DOI: 10.7641/CTA.2016.50266

# 根据混合选择策略的直觉模糊核匹配追踪集成算法

雷英杰[1], 余晓东[2†], 王　睿[1], 王　毅[1]

(1. 空军工程大学 防空反导学院, 陕西 西安 710051; 2. 空军装备研究院 装备总体论证研究所, 北京 100076)

**摘要**: 为了从分类器集成系统中选择出一组差异性大的子分类器, 从而提高集成系统的泛化能力, 提出了一种基于混合选择策略的直觉模糊核匹配追踪算法. 基本思想是通过扰动训练集和特征空间生成一组子分类器; 然后采用$k$均值聚类算法将对所得子分类器进行修剪, 删去其中的冗余分类器; 最后根据实际识别目标动态选择出较高识别率的分类器组合, 使选择性集成规模能够随识别目标的复杂程度而自适应地变化, 并基于预期识别精度实现循环集成. 实验结果表明, 与其他常用的分类器选择方法相比, 本文方法灵活高效, 具有更好的识别效果和泛化能力.

**关键词**: 直觉模糊匹配追踪; 选择性集成; 混合选择策略; 差异性度量; 泛化性能

**中图分类号**: TP182, TP391　　**文献标识码**: A

# Intuitionistic fuzzy kernel-matching pursuit ensemble algorithm based on hybrid selection strategy

LEI Ying-jie[1], YU Xiao-dong[2†], WANG Rui[1], WANG Yi[1]

(1. Air and Missile Defense Institute, Air Force Engineering University, Xi'an Shaanxi 710051, China;
2. Research Institute on General Development and Evaluation of Equipment, Air Force Equipment Academy, Beijing 100076, China)

**Abstract:** In order to improve the generalization ability of a classifier ensemble, we propose an intuitionistic fuzzy kernel-matching pursuit algorithm based on the hybrid selection strategy for target recognition to select a subset of optimal individuals from the given classifier ensemble. The basic idea of this algorithm is to produce a preliminary subset of classifiers by disturbing the training set and the feature space, and then trim this subset by eliminating the redundant classifiers based on $k$-means clustering algorithm and dynamically singling out classifiers with high differentiability from practical object recognition, making the size of the subset adaptively change according to the complexity of the objects and the expected accuracy of recognition be determined recursively. Experimental results show that the performance of the proposed algorithm is more flexible, efficient and accurate, with higher generalization, in comparison to other classifier selection methods.

**Key words:** intuitionistic fuzzy kernel matching pursuit; selective ensemble; hybrid selection strategy; diversity measure; generalization performance

## 1　引言(Introduction)

2002年, Pascal Vincent和Yoshua Bengio[1]提出了一种新的核机器学习方法—核匹配追踪(kernel matching pursuit, KMP), 其主要思想源自于信号处理中的匹配追踪(matching pursuit, MP)算法及支持向量机(support vector machine, SVM)中的核方法. KMP学习机的性能与SVM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]. KMP理论虽然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则远远大于对诱饵、碎片等其他目标的识别精度要求. 然而经典KMP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KMP理论在很多有特殊性要求问题中的应用[6]. 针对这个问题, 文献[7]提出了直觉模糊核匹配追踪(intuitionistic fuzzy kernel matching pursuit,

收稿日期: 2015−04−02; 录用日期: 2015−10−09.
†通信作者. E-mail: agosoa@163.com; Tel.: +86 18165296878.
本文责任编委: 王耀南.
国家自然科学基金项目(61272011, 61309022)资助.
Supported by National Natural Science Foundation of China (61272011, 61309022).

IFKMP)学习机, 把KMP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, IFKMP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. 2002年, Zhou等[10]指出通过选择部分分类器构建的集成学习系统的性能或许要优于使用全体分类器构建的集成学习系统, 并提出了选择性集成的概念. 目前, 针对集成学习系统的研究已不再局限于集成方法的提出及改进, 而更多的着眼于分类器的选择方法[11]. 从目前来看, 选择性集成方法可主要分成4大类: 基于聚类的选择方法, 基于顺序的选择方法, 基于优化的选择方法以及其他方法[12]. 文献[13]利用聚类算法对候选分类器进行聚类, 然后从每个类中选择一个分类器用于集成. 文献[14]通过前向或后向顺序选择方法将那些不能提高集成性能的分类器从集合中删除. 文献[15]基于不同分类器模型之间的互补性, 提出了一种分类器的动态选择与循环集成方法. 同时, 大量的分类器选择标准也相继被提出和使用, 如集成精度、分类器差异性度量等. 文献[16–17]对多种分类器差异性度量方法进行了总结、并验证了这些方法的有效性. 但上述研究工作基本上都只是对某一种选择性集成策略进行研究或改进, 因此或多或少都存在其局限性, 例如聚类方法的不稳定性会导致集成系统性能的不稳定, 而顺序选择方法则需要大量的时间及存储空间来训练分类器, 而优化选择方法需要经过大量的尝试才能找到最优解[18]. 此外, 这些方法基本上系统结构都相对固定, 分类器一旦选定好了就不再变动, 缺乏足够的灵活性[19].

事实上, 通过采用混合策略来提升系统性能也是机器学习领域的一种有效手段[20]. 鉴于此, 本文尝试结合Bagging集成方法、$k$均值聚类算法以及动态选择和循环集成算法这3种策略的优势, 并基于此提出一种基于混合选择策略的直觉模糊核匹配追踪集成方法. 为了对本文方法进行验证, 选取UCI数据集、人工含噪数据集及弹道目标数据集进行仿真实验, 并与其他分类器选择方法进行比较, 实验表明本文方法具有更好的识别效果及泛化性能.

## 2　集成直觉模糊核匹配追踪学习机的理论分析(Theories analysis for intuitionistic fuzzy kernel matching pursuit ensemble)

### 2.1　直觉模糊核匹配追踪算法(Intuitionistic fuzzy kernel matching pursuit)

直觉模糊核匹配追踪算法的基本思想[7]: 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$\omega$, 对$\omega$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$\omega$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$\odot$运算定义为

$$x\odot y=(x_1\cdot y_1,\cdots,x_m\cdot y_m). \tag{1}$$

同时

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

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$x_i\in\mathbb{R}^N$为样本特征值, $y_i\in\mathbb{R}$为训练样本观测值, $\omega_i$为直觉模糊参数, 给定核函数$K:\mathbb{R}^d\times\mathbb{R}^d\to R$, 利用训练样本处的核函数值生成函数字典库$D=\{g_i=k(\cdot,x_i)\,|i=1,\cdots l\}$. 定义残差为

$$r_N=\omega\odot(y-f_N)=\begin{bmatrix}\omega(y_1)(y_1-f_N(x_1))\\ \vdots\\ \omega(y_l)(y_l-f_N(x_l))\end{bmatrix}, \tag{3}$$

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

$$\|r_N\|^2=\|\omega\odot(y-f_N)\|^2=\sum_{i=1}^{l}(\omega(y_i)(y_i-f_N(x_i)))^2. \tag{4}$$

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

$$g_{N+1}=\arg\max_{g\in D}|(\frac{\langle r_N,\omega\odot g\rangle}{\|\omega\odot g\|})|, \tag{5}$$

$$\alpha_{N+1}=\frac{\langle r_N,\omega\odot g_{N+1}\rangle}{\|\omega\odot g_{N+1}\|^2}. \tag{6}$$

最终得到判决函数为

$$f_N(x)=\sum_{i=1}^{N}\alpha_i g_i(x)=\sum_{i\in\{sv\}}\alpha_i k_i(x,x_i), \tag{7}$$

其中$\{sv\}$为直觉模糊核匹配追踪学习机所得的支持向量集.

### 2.2 理论分析(Theoretical analysis)

理论上, 相比传统的学习机, 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两个问题.

1) 泛化性能的问题. 当面对大规模训练样本时, 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, 这样虽然减少了训练规模, 但是由于训练集没有包含整个样本集信息, 其最终泛化性能必然会下降.

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阀值. 同时, 该算法往往设置了最大搜索次数, 这样虽然加快了算法的训练速度, 但是所求的解通常难以达到最优, 使得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, 使得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等[21]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)[21] 若要使集成学习机的错分误差减小, 须满足如下条件: 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1/2.

只要满足优越性条件, 应用集成学习方法, 是可以有效的解决直觉模糊核匹配追踪学习机存在的问题[22]). 当然, 实际应用中, 集成学习机的错分误差也不可能无限降低, 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, 从而导致集成系统的性能的下降. 因此, 选择那些差异性大的子学习机才是提升集成泛化性能的有效途径.

## 3 基于混合选择策略的 IFKMP 集成学习机 (IFKMP classifier ensemble based hybrid selection strategy)

针对如何选择差异性大的分类器的问题, 文献[23]提出了一种“overproduce and choose”的筛选策略, 即先产生的大量的冗余分类器, 然后从中选择选择差异度较大的个体形成一个子集, 该策略允许同时采用多种方法对分类器进行选择, 以增加子分类器间的差异性). 基于这个思想, 本文尝试结合Bagging集成算法、$k$ 均值聚类算法以及动态选择和循环集成算法3种集成策略的优势来对分类器进行选择, 并提出一种基于混合选择策略的直觉模糊核匹配追踪集成方法. 该算法包含3个阶段: 第1阶段是子分类器生成阶段, 采用基于Bagging框架的双重扰动策略生成一定数量的子训练集, 然后通过直觉模糊核匹配追踪学习算法训练出相应的子分类器; 第2阶段则采用$k$均值聚类算法将对训练所得的多个子分类器进行聚类, 然后从每个类别中选出一个分类器进入到第3阶段候选, 目的是在保持分类器多样性的同时, 删去一些冗余的分类器; 第3阶段则采用动态选择和循环集成策略对已获得的子分类器进行二次选择, 先基于集成系统的差异度对候选分类器进行排序, 然后动态选择出对实际测试目标有较好识别结果的分类器组合, 使参与集成的分类器的个数能够随识别问题的复杂程度而自适应变化, 若候选分类器都参与集成仍无法达到识别要求时, 则通过降低预期识别精度来实现分类器的循环集成. 整个算法流程如图1所示.

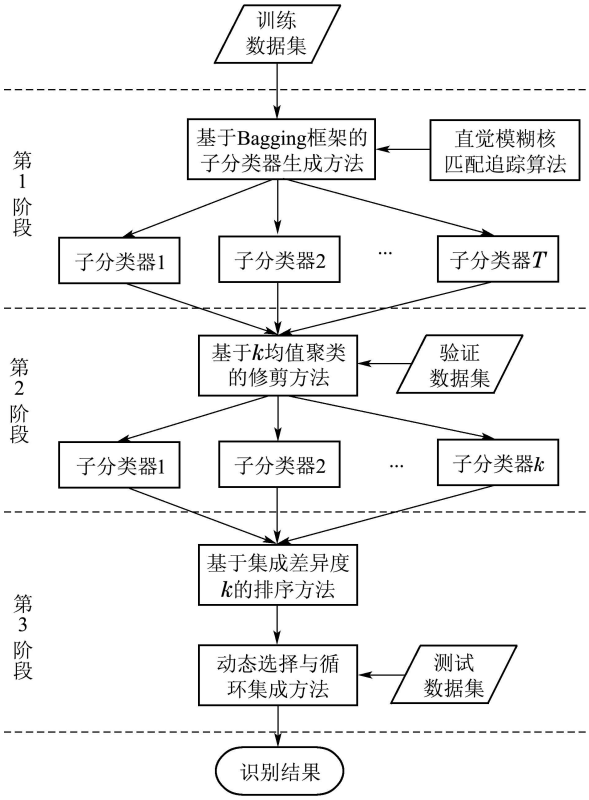


图 1 算法流程

Fig. 1 The algorithm flow

### 3.1 子分类器的生成(The production of individual classifier)

第1阶段是子分类器的生成阶段. 为了降低训练样本中冗余特征及噪声数据对子分类器性能的影响, 首先对训练样本进行预处理. 偏最小二乘(partial least squares, PLS)是近年来提出的一种基于主成分分析(principal component analysis, PCA)的多元数据处理方法, 它可以同时实现降维及两组特征变量间的相关性分析, 具有广泛的适用性. PLS方法目前已被证明能较好的解决高维样本问题及变量间的多重相关性, 且其算法复杂度要低于PCA算法, 因此本文使用PLS

方法对数据进行预处理. 此外, 仅依靠扰动训练集的Bagging重采样策略难以保证子分类器之间的差异性. 而通过对特征子空间进行随机选取, 使不同子分类器更加倾向于问题域的不同侧面, 能更加有效地增加分类器之间的差异性. 因此本文采用二重扰动机制对训练集进行扰动.

### 3.2 基于$k$均值聚类的修剪方法(Ensemble pruning based on $k$-means clustering)

第2阶段是子分类器的剪枝阶段. 文献[10]指出采用部分分类器进行集成可能会比使用全部分类器进行集成的效果更好. 此外, 使用全部分类器进行集成往往需要消耗大量的存储空间. 因此实际应用中, 在进行选择性集成前, 通常对分类器集合进行剪枝处理, 目的是在保持分类器多样性的同时, 删去一些冗余的分类器. 本节主要对基于$k$均值聚类的修剪方法进行介绍[24]). 设$C=\{c_1,c_2,\cdots,c_T\}$为已经训练好的子分类器集合, $S=\{(x_1,y_1),\cdots,(x_N,y_N)\}$为一组规模为$N$的验证样本集. 让子分类器$c_i$对全部验证样本进行识别, 并将其判决结果与验证样本的实际类别标签进行对比, 若子分类器$c_i$ 对样本$(x_j,y_j)$识别正确, 则输出结果$c_i(x_j,y_j)=1$, 反之$c_i(x_j,y_j)=0$. 因此, 分类器$c_i$对整个验证样本集的输出结果可以用向量$h_i=\{h_{i1},h_{i2},\cdots,h_{iN}\}$表示, 其中$h_{ij}\in\{0,1\}$. 若分类器$c_i$和$c_j$的输出结果分别为$h_i=\{h_{i1},h_{i2},\cdots,h_{iN}\}$和$h_j=\{h_{j1},h_{j2},\cdots,h_{jN}\}$. 当且仅当$h_{ik}$和$h_{ik}$同时为零时, 令$h_{ik}h_{ik}=1$, 其余情况$h_{ik}h_{ik}=0$, 则子分类器$c_i$和$c_j$间的重合错误率可定义如下[25]:

$$\mathrm{Prob}(c_i\mathrm{fails},c_j\mathrm{fails})=\frac{1}{N}\sum_{k=1}^{N}h_{ik}h_{jk},\tag{8}$$

则分类器$c_i$和$c_j$的之间的距离度量可定义如下:

$$D(c_i,c_j)=1-\mathrm{Prob}(c_i\mathrm{fails},c_j\mathrm{fails}).\tag{9}$$

本文采用$k$均值聚类法将分类器集合分类器. 本节主要对基于$k$均值聚类的修剪方法进行介绍. 设$C=\{c_1,c_2,\cdots,c_T\}$分成$k$类, 通过寻找到$k$个聚类中心点$\{M_j\}_{j=1}^{k}$令目标函数

$$J=\sum_{i=1}^{T}\min_j D(c_i,M_j)\tag{10}$$

达到最小, 同时输出$k$个聚类中心点作为新的子分类器集合$C'=\{c_1,c_2,\cdots,c_k\}$. 此外, 由于$k$均值聚类的聚类数目$k$需要预先给出, 因此本文通过逐步递增分类数目$k$直至目标函数值$J$升高的方法来获得最佳分类数$k$.

### 3.3 子分类器的动态选择与循环集成(Dynamic selection and circulating combination of individual classifier)

第3阶段是对分类器进行动态选择和循环集成. 大量研究表明只有当子分类器之间的错误不相关时, 集成学习才有意义[12]. 因此, 选择或者生成更具差异性的子分类器是选择性集成系统能取得成功的关键之处. 目前, 虽然大量文献对子分类器之间的差异性度量进行了研究, 但至今仍没有形成统一的标准. 这些差异性度量大致可以分为两类: 成对的(pairwise)和非成对的(non-pairwise). 成对的差异性度量只计算每一对子分类器之间的差异性, 而非成对的差异性度量则直接计算整个集成系统的差异性[17]. 假设已获得训练好的$k$个子分类器, $c_i$和$c_j$为其中两个不同的分类器; $N^{11}$($N^{00}$)为分类器和都分类正确(错误)的样本数目, $N^{10}$($N^{01}$)则为分类器$c_i$($c_j$)对其分类正确而分类$c_j$($c_i$)对其分类错误的样本数目, 因此总的测试样本数目为$N=N^{11}+N^{00}+N^{10}+N^{01}$.

不一致度量Dis主要反映了分类器$c_i$和$c_j$之间不一致度量$\mathrm{Dis}_{ij}$定义为[16]

$$\mathrm{Dis}_{ij}=\frac{N^{10}+N^{01}}{N},\tag{11}$$

$\mathrm{Dis}_{\mathrm{av}}$则为整个集成系统的不一致度量平均值, 本文在这里使用文献[26]定义的差异性度量$\kappa$, 其定义如下:

$$\kappa=1-\frac{1}{2\bar{p}(1-\bar{p})}\mathrm{Dis}_{\mathrm{av}},\tag{12}$$

其中$\bar{p}$为子分类器集合的平均分类精度, 有

$$\bar{p}=\frac{1}{Nk}\sum_{j=1}^{N}\sum_{i}^{k}c_i(x_j,y_j).\tag{13}$$

根据集成差异性度量$\bar{p}$的定义, 本文先给出一种分类器排序方法, 其核心思想是根据集成差异度$\bar{p}$, 按照集成前序选择法对$k$个子分类器进行排序. 按照排序算法对分类器完成排序之后, 就可以根据实际识别精度需求对子分类器进行动态选择与循环集成. 当少数分类器集成能满足识别需求时, 则无需再集成其他子分类器. 若达不到预期识别需求则按顺序添加其他分类器, 并进行循环集成. 其过程为: 首先设置初始识别精度阀值及其步长, 然后分类器序列中选择第一个分类器对测试集进行识别, 若识别结果满足识别要求则终止并输出识别结果. 否则按序列顺序依次添加其他子分类器进行集成直至满足识别需求. 若集成了所有子分类器仍不能满足识别需求, 则按照设置的步长降低识别精度阀值, 重复以上过程, 直至满足识别需求并输出识别结果. 分类器的动态选择与循环集成方法步骤如下所示:

**步骤 1**　输入分类器序列$P$, 测试样本集$\mathrm{Test}=\{(x_1,y_1),\cdots,(x_m,y_m)\}$, 阀值步长$\Delta\lambda$, 初始识别精度阀值$\lambda$, 令计数器$i=1$.

**步骤 2**　从分类器序列$P$中选择前$i$个分类器对测试集Test进行测试, 并得到识别结果$R$.

**步骤 3**　判断, 若识别结果$R\geqslant\lambda$, 则中断算法并输出识别结果$R$; 否则跳转至步骤4.

**步骤 4** 判断, 若$i < k$则$i++$并跳转至步骤2, 若$i \geqslant k$则跳转至步骤5.

**步骤 5** 令识别精度阈值$\lambda = \lambda - \Delta\lambda, i = 1$, 并跳转至步骤2.

显然, 若初始识别精度阈值$\lambda$越大, 阈值步长$\Delta\lambda$越小, 则集成系统的整体识别效果较好, 但算法的循环次数也会因此增加, 识别所需时间也相对较长. 因此在应用中, 应根据识别目标的复杂程度及实际识别需求来对初始精度阈值及步长取值, 以兼顾识别精度及时效性.

## 4 实验结果及分析(Experimental results and analysis)

为了验证算法的有效性, 将本文算法与基于Bagging的直觉模糊核匹配追踪集成算法(Bagging-IFKMP)[27]、基于遗传选择的IFKMP集成算法(GASEN–IFKMP)[10] 以及基于前向顺序选择的直觉模糊核匹配追踪集成算法(SFS–IFKMP)[28]进行了对比. 其中, Bagging方法是传统的集成方法, 即直接对所有子分类器的识别结果进行多数投票组合, GASEN和SFS方法则是两种相对典型的选择性集成算法. 实验过程中选取高斯核作为基分类器的核函数. 其中, GASEN采用二进制编码, 为了对算法性能进行验证, 本文选择不同的样本集合进行试验. 为了避免随机误差, 每次试验分别进行20次蒙特卡洛仿真. 仿真环境: 操作系统 Window XP, 编译软件 MATLAB7.6, Pentium(R) Dual-Core CPU E5500 @ 2.8 GHz, 内存 2 GB.

### 4.1 Shuttle数据识别(Test on Shuttle data)

实验首先选取UCI数据集中的Shuttle数据S集进行验证. Shuttle是一个7类数据集, 包含9维特征, 共有43500个训练样本及14500个测试样本. 其中大约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, 本文将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). 将训练样本中的前40000个样本做为训练集, 用于子分类器的生成, 将剩余的3500个样本作为验证集, 用于对子分类器进行修剪及排序. 实际应用中, 更多的是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重要样本(即异常样本)的检测概率, 因此从14500个测试样本中随机选取其中1000个负类样本作为测试集进行测试.

参数设置: 直觉模糊参数$\omega(y_i)$根据文献[24]方法进行选取, 得: 1) 正类样本为非指定类别样本$y_1$的直觉模糊参数$\omega(y_1) = 0.9$; 2) 负类样本为指定类别样本$y_2$的直觉模糊参数$\omega(y_2) = 1.1$; 核函数通过交叉验证得到$\sigma = 1$; 令子训练集样本规模$L = 2000$; 初始集成规模$T = 100$, 初始识别精度为0.99, 阈值步长$\Delta\lambda = 0.005$. 采用GASEN选择时, 令交叉概率为0.7, 变异概率为0.01. 实验结果见表1, 其结果为20次蒙特卡洛仿真的均值. 在这里需要说明的是, 由于Bagging方法为直接集成法, 无需对子分类器进行搜索选择, 其初始集成规模即为最终集成规模.

表 1 Shuttle数据集识别实验

Table 1 Testing of Shuttle data

| 算法 | 训练集 | 验证集 | 测试集 | 集成规模 | 搜索时间/s | 识别率/% | 偏差/% |
|---|---|---|---|---|---|---|---|
| Bagging–IFKMP | 40000 | 3500 | 负类: 1000 | 100 | — | 97.32 | 0.61 |
| SFS–IFKMP | 40000 | 3500 | 负类: 1000 | 26.2 | 1520.7 | 97.65 | 0.16 |
| GASEN–IFKMP | 40000 | 3500 | 负类: 1000 | 19.6 | 2910.2 | 97.88 | 0.23 |
| 本文方法 | 40000 | 3500 | 负类: 1000 | **11.3** | **343.4** | **98.63** | **0.12** |

表1中, 从集成规模及子集搜索时间来看, 本文方法集成所需的子分类器数目和子集搜索时间相对其他方法明显减少, SFS–IFKMP 方法及 GASEN–IFKMP方法相对直接集成方法所需的子分类器数目分别下降了73.8%和80.4%, 而本文方法则下降了88.7%, 同时本文方法所需的子集搜索时间也相对SFS–IFKMP方法及GASEN–IFKMP方法分别下降了77.3%和88.2%. 这是由于本文方法先采用$k$均值聚类对大量的冗余分类器进行了剔除, 因此在后期集成过程中所需的子分类器数目及子集搜索时间均相对没有采用剪枝的集成选择策略明显下降. 同时还注意到3种选择集成方法的搜索时间均较长, 这是因为搜索过程中各子分类器还需要对验证数据集进行识别, 而本例中选取的验证集规模较大, 导致其子集搜索时间较长. 综合实验结果可得, 本文方法不仅具有更好的识别率, 同时也具备更好的泛化性能, 也是因为其他集成选择策略往往是针对某一验证集选定子分类器, 且选定后就不再改变. 而测试集则往往不同于验证数据集, 这使得其他一些有可能对测试集识别效果更好的子分类器无法入

选, 致使整个集成系统缺乏足够的灵活性, 同时也影响了识别效果及泛化性能. 而本文采取的分类器动态选择和循环集成策略则有效地避免了上述问题, 该策略能够根据测试集的情况动态选择出所需子分类器并进行循环集成, 不仅使系统结构更为灵活, 也大大增加集成系统的泛化能力和效率.

### 4.2 人工数据集识别(Test on artificial data)

模式识别领域中, 双螺旋曲线的分类问题一直是公认的有相当难度的问题, 因此它也被经常用作检测识别算法分类性能的试金石.

随机分别产生两类双螺旋曲线的样本共10000个, 样本分布如图2所示. 从样本集中的随机选取9000个样本做为训练集, 用于子分类器的生成, 将剩余的1000个样本作为验证集, 用于对子分类器进行修剪及排序. 为了检验算法对重要目标的识别效果, 本文将正类样本(即图2中的黑色样本)标记为重要样本类别, 并从5000个正类样本中随机抽取1000个样本作为测试集进行测试.

参数设置: 令正类样本的直觉模糊参数$\omega(y_1)=1.2$, 负类样本的直觉模糊参数$\omega(y_2)=0.9$; 核函数通过交叉验证得到$\sigma=8$; 令子训练集样本规模$L=500$; 初始集成规模$T=60$, 初始识别精度为0.99, 阀值步长$\Delta\lambda=0.005$. 采用GASEN选择时, 令交叉概率为0.7, 变异概率为0.02. 实验训练前先对训练数据进行加噪处理, 随机改变20%样本的类别属性, 然后进行训练, 实验结果如表2所示.

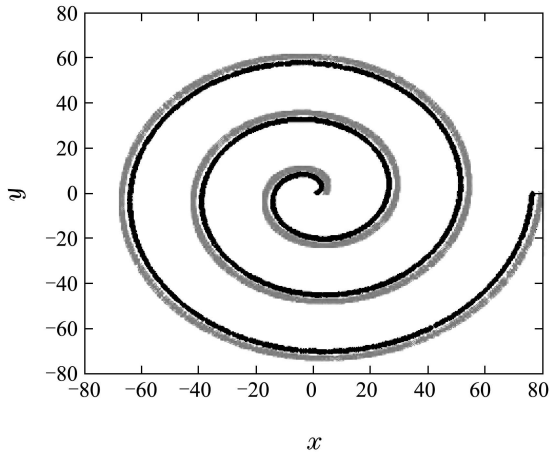


图 2 样本分布图

Fig. 2 The figure of sample distribution

表 2 人工数据集识别实验

Table 2 Testing of artificial data

| 算法 | 训练集 | 验证集 | 测试集 | 集成规模 | 搜索时间/s | 识别率/% | 偏差/% |
|---|---|---|---|---|---|---|---|
| Bagging-IFKMP | 10000 | 1000 | 正类: 1000 | 100 | — | 97.26 | 1.33 |
| SFS–IFKMP | 10000 | 1000 | 正类: 1000 | 16.3 | 791.9 | 98.33 | 0.31 |
| GASEN–IFKMP | 10000 | 1000 | 正类: 1000 | 26.6 | 1115.9 | 98.65 | 0.26 |
| 本文方法 | 10000 | 1000 | 正类: 1000 | **8.4** | **163.5** | **99.23** | **0.06** |

表2的实验结果表明, 在训练样本含噪声的情况下, 本文方法也能对指定的重要样本仍能保持较高的识别精度, 且其识别效率、识别性能及泛化能力均明显优于传统的集成方法和其他两种集成选择策略, 这也与Shuttle数据实验结果相吻合.

### 4.3 弹道目标识别(Test on ballistic target)

未来防空反导作战中, 弹道导弹的威胁程度明显要比其他来袭目标威胁程度更大, 因此本实验将这弹道目标设为重要识别目标类别, 以此来检验本文算法应用于弹道目标识别领域的效果. 本文采用FEKO电磁仿真软件生成弹道目标的雷达截面积(rader cross section, RCS)数据进行仿真实验, 并选取目标RCS序列的极大值、极小值、均值及方差作为一个样本的类别特征指标. 图3为FKEO计算的弹道目标全姿态RCS数据, 图4所示为采用插值法获得的弹道目标在30 s内的动态RCS序列.

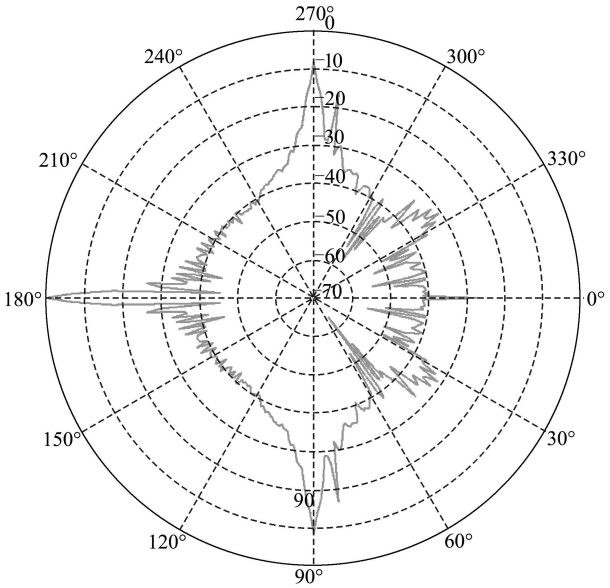


图 3 RCS 随方位角变化

Fig. 3 RCS via azimuth

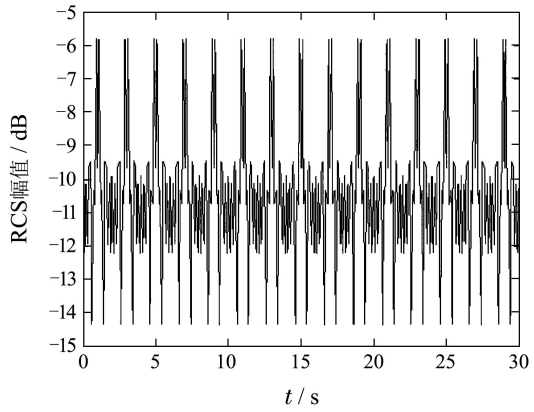


图 4 弹道目标RCS序列图

Fig. 4 The RCS series of ballistic target

除了弹道目标外, 选取空中诱饵、碎片及各式飞行器作为干扰目标类(负类目标), 生成两类样本共1000个. 从样本集中的随机选取900个样本做为训练集, 用于子分类器的生成, 并将剩余的100个样本作为验证集, 用于对子分类器进行修剪及排序. 为了对算法进行检验, 从500个正类样本中随机抽取200个样本作为测试集进行测试.

参数设置: 令正类样本的直觉模糊参数$\omega(y_1) = 1.4$, 负类样本的直觉模糊参数$\omega(y_2) = 0.9$; 核函数通过交叉验证得到$\sigma = 1.6$; 令子训练集样本规模$L = 500$; 初始集成规模$T = 60$, 初始识别精度为0.9, 阀值步长$\Delta\lambda = 0.005$. 采用GASEN选择时, 令交叉概率为0.85, 变异概率为0.01. 实验结果如表3所示.

表 3 弹道导弹目标识别实验

Table 1 Testing of ballistic target

| 算法 | 训练集 | 验证集 | 测试集 | 集成规模 | 搜索时间/s | 识别率/% | 偏差/% |
|---|---|---|---|---|---|---|---|
| Bagging-IFKMP | 900 | 100 | 正类: 200 | 100 | — | 82.36 | 2.23 |
| SFS–IFKMP | 900 | 100 | 正类: 200 | 18.9 | 10.2 | 83.65 | 0.83 |
| GASEN–IFKMP | 900 | 100 | 正类: 200 | 29.6 | 28.6 | 83.85 | 1.02 |
| 本文方法 | 900 | 100 | 正类: 200 | 5.3 | 5.3 | 84.42 | 0.37 |

表3的实验结果表明, 针对复杂的弹道目标识别问题, 采用混合选择策略的集成方法后, 相比较其他3种方法, 识别的精确度得到了提升, 且经过剪枝后集成系统所包含的候选IFKMP学习机数量明显减少, 这意味后期采用动态选择和循环集成策略时, 对测试样本的识别速度将有较大的提高. 此外, 还根据实际需求对初始精度阀值及步长进行调节, 以实现效率与精度的平衡. 因此, 对于需要兼顾识别率及时效性的弹道目标识别领域, 本文方法不失为一种较好的选择.

## 5 结论(Conclusions)

为了从子分类器集合中选择一组差异性大的子分类器, 从而改进集成学习系统的性能, 本文提出了一种基于混合选择策略的直觉模糊核匹配追踪集成算法. 该方法在首先采用训练集和特征空间双重扰动的方式来生成子分类器集合; 然后采用$k$均值聚类进行修剪, 删除冗余的子分类器; 最后采用动态选择和循环集成策略候选子分类器进行二次选择, 使参与集成的分类器的个数不仅能够随识别问题的复杂程度而自适应变化, 而且还可以根据识别精度的要求进行循环集成, 从而实现识别精度和识别效率的折衷. 实验结果表明, 与传统方法相比, 本文方法不仅具有更好的识别效果和泛化能力, 同时系统结钩也更为灵活, 效率更高.

## 参考文献(References):

[1] PASCAL V, BENGIO Y. Kernel matching pursuit [J]. *Machine Learning*, 2002, 48(1/2/3): 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] FU Lihua, LI Hongwei, ZHANG Meng. Fast orthogonal kernel matching pursuit based on greedier strategy [J]. *Acta Electronica Sinica*, 2013, 41(8): 1580 – 1585.
(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.)

[5] LEI Yang, KONG Weiwei, LEI Yingjie. Technique for target recognition based on intuitionistic fuzzy c-means clustering and kernel matching pursuit [J]. *Journal on Communications*, 2012, 33(1): 136 – 143.
(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(1): 136 – 143.)

[6] LI Qing, JIAO Licheng, ZHOU Weida. Pattern recognition based on the fuzzy kernel matching pursuit [J]. *Chinese Journal of Computers*, 2009, 32(8): 1687 – 1694.
(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.)

[7] LEI Yang, LEI Yingjie, ZHOU Chuangming. Techniques for target recognition based on intuitionistic fuzzy kernel matching pursuit [J].

*Acta Electronica Sinica*, 2011, 39(6): 1441 – 1446.
(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.)

[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] ZHOU Z H, WU J X, TANG W. Ensembling neural networks: many could be better than all [J]. *Artificial Intelligence*, 2002, 137 (1/2): 239 – 263.

[11] RAFAL L, MAREK K, TOMASZ W. Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers [J]. *Neurocomputing*, 2014, 126(Complete): 29 – 35.

[12] ZHANG Chunxia, ZHANG Jiangshe. A survey of selective ensemble learning algorithms [J]. *Chinese Journal of Computers*, 2011, 34(8): 1399 – 1410.
(张春霞, 张讲社. 选择性集成学习算法综述 [J]. 计算机学报, 2011, 34(8): 1399 – 1410.)

[13] ELAHEH R, ABDOLREZA M. A hierarchical clusterer ensemble method based on boosting theory [J]. *Knowledge-Based Systems*, 2013, 45(Complete): 83 – 93.

[14] ROBERT B, RICARDO G O, FRANCIS Q. Attribute bagging: improving accuracy of classier ensembles by using random feature subsets [J]. *Pattern Recognition*, 2003, 36(6): 1291 – 1302.

[15] HAO Hongwei, WANG Zhibin, YIN Xucheng, et al. Dynamic selection and circulating combination for multiple classifier systems [J]. *Acta Automatica Sinica*, 2011, 37(11): 1291 – 2295.
(郝红卫, 王志彬, 殷绪成, 等. 分类器的动态选择与循环集成方法 [J]. 自动化学报, 2011, 37(11): 1290 – 1295.)

[16] YANG Chun, YIN Xucheng, HAO Hongwei. Classifier ensemble with diversity: effectiveness analysis and ensemble optimization [J]. *Acta Automatica Sinica*, 2014, 40(4): 660 – 674.
(杨春, 殷绪成, 郝红卫. 基于差异性的分类器集成: 有效性分析及优化集成 [J]. 自动化学报, 2014, 40(4): 660 – 674.)

[17] SUN Bo, WANG Jiandong, CHEN Haiyan, et al. Diversity measures in ensemble learning [J]. *Control and Decision*, 2014, 29(3): 387 – 395.
(孙博, 王建东, 陈海燕, 等. 集成学习中的多样性度量 [J]. 控制与决策, 2014, 29(3): 387 – 395.)

[18] BARTOSZ K, MICHAL W, BOGUSLAW C. Clustering-based ensembles for one-class classification [J]. *Information Sciences*, 2014, 264(Complete): 182 – 195.

[19] MONTHER A, WANG D H. Fast decorrelated neural network ensembles with random weights [J]. *Information Sciences*, 2014, 264(Complete): 104 – 117.

[20] CHEN L, WEN Q C, CHENG Q, et al. LibD3C: ensemble classifiers with a clustering and dynamic selection strategy [J]. *Neurocomputing*, 2014, 123(Complete): 424 – 435.

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

[22] JIAO L C, LI Q. Kernel matching pursuit classifier ensemble [J]. *Pattern Recognition*, 2006, 39(4): 587 – 584.

[23] GIACINTO G, FABIO R. An approach to the automatic design of multiple classifier system [J]. *Pattern Recognition Letters*, 2001, 22(1): 25 – 33.

[24] LI Chunsheng, WANG Yaonan. New initialization method for cluster center [J]. *Control Theory & Applications*, 2010, 27(10): 1435 – 1440.
(李春生, 王耀南. 聚类中心初始化的新方法 [J]. 控制理论与应用, 2010, 27(10): 1435 – 1440.)

[25] ZHOU Z H, TANG W. Clusterer ensemble [J]. *Knowledge-Based Systems*, 2006, 19(1): 77 – 83.

[26] KUNCHEVA L I, WHITAKER C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy [J]. *Machine Learning*, 2003, 51(2): 181 – 207.

[27] BREIMAN L. Bagging predictors [J]. *Machine Learning*, 1996, 24(2): 123 – 140.

[28] FAN W, CHU F, WANG H, et al. Pruning and dynamic scheduling of cost-sensitive ensembles [C] //*Proceedings of the 19th National Conference on Artificial Intelligence*. Menlo Park: American Association for Artificial Intelligence, 2002: 146 – 151.

## 作者简介:

**雷英杰**　(1956−), 男, 教授, 博士生导师, 目前研究方向为智能信息处理与智能决策, E-mail: leiyjie@163.com;

**余晓东**　(1989−), 男, 博士研究生, 目前研究方向为模式识别与智能信息处理等, E-mail: agosoa@163.com;

**王　睿**　(1964−), 女, 副教授, 硕士生导师, 目前研究方向为智能信息处理与多传感器信息融合, E-mail: wangrui@163.com;

**王　毅**　(1979−), 男, 讲师, 目前研究方向为智能信息处理, E-mail: wangyi.kgd@gmail.com.