英杰[1]，宋亚飞[1]，岳韶华[1]，胡军红[2]
（1.空军工程大学 防空反导学院，陕西 西安 710051；2.中国人民解放军 94936 部队，浙江 杭州 310021）

摘 要：针对现有直觉模糊核匹配追踪算法采用部分样本进行训练和停机策略而导致学习机泛化能力下降的缺陷，结合集成学习的思想，提出了一种基于直觉模糊核匹配追踪集成的目标识别方法。该算法通过采用样本扰动和参数扰动的二重扰动策略产生子学习机，并利用多数投票法对其识别结果进行融合，从而提高了集成学习机的分类精度和泛化性能。实验结果表明，与传统方法相比，该方法可获得更优的识别效果，有效提高了识别精度，并能避免采样学习带来的不稳定性。

关键词：直觉模糊集；核匹配追踪；集成学习；目标识别
中图分类号：TP182; TP391 **文献标识码**：A

Intuitionistic fuzzy kernel matching pursuit ensemble based target recognition

YU Xiao-dong[1], LEI Ying-jie[1], SONG Ya-fei[1], YUE Shao-hua[1], HU Jun-hong[2]
(1. Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China; 2. Troop 94936, PLA, Hangzhou 310021, China)

Abstract: Considering that the generalization of the learning machine performed poorly in the present intuitionistic fuzzy kernel matching pursuit algorithm (IFKMP) due to its training method and stopping criteria, a new recognition method based on intuitionistic fuzzy kernel matching pursuit ensemble (IFKMPE) was proposed by introducing the idea of ensemble learning. In IFKMPE, the double perturbation strategy including sample and parameter perturbation was applied to generate the sub-learning machine, the recognition results were fused by the principle of majority voting, and therefore both the classify accuracy and generation ability were enhanced. Simulation results show the new algorithm IFKMPE performs better in terms of recognition accuracy and stability of sample learning compared with the traditional ones.
Key words: intuitionistic fuzzy set; kernel matching pursuit; ensemble learning; target recognition

1 引言

2002 年，Pascal Vincent 和 Yoshua Bengio[1]提出了一种崭新有效的核机器学习方法——核匹配追踪(KMP, kernel matching pursuit)，其主要思想源自于信号处理中的匹配追踪(MP, matching pursuit)算法及支持向量机中的核方法。通过核函数将训练样本集投影到高维空间，从而形成一组基函数字典库，然后利用贪婪算法在函数字典库中搜索一组基函数的线性组合来逼近目标函数。同其他核方法相比较，KMP 可以使用任意的核，而不必满足 Mercer 条件，并可采用多个核函数生成函数字典库。核匹配追踪学习机的性能与支持向量机相当，却有着更为稀疏的解。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[2~5]。

虽然核匹配追踪理论已经在模式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）比其余目标的更高，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降低识别精度要求。例如反导作战中，对真弹头的识别精度要求远远大于对诱饵、碎片等其他目标的识别精度要求。此外，在很多

收稿日期：2015-02-05；**修回日期**：2015-08-27
基金项目：国家自然科学基金资助项目(61272011, 61309022, 61402517)
Foundation Item: The National Natural Science Foundation of China (61272011, 61309022, 61402517)

实际问题中，两类样本的数目往往是不平衡的，例如医学实验的病例样本，阴性样本的数量明显大于阳性样本，这样对弱势阳性样本的识别就异常困难。传统的核匹配追踪学习机在进行学习的时候对待所有训练样本均是平等的，因此，判决函数是针对所有训练样本的一个综合考虑，预期达到总识别误差最小，而无法针对某一类指定样本进行有效识别，这就限制了核匹配追踪理论在很多特殊性问题中的应用[6]。针对这个问题，文献[7]提出了直觉模糊核匹配追踪学习机，把核匹配追踪算法拓展到直觉模糊理论领域，通过将直觉模糊参数有效地赋值给不同的目标样本，解决了对特殊样本进行高精度识别这一难题。但是面对大规模样本数据时，直觉模糊核匹配追踪学习机仍然只是选取部分样本进行训练，同时由于采用贪婪策略及在优化过程中使用停机条件，因此该学习机泛化性能下降的问题并没有得到解决。

由于外界条件的限制及自身存在的各种缺陷，单一学习机的泛化能力往往难以满足实际应用的需求。1990 年，Hansen 和 Salamon[8]提出了神经网络集成，显著地提高了系统的泛化能力，从而将研究者带入了集成学习这一重要领域。特别是 Schapire[9]证明了多个弱分类器可以集成为一个强分类器，从而奠定了集成学习的理论基础，近年来，集成学习的研究已成为机器学习领域的一个热门方向[10~12]。文献[13]将核匹配追踪学习机与集成学习理论相结合，提出了集成核匹配追踪学习机，提升了算法的识别精度及泛化性能，但是该方法没有拓展到直觉模糊领域，实际应用中无法对指定重要目标进行高精度识别。鉴于此，本文在对现有直觉模糊核匹配追踪算法和集成学习方法研究的基础上，从理论上分析了建立直觉模糊核匹配追踪集成学习机的可行性，并依此构建了直觉模糊核匹配追踪集成学习机，提出了一种基于直觉模糊核匹配追踪集成（IFKMPE, intuitionistic fuzzy kernel matching pursuit ensemble）的目标识别算法。为了对 IFKMPE 算法实际分类效果及有效性进行验证，选取 UCI 数据集及人工含噪数据集进行仿真实验，并将其与 KMP、核匹配追踪集成算法（KMPE）及直觉模糊匹配追踪算法（IFKMP）的分类效果进行比较，充分表明了 IFKMPE 算法的优越性和有效性。

2 直觉模糊核匹配追踪算法

直觉模糊核匹配追踪算法的基本思想：针对样本重要性程度的不同，对不同类别的样本赋予不同的直觉模糊参数 ω_i，ω_i 较大的样本则对其进行充分学习，尽可能保持对该类样本识别正确，对 ω_i 较小的样本则只进行粗略学习，这样直觉模糊核匹配追踪学习机就能对指定样本进行高精度识别。

定义 1 （$\odot$ 运算[6]）：对于 2 个向量 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_m\}$ 和 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_m\}$，向量之间的 $\odot$ 运算定义为

$$\boldsymbol{x}\odot\boldsymbol{y}=(x_1y_1,\cdots,x_my_m) \tag{1}$$

同时

$$\|\boldsymbol{x}\odot\boldsymbol{y}\|^2=\sum_{i=1}^{m}(x_iy_i)^2 \tag{2}$$

训练样本集为 $\{(x_1,y_1,\omega(y_1)),\cdots,(x_l,y_l,\omega(y_l))\}$，其中，$x_i\in\boldsymbol{R}^N$ 为样本特征值，$y_i\in\boldsymbol{R}$ 为训练样本观测值，ω_i 为直觉模糊参数，给定核函数 $K:\boldsymbol{R}^d\times\boldsymbol{R}^d\to\boldsymbol{R}$，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：$D=\{g_i=k(\cdot,x_i)\,|\,i=1,\cdots,l\}$。

为了有效地将直觉模糊参数 ω_i 赋予不同类别的样本，通过 $\odot$ 运算重新定义残差

$$\boldsymbol{r}_N=\boldsymbol{\omega}\odot(\boldsymbol{y}-\boldsymbol{f}_N)=\begin{bmatrix}\omega(y_1)(y_1-f_N(x_1))\\ \cdots\\ \omega(y_l)(y_l-f_N(x_l))\end{bmatrix} \tag{3}$$

其中，N 为迭代次数，$f_N(\boldsymbol{x}_i)=\sum_{i=1}^{N}\alpha_ig_i(x_i)$ 为 x_i 的观测估计值 $\hat{y}_i$，则重构误差为

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

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

$$\boldsymbol{g}_{N+1}=\arg\max_{\boldsymbol{g}\in D}\left|\left(\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}\odot\boldsymbol{g}\rangle}{\|\boldsymbol{\omega}\odot\boldsymbol{g}\|}\right)\right| \tag{5}$$

$$\alpha_{N+1}=\frac{\langle\boldsymbol{r}_N,\boldsymbol{\omega}\odot\boldsymbol{g}_{N+1}\rangle}{\|\boldsymbol{\omega}\odot\boldsymbol{g}_{N+1}\|^2} \tag{6}$$

最终得到判决函数为

$$f_N(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i\boldsymbol{g}_i(\boldsymbol{x})=\sum_{i\in\{sv\}}\alpha_ik_i(\boldsymbol{x},x_i) \tag{7}$$

其中，$\{sv\}$ 为直觉模糊核匹配追踪学习机所得的支持向量集。

3　直觉模糊核匹配追踪集成学习机

直觉模糊核匹配追踪学习机虽然有效地解决了传统核匹配追踪算法无法针对指定样本类别进行不同精度识别这一问题，但是面对大规模样本数据时，直觉模糊核匹配追踪学习机仍然只是选取部分样本进行训练，同时由于采用贪婪策略及在优化过程中使用停机条件，该学习机泛化性能下降的问题并没有得到解决。本节拟从理论上分析建立直觉模糊核匹配追踪集成学习机的可行性，并依此构建直觉模糊核匹配追踪集成学习机。

3.1　直觉模糊核匹配追踪集成学习机的理论分析

理论上，相比决策树、神经网络等传统的学习机，直觉模糊核匹配追踪学习机的性能更加稳定，其判决能力与支持向量机相当，却具有更为稀疏的解。但是直觉模糊核匹配追踪学习机在实际应用上仍然存在以下 2 个问题。

1) 推广性能的问题。当面对大规模训练样本时，算法所需的存储空间和训练时间会成倍增加，此时，直觉模糊核匹配追踪学习机通常只随机选择部分样本进行训练，这样虽然减小了训练规模，解决了计算量和存储量过大的问题，但是由于训练集没有包含整个样本集信息，其最终推广性能必然会下降。

2) 计算误差的问题。直觉模糊核匹配追踪学习机在搜索过程中实际采用的是贪婪算法，并且给出了迭代误差阈值。同时，该算法往往设置了最大迭代次数，这样虽然加快了算法的训练速度，但是所求的判决函数往往达不到最优，使学习机的性能进一步下降。

正是因为直觉模糊核匹配追踪学习机存在以上缺点，使其难以达到预期的分类效果。因此，考虑将集成方法和直觉模糊核匹配追踪学习机进行有效结合，从而达到提高分类性能的目的。Kim[12] 针对多个弱分类器集成为一个强分类器，提出了如下优越性条件定理。

定理 1　（优越性条件[9]）若要使集成学习机的错分误差减小，需要满足如下条件：1）集成学习机中的各子学习机互异；2）集成学习机中的各子学习机的错分误差均小于 $\frac{1}{2}$。

下面对采用集成学习方法是否能解决直觉模糊核匹配追踪学习机存在的 2 个问题分别进行分析。

1) 推广性能的角度。直觉模糊核匹配追踪学习机选择部分样本进行训练时，所得判决函数的性能则会下降。这是因为训练集只包含了全部样本的部分信息，所以该直觉模糊核匹配追踪学习机所得的判决域只是真正判决域的一个近似。但采用集成学习方式时，每个学习机的差异性尽可大，则直觉模糊核匹配追踪集成学习机的判决域则会得到扩展。

图 1 中，U 表示整个训练样本空间，T_i 为集成学习系统中第 i 个直觉模糊核匹配追踪学习机 h_i 的训练集空间，G_i 为第 i 个学习机的判决区域。假设集成学习系统由 $h_1(x)$、$h_2(x)$、$h_3(x)$ 3 个不相关的学习机组成，那么若子学习机 $h_1(x)$ 错分时，而 $h_2(x)$、$h_3(x)$ 却有可能正确，这样在决策阶段采用多数投票法就可以消除子学习机 $h_1(x)$ 错分的影响，做出正确的判决结果。

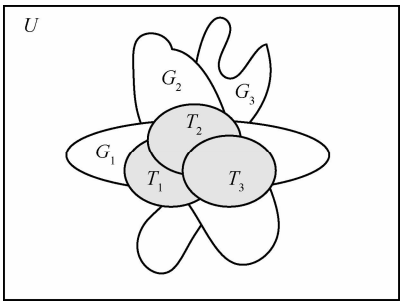


图 1　集成学习系统

假设有 l 个相互独立的子学习机，每个子学习机的错分误差均为 p，如果采用投票策略，则整个集成学习机的错分误差可表示为

$$\tilde{p}=\sum_{k=\left\lceil \frac{l}{2} \right\rceil}^{l} C_l^{\left\lceil \frac{l}{k} \right\rceil} p^k (1-p)^{(l-k)} \tag{8}$$

可得，$\tilde{p}$ 为一个二项分布，当 $p<\frac{1}{2}$ 时，满足

$$\tilde{p}<\sum_{k=\left\lceil \frac{l}{2} \right\rceil}^{l} \left(\frac{1}{2}\right)^l \tag{9}$$

从式（9）可以看出，只要满足 $p<\frac{1}{2}$，随着子学习机个数 l 的增加，集成学习机的错分误差会越来越小，直至趋近于 0。因此，只要满足优越性条

件，应用集成学习方法，可以有效地解决直觉模糊核匹配追踪学习机存在的问题。当然，实际应用中，集成学习机的错分误差 $\tilde{p}$ 也不可能无限降低，训练样本是有限的，随着子学习机数目的增加，子训练集之间的相关性也会随之增加，从而导致子学习机之间的互异性减弱，由定理 1 可知，这也会导致集成学习机错分误差增加，但相对单个学习机而言，集成学习机仍可以在一定程度上提高泛化性能。

2) 计算补偿的角度。觉模糊核匹配追踪学习机采用贪婪算法来搜索一种基函数的线性组合来匹配观测值。但实际应用过程中，这组基函数的线性组合只能无限逼近观测值，却不能等价于观测值。因此通过 2 种方式作为停机条件：①设置残差阈值，在每次迭代后计算残差，判断残差是否达到阈值，若达到则终止训练并输出当前结果；②设置最大迭代次数 N，当迭代次数达到 N 时，算法停止训练。这样虽然加速了训练速度，但所求解往往只是最优解得一个近似值，从而降低学习机的判决性能。

如图 2 所示，M 是理论最优值，f_1、f_2 和 f_3 分别是单个直觉模糊核匹配追踪学习机的最优解，虽然 f_1、f_2 和 f_3 距离理论最优值 M 很近，但均距最优值有一定差距；因此采用 f_1、f_2 和 f_3 分别进行预测，其识别误差必然大于期望误差。而采用集成学习方法则可能平均单个学习机的识别误差，并将其进一步逼近期望误差。

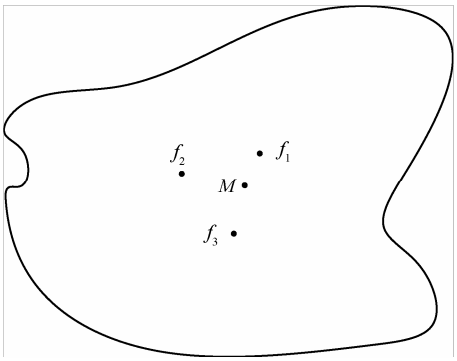


图 2 计算补偿

3.2 直觉模糊核匹配追踪集成学习机的实现

本节详细叙述如何构建直觉模糊核匹配追踪集成学习机。图 3 给出了直觉模糊核匹配追踪集成学习机的结构。

在训练阶段，需要尽可能按一定策略从原始训练数据集中选出 l 组互异的数据集作为各子学习机的训练样本。针对如何生成子训练集可以采用如下策略。

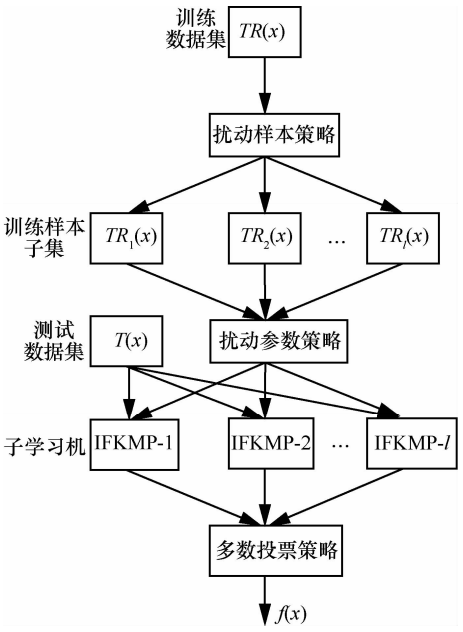


图 3 直觉模糊核匹配追踪集成学习机结构

1) 扰动样本策略。Bagging 算法[14]是目前最流行的一种扰动样本策略，也被称为随机采样策略。假设原始训练样本集包含 N 个训练样本，Bagging 算法运行一次就随机从原始训练样本集选取 m 个样本，反复运行 l 次算法就得到 l 个子训练集。但是，由于采用的是随机策略，原始训练样本集中的部分样本有可能被采样多次，而另外部分样本则可能不会被采样。为了避免这种情况，也可以采取另外一种采样策略——等间距采样策略(equidistance sampling)[13]，即按照设定好的间距，按顺序从原始训练样本集中选取子训练集。为了表示方便，分别把基于随机采样和等间距采样的 IFKMP 集成算法表示为 Bagging-IFKMPE 算法和 ES-IFKMPE 算法。

2) 扰动参数策略。基于扰动样本策略的各个子训练集所带来的数据扰动必然导致最优核参数的扰动。各个子训练集的最优核参数可以通过自动模型选择法搜索获得，但计算量较大[15]。而采用各子训练集的标准差作为核参数，也可以实现较高的分类精度[16]。因此本文将各子学习机训练样本集的标准差设置为该子学习机的核参数。

在测试阶段，将测试数据集输入 l 个训练好的子学习机，得到 l 个判决结果，再根据设定好的集成策略将 l 个判决结果进行融合，作出最终判决结果进行输出。本文采用最常用的集成策略——多数投票法。

若 $h_1, h_2, \cdots, h_l \left(h_i \in \{-1,1\}\right)$ 为各个子直觉模糊核

匹配追踪学习机的判决结果，则集成学习机的最终判决结果为

$$h_{\text{ensemble}} = \text{sgn}\left(\sum_{i=1}^{I} h_i\right) \tag{10}$$

具体的直觉模糊核匹配追踪集成算法描述如下。

输入：样本数据集 $\boldsymbol{X}=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，测试样本集 $\boldsymbol{D}$，核函数 $K(\boldsymbol{x},\boldsymbol{y})$，直觉模糊参数 $\omega(y_i)$，最大迭代次数 $T_{\max}$，迭代停止阈值 ε，集成规模 T，子训练集样本规模 L。

输出：最终判决结果 h_{ensemble}。

过程：

Step1　按照随机采样策略或等间距采样策略从训练样本集中选取 N 组训练子集 $\{(d_1,d_2,\cdots,d_N\}$；

Step2　计算各训练子集的标准差，并将其设置为各子学习机的核参数 σ；

Step3　对每个子学习机进行训练，满足停机条件后输出判决函数；

Step4　输入测试样本集 $\boldsymbol{D}$ 对每个子学习机进行测试，得到 N 组判决结果 $\{h_1,h_2,\cdots,h_N\}$；

Step5　按照多数投票法对 N 组判决结果进行融合，输出最终判决结果 h_{ensemble}。

4　实验结果及分析

本文实验过程中选取高斯核作为核函数。为了验证算法的有效性，将标准核匹配追踪算法、核匹配追踪集成算法、直觉模糊核匹配追踪算法和本文方法进行了对比。为了对算法性能进行验证，本文选择不同的样本集合进行实验。为了避免随机误差，每次实验分别进行 50 次蒙特卡洛仿真，并给出了实验偏差。仿真环境：操作系统 Windows XP，编译软件 Matlab7.6，Pentium(R)Dual-Core CPU E5500@2.8 GHz，内存 2 GB。

4.1　Musk 数据集测试

实验首先选取公共数据集 UCI 中的 Musk 进行验证，Musk 是描述麝香分子的数据集，具有 166 个特征属性，6 598 个样本包含 1 017 个正类样本和 5 581 个负类样本。从数据集样本的组成来看，正类样本在数量上处于弱势地位，因此本文将正类样本类别设定为指定类别样本。实验从全体数据集中随机选取 300 个样本作为训练集，并从 1 017 个正类样本中随机抽取 200 个样本为测试集。

参数设置：选取高斯核作为核函数，并将训练样本集的标准差设置为核参数 σ，设置最大迭代次数 L=200，迭代误差阈值 ε=0.05，直觉模糊参数根据文献[7]进行选取，可得：1)正类样本为指定类别样本 y_1 的直觉模糊参数 $\omega(y_1)$=1.5；2)负类样本为非指定类别样本 y_2 的直觉模糊参数 $\omega(y_2)$=0.3。分别采用随机采样策略和等间隔采样策略各生成 15 个子训练集，每个子训练集 300 个样本数据。实验结果如表 1 所示。

表 1　**Musk 数据集识别结果比较**

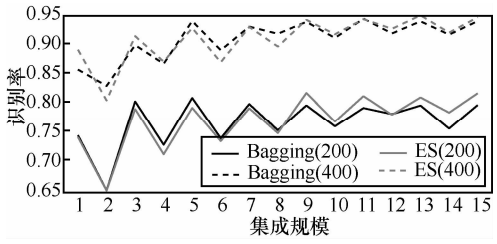
算法	训练规模	测试规模	集成规模	训练时间/s	识别率/%	偏差/%
KMP			1	2.863	75.58	2.21
KMPE			15	42.946	78.69	0.79
IFKMP	400	＋:500	1	4.443	93.24	2.17
Bagging-IFKMPE			15	68.199	1	0
ES-IFKMPE			15	65.692	1	0

从表 1 可以看出，传统核匹配追踪学习机由于其平等对待所有训练样本的特点，对弱势样本类别识别效果不好，无法达到提高对重要样本类别的识别精度的要求。改进后的核匹配追踪集成算法虽然改善了算法的泛化性能，但仍无法对指定类别样本进行高精度识别。直觉模糊核匹配追踪学习机通过对重要类别样本进行充分学习，对次要类别样本进行粗略学习，改善了对指定样本类别的识别效果，但其识别精度和泛化性能还有待于进一步提升。而通过采用集成策略，不论是随机采样集成还是等间采样均有效地提高了直觉模糊核匹配追踪学习机的识别性能和泛化能力。

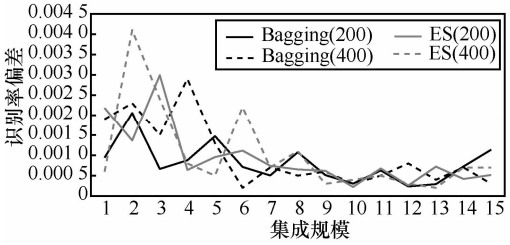
为了验证集成规模对本文算法的影响，其他参数不变，分别令子训练集样本规模 L 分别为 200 和 400，集成规模 T 在 1~15 之间对算法进行验证，具体实验结果如图 4 所示。

由图 4 可知，随着集成规模的增大，算法性能会迅速提升，并逐渐达到稳定。但算法稳定状态的性能与训练子集规模相关，子集规模越大，则集成学习系统处于稳定状态的性能越好。此外，在同等条件下，ES-IFKMPE 算法在识别率和稳定性方面可能要略优于 Bagging-IFKMPE 算法，其原因在于采用随机策略，训练样本集中的部分样本有可能被采样多次而另外部分样本则可能不会被采样，从而导致各自学习机间的互异性减弱，致使集成学习机性能下降。而 ES-IFKMPE 算法的虽然各子学习机的

互异性较强，但其集成规模则会因为样本规模的限制而无法大规模增加。此外，还观察到算法识别率随集成规模呈锯齿状上升趋势，其识别率波谷正好处于偶数集成规模状态，这是因为当集成规模为偶数时，有可能出现判决错误和判决正确的子分类器数目相等的情况从而导致识别率下降。



(a) 识别率



(b) 偏差

图 4 集成算法性能随集成规模变化情况

4.2 人工含噪数据集测试

实验选取三维空间内线性不可分的同心球样本进行测试，采用如下参数方程产生 2 类交错的同心球样本。

$$\begin{cases} x = \rho\cos\theta\sin\varphi \\ y = \rho\sin\theta\sin\varphi \\ z = \rho\cos\varphi \end{cases} \tag{11}$$

2 类样本的半径参数 ρ 均服从均匀分布，分别为[0,60]和[40,100]，随机产生 2 类样本共 12 000 个。

实验参数设置：选取高斯核作为核函数，并将训练样本集的标准差设置为核参数 σ，设置最大迭代次数 L=150，迭代误差阈值 ε=0.02，指定样本类别 y_1 的直觉模糊参数 $\omega(y_1)$=1.45，非指定类别样本 y_2 的直觉模糊参数 $\omega(y_2)$=0.35，要求尽可能提高对指定样本类别 y_1 的识别精度。直觉模糊核匹配追踪集成学习机参数：采用 2 种样本扰动策略，随机采样策略和等间隔采样策略各生成 25 个子训练集，每个子训练集 300 个样本数据。实验前先对训练数据进行加噪处理，随机改变 25%样本的类别属性，然后进行训练。50 次蒙特卡洛仿真实验结果如表 2 所示。

表 2 人工含噪数据集识别结果比较

算法	训练规模	测试规模	集成规模	训练时间/s	识别率/%	偏差/%
KMP	500		1	5.825	85.42	1.11
KMPE	500		25	109.26	89.59	0.69
IFKMP	500	+:500	1	6.474	93.36	1.02
Bagging-IFKMPE	200		15	31.994	97.72	0.34
ES-IFKMPE	200		15	31.238	97.50	0.38

从表 2 可以看出，在训练样本含噪声的情况下，本文提出的直觉模糊核匹配追踪集成学习机仍能对指定的重要样本仍能保持 97%以上的识别精度，其识别性能和泛化能力明显优于单一学习机以及传统的核匹配追踪集成学习机。

4.3 时间/性能代价测试

通过对上述的实验结果进行分析，采用集成策略，直觉模糊核匹配追踪学习机的识别性能和泛化能力确实得到了提升，但是随之也带来训练时间的大幅度增加。这是否说明直觉模糊核匹配追踪集成学习机的优越性能是牺牲训练时间换取的？鉴于此，本组实验拟对本节算法的性能/时间代价进行测试，令 KMP 算法、KMPE 算法及 IFKMP 算法选择较多的样本进行充分学习，而 IFKMPE 算法的各子分类器则选择较少的样本进行粗略学习，使集成学习机的训练时间小于 KMP 及 IFKMP 算法时，对三者的性能进行对比。

实验选取 UCI 数据库中的 Diabetes 数据集对集成学习机的性能/时间代价进行测试。Diabetes 数据集描述了印度妇女糖尿病病例检测情况，786 个样本包含 268 个正类（阳性）样本和 500 个负类（阴性）样本。从确诊病症的角度来看，由于正类样本描述了检测呈阳性的病理特征，对该类样本应该具备比负类样本更高的识别精度，因而本文将正类样本类别设定为指定类别样本

参数设置：选取高斯核作为核函数，并将训练样本集的标准差设置为核参数 σ，最大迭代次数 L=120，迭代误差阈值 ε=0.05，直觉模糊参数 $\omega(y_i)$ 根据算法 1 进行选取，此处 $\delta(y_i)$=0.8，可得：1)阳性样本为指定类别样本 y_1 的直觉模糊参数 $\omega(y_1)$=1.8；2)阴性样本为非指定类别样本 y_2 的直觉模糊参数 $\omega(y_2)$=0.1。其他参数不变，采用 2 种样本扰动策略，随机采样策略和等间隔采样策略各生成

5 个子训练集，每个子训练集 60 个样本数据。实验结果如表 3 所示。

表 3　　Diabetes 数据集识别结果比较

算法	训练规模	测试规模	集成规模	训练时间/s	识别率/%	偏差/%
KMP	300		1	1.357 8	67.63	5.68
KMPE	300		5	6.236	70.59	1.36
IFKMP	300	+:150	1	1.421 1	92.77	6.06
Bagging-IFKMPE	60		5	0.630 2	96.04	2.21
ES-IFKMPE	60		5	0.644 5	97.32	0.89

实验结果表明，通过参数设置，使核匹配追踪学习机及直觉模糊核匹配追踪学习机所耗费的训练时间超过集成直觉模糊核匹配追踪学习机时，集成学习机的识别性能和泛化能力仍明显优于经典的单一学习机，这也说明了直觉模糊核匹配追踪集成学习机的优越性能并不是以牺牲训练时间换取的。

5　结束语

本文首先对直觉模糊核匹配追踪算法及集成学习方法的原理进行了研究，从理论上分析了建立直觉模糊核匹配追踪集成学习机的可行性，并按照一定策略建立了具体的直觉模糊核匹配追踪集成学习机，从而克服了原有学习机在面对大规模数据样本时，仅采用部分样本进行训练和停机策略而导致泛化能力下降的缺陷。实验结果表明，直觉模糊核匹配追踪集成学习机相对单一学习机而言，具有更好的识别效果和泛化能力，并且算法性能会随着集成规模的增大逐渐提升并趋于稳定。但是，该算法仍有一些需要完善的地方，如何构建差异性更大的子学习机以及对子学习机进行选取等，这些都将是下一步重点研究的课题。

参考文献：

[1] PASCAL V, BENGIO Y. Kernel matching pursuit[J]. Machine Learning, 2002, 48(1-3): 165-187.

[2] CEVHERV, KRAUSE A. Greedy dictionary selection for sparse representation[J]. IEEE Journal 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 Transaction on Neural Networks, 2010, 21(6):883-894.

[4] 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.

[5] 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]. Journal on Communications, 2012, 33(11):136-143.

[6] 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 Journal of Computers, 2009, 32(8):1687-1694.

[7] 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.

[8] HANSEN L K, SALAMON P. Neural network ensemble[J]. IEEE Transactions on Pattern analysis and Machine Intelligence, 1990, 12(10):993-1001.

[9] SCHAPIRE R E. The strength of weak learn ability[J]. Machine Learning, 1990, 5(2): 197-227.

[10] BARTOSZ K, MICHAŁ W, BOGUSŁAW C. Clustering-based ensembles for one-class classification[J]. Information Sciences, 2014, 264(Complete):182-195.

[11] MONTHER A, WANG D H. Fast decorrelated neural network ensembles with random weights[J]. Information Sciences, 2014, 264:104-117.

[12] KIM H C, PAN S N. Constructing support vector machine ensemble[J]. Pattern Recognition, 2003, 36(12):2757-2767.

[13] JIAO L C, LI Q. Kernel matching pursuit classifier ensemble[J].Pattern Recognition, 2006, 39(4):587-594.

[14] MORDELET F, VERT J P. A bagging SVM to learn from positive and unlabeled examples[J]. Pattern Recognition Letters, 2014, 37: 201-209.

[15] YANG K H, ZHAO L L. A new optimizing parameter approach of LSSVM multiclass classification model[J]. Neural Computing & Applications, 2012, 21(5):954-955.

[16] 王晓丹, 孙东延, 郑春颖,等. 一种基于AdaBoost的SVM分类器[J]. 空军工程大学学报(自然科学版), 2006, 7(6):54-57.
WANG X D, SUN D Y, ZHENG C Y, *et al*. A combined SVM classifier based on AdaBoost[J]. Journal of Air Force Engineering University(Natural Science Edition), 2006, 7(6):54-57.

作者简介：

余晓东（1989-），男，江西九江人，空军工程大学博士生，主要研究方向为模式识别、智能信息处理等。

雷英杰（1956-），男，陕西渭南人，空军工程大学教授、博士生导师，主要研究方向为智能信息处理与智能决策。

宋亚飞（1988-），男，河南汝州人，博士，空军工程大学讲师，主要研究方向为信息融合、证据推理等。

岳韶华（1968-），女，湖北黄梅人，空军工程大学高级实验师、硕士生导师，主要研究方向为智能信息处理与智能决策。

胡军红（1978-），女，山东莱芜人，博士，94936 部队工程师，主要研究方向为智能信息处理。