# 一种基于遗传算法的 SVM 决策树多分类方法

王 一[1,2] 杨俊安[1,2] 刘 辉[1,2]
（1. 电子工程学院 合肥 230037；2. 安徽省电子制约技术重点实验室 合肥 230037）

**摘 要**：在当前的机器学习领域，如何利用支持向量机（SVM）对多类目标进行分类，同时提高分类器的分类效率已经成为研究的热点之一，有效地解决此问题对于提高目标的识别概率具有较大意义。本文针对 SVM 多分类问题提出了一种基于遗传算法的 SVM 最优决策树生成算法。算法以随机生成的决策树构建的 SVM 分类器对同一测试样本的分类正确率作为遗传算法的适应度函数，通过遗传算法寻找到最优决策树，再以最优决策树构建 SVM 分类器，最终实现 SVM 的多分类。将该算法应用于低空飞行声目标识别问题，实验结果表明，新方法比传统的 1-a-1、1-a-r、SVM-DL 和 GADT-SVM 方法有更高的分类精度和更短的分类时间。

**关键词**：支持向量机；遗传算法；决策树
**中图分类号**：TN959·1 文献标识码：A 文章编号：1003-0530(2010)10-1495-05

# A GA-based SVM Decision-tree Multi-Classification method

WANG Yi[1,2] YANG Jun-an[1,2] LIU Hui[1,2]
（1. Electronic Engineering Institute, Hefei, 230037; 2. Key Laboratory of Electronic Restriction, Anhui Province Hefei, 230037）

**Abstract**: Recently, in the fields of machine learning, how to use support vector machine for multi-class objects classification while improving the classification efficiency of the classifier has become one of the main study points, effective solutions to this problem have great significance for improving the probability of target recognition. In this paper we present a GA-based SVM decision tree algorithm. In our algorithm, we randomly generate a decision tree to build the SVM classifier on the same test samples of the classification accuracy rate as the genetic algorithm fitness function, then with the help of genetic algorithm, we can find the optimal decision tree, and then construct an optimal decision tree SVM classifier as the optimal SVM classifier. We use this algorithm to deal with the low altitude flying passive acoustic target identify problem. Experiment results show that the proposed method is more precise and less testing time cost than the traditional 1-a-1, 1-a-r, SVM-DL, GADT-SVM methods.

**Key words**: support vector machine; genetic algorithm; decision tree

## 1 引言

战场声目标识别技术相对于传统的雷达和光电探测技术而言具有抗干扰性强、隐蔽性好、成本低、功耗小等特点，而且声波相对于雷达波、光波，具有不受视线和能见度限制，能够探测到障碍物后方目标的特点。因此作为国土防空体系良好补充的低空声目标实时识别系统将在未来信息化战场中发挥越来越重要的作用[1]。

战场声目标识别技术突破的关键在于特征的提取和分类器设计，而对于分类器设计部分而言，支持向量机（SVM）由于其在解决小样本、非线性及高维模式问题中所表现出来的独特优越性，已在语音识别、文本分类等方面得到了成功应用[2]，并在战场声目标识别技术中得到了初步应用[3]。但是支持向量机总体来说是一个二分类器[4]，如何利用支持向量机对多类目标进行分类，同时提高分类器的精度和运算速度已成为当前机器学习领域研究的热点之一，此问题的有效解决对于提高低空声目标的识别概率具有较大意义。在现有的 SVM 多分类方法中，经典的一对一分类（one-a-against-one）、一对多分类（one-against-rest）、有向非循环图支持向量机（DAG-SVM）、决策树支持向量机（DT-SVM），由于存在需要训练的支持向量机个数过多、测试的时间过长等缺点，导致这些方法识别的精度不高，所耗时间过长[5]，近年来国内外研究人员提出的 SVM-DL 技术[6]、GADT-SVM[7]，虽然对此做出了一定改进，但也存在着识别精度降低以及产生的 SVM 分类器不是最优等问题，导致达不到最佳分类效果。

本文针对上述问题，提出了一种基于遗传算法的 SVM 决策树多分类方法：首先利用随机二分的方式产生决策树，然后以该决策树构建的 SVM 分类器对同一测试

收稿日期：2009 年 11 月 30 日；修回日期：2010 年 5 月 6 日
基金项目：国家自然科学基金（No. 60872113）

样本的分类正确率作为遗传算法的适应度函数,同时采用精英保留策略,经过若干代遗传操作后最终得到最优决策树,从而构建出最优的 SVM 分类器。对战场声目标实际数据的测试结果表明了该方法的有效性。

## 2 Mel 倒谱系数特征

本文采用 Mel 倒谱系数特征作为输入,送入 SVM 分类器进行分类识别。Mel 倒谱系数(简称 MFCC)是将人耳的听觉感知特性和语音的产生机制相结合,已经在低空声目标识别系统中得到了广泛应用[3]。

Mel 倒谱系数设计是基于以 Mel 频率尺度为线性标度的频率映射,与线性频率的转换公式如下,其中 f 为线性频率:

$$Mel(f)=\left(\frac{1000}{\log_{10}(1+1000/700)}\right)\log_{10}(1+f/700)$$
$$=2595\log_{10}(1+f/700) \quad (1)$$

从三种典型战场声目标信号提取一段得到 MFCC 参数如图 1 所示,其中立方体图为 25 帧,每帧 24 维的 MFCC。对应的平面图为其中一帧的 MFCC,其中横轴代表为特征维数。

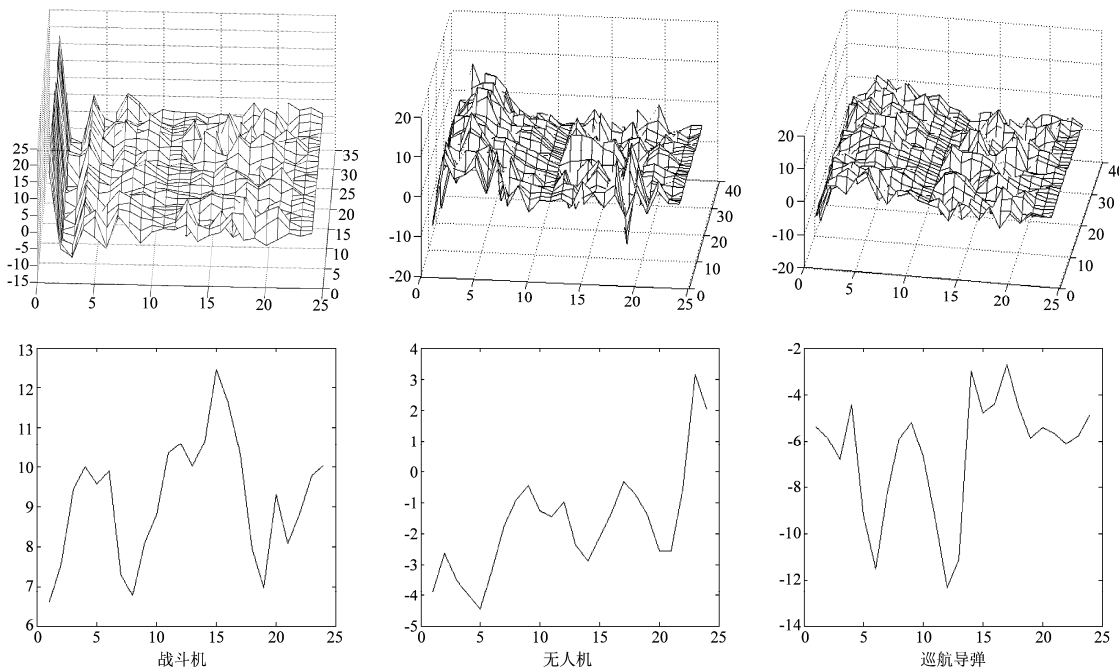



图 1 三种典型低空声目标的 MFCC 参数

从图中可以看到不同目标的 MFCC 参数明显不同,因此用此参数可以区别不同声目标。

## 3 支持向量机(SVM)

SVM 的主要思想是将在低维空间线性不可分的样本通过核函数(非线性映射)映射到高维特征空间,在高维空间构造最优分类超平面。最优分类超平面就是不但能将两类样本正确划分,并且使每一类样本与超平面距离最近的点与分类线之间距离最大,即分类间距(margin)最大。如图 2 为最优分类面,其中 ▲ 形和 ● 形代表两类样本,$H$ 为分类面,$H_1$ 和 $H_2$ 分别为两类中离分类面最近的样本与分类面的平行线,它们之间的距离就是分类间隔,$H_1$、$H_2$ 上的训练样本点称作支持向量(Support Vector)。

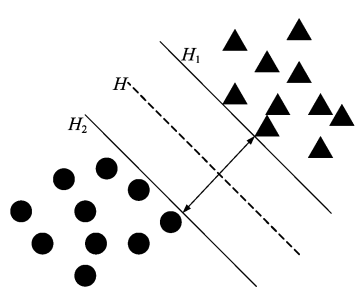



图 2 最优分类面

用数学式表示,求解最优分类面问题转化为:

$$\min \frac{1}{2}\|w\|^2+C\left[\sum_{i=1}^{n}\zeta_i\right]$$
$$s.t\ y_i[(w\cdot x_i)+b]\geq 1-\zeta_i,\quad i=1,\cdots,n \qquad \zeta_i\geq 0 \quad (2)$$

式中 $\zeta_i$ 为非负松弛变量,$C$ 为惩罚因子,它控制对错分样本惩罚的程度。通过拉格朗日乘子法求解上述优化问题,最后可得优化问题的对偶形式为:

$$\max \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i x_j)$$
$$s.t\ \sum_{i=1}^{n}\alpha_i y_i=0 \qquad 0\leq\alpha_i\leq C, i=1,..n \quad (3)$$

最后得到判别函数:

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b\} \quad (4)$$

## 4 基于遗传算法的 SVM 决策树多分类方法

本文利用遗传算法寻找最优分类决策树,从而构

建 SVM 分类器。以由决策树构建的 SVM 分类器对同一测试样本的分类正确率作为适应度函数,并采用精英保留策略经过若干代遗传操作后得到最优决策树,使得构成的 SVM 能够得到最大的分类正确率。

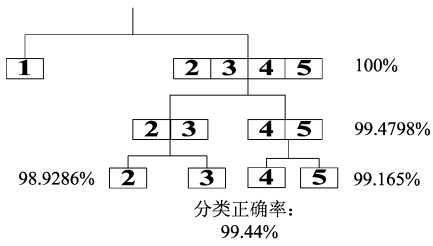
### 4.1　算法设计

#### 4.1.1　编码策略

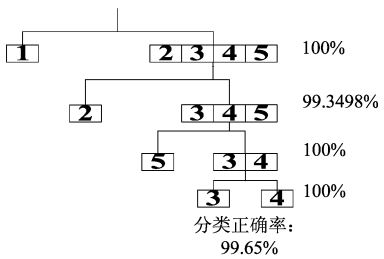
采用实值编码的策略来实现原始训练样本集的编码,决策树根节点的染色体编码为:{1,2,…,K},其中 K 为原始训练样本集的类别总数,染色体中每个基因对应的是原始训练样本集的类别编号:对于根节点以下的子节点,根据其父节点的划分结果,剔除父节点染色体中本节点不包含的类别对应的基因,形成新的染色体。

#### 4.1.2　适应度函数的确定

若采用文献[7]所述的适应度函数,即以 SVM 分类算法的分类间隔作为 GA 适应度函数,虽然计算速度较快,但是仅考虑了每个内部节点(Internal Node)下的最优分类方式,忽略了全局信息,即局部最优的叠加并不等同于全局最优,很可能会陷入局部最优,如图3(a)所示,应用文献[7]所述的方式寻找到的决策树虽然在每一层都是最优的,但是由于叶节点(Leaf)分类能力较差,影响了整体性能,正确率反而不如图 3(b)所示的决策树。




(a) 运用文献[7]方法生成的最优决策树




(b) 运用本文方法生成的最优决策树

图 3　两种适应度函数产生的最优决策树

为此,我们从考虑整体的角度出发,以生成的决策树构建的SVM对同一测试样本的预测正确率作为适应度函数fit($a_i$)($i=1,2,\cdots k$)(其中个体 $a_i$ 为随机生成的决策树)。

#### 4.1.3　遗传操作

本文中选择的遗传算子包括选择、交叉和变异算子。

(1)选择算子

采用轮盘赌的选择方式,若个体 $a_i$ 的适应度函数为 fit($a_i$),种群规模为 popsize,则选中 $a_i$ 为下一代个体的概率为:

$$p(a_i)=\mathrm{fit}(a_i)/\sum_{j=1}^{\mathrm{popsize}}\mathrm{fit}(a_j) \qquad (5)$$

显然适应度高的个体,繁殖的下一代数目较多;而适应度小的个体,繁殖的数目少,甚至被淘汰。

(2)交叉算子

交叉策略与问题的编码方式是密不可分的,对于本问题若采取简单的一点或多点交叉策略,必然会导致染色体基因出现重复,为此,本文采取了一种类似顺序交叉法(OX)的交叉策略:(a)随机在串中选择一个交配区域,如图 4 中 $A$、$B$ 串中用竖线划分的区域;(b)将 $B$ 的交配区域加到 $A$ 的前面,$A$ 的交配区域加到 $B$ 的前面;(c)在 $A'$、$B'$ 中自交配区域后依次删除与交配区相同的类别号,得到交叉后的两子串。交叉策略示意图如图 4 所示

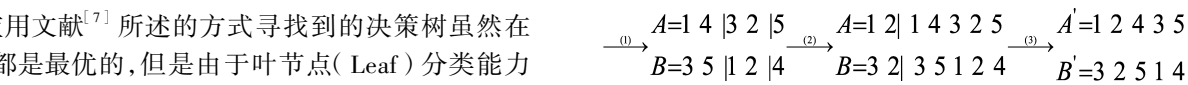



图 4　交叉策略示意图

(3)变异算子

为了在遗传操作初期取得较大的变异算子以维持种群的多样性,防止出现早熟现象;在算法已接近最优解邻域时,减小变异算子,确保其局部搜索能力,本文采用自适应变异概率:

$$p_m=\exp(-1.5\times 0.5t)/\mathrm{popsize}\times\sqrt{L} \qquad (6)$$

其中:t 是进化代数,L 是染色体长度。

为了防止同一条染色体中相同基因重复出现,采用对换变异的思想,即染色体中某一基因座上的基因发生变异时,变异后的基因编码对应的基因座上的基因相应地变换为变异基因座上的原基因编码。假设串 $A$ 包含 5 个基因,$A=1\ 2\ 3\ 4\ 5$,若第二基因座上的基因发生变异,变异后为 4,则得到的新染色体为:$A'=1\ 4\ 3\ 2\ 5$。

### 4.2　算法流程

步骤 1:将全部训练样本集所属类别按实值编码策略进行编码。

步骤 2:从编码好的样本集中随机产生 K 个按照不同顺序排列样本作为初始样本,并进行随机二分。

步骤 3:判断各子节点是否只包含一类样本,若还包含一类以上的样本则转步骤 2 继续进行二分,直到生成决策树。

步骤 4:以生成的决策树构建 SVM,以对同一测试样本的预测正确率为适应度函数fit($a_i$)($i=1,2,\cdots K$),并按照适应度高低排序,并将产生的一个最佳决策树作为精英保存。

步骤 5:对初始种群进行选择、交叉、变异操作,产生新种群。

步骤 6:以最佳决策树是否 N 代未变以及是否达到最大进化代数 $T_{MAX}$ 作为终止条件,若满足,则算法停止,生成最优生成树;若不满足,转步骤 2 继续。

## 5 实验结果与分析

实验中采用实测噪声环境下五种飞行器 A、B、C、D、E 的声信号( 采样频率分别为 48kHz、48kHz、10kHz、50kHz、10kHz )作为训练、识别数据。声音信号分帧时每 256 个点分成一帧,帧移为 80 个点;选取 24 个数字滤波器组用于 24 阶的 MFCC 参数提取。我们选择目前应用最广的径向基核函数作为 SVM 的核函数,在径向基核函数参数 $\sigma^2$ 和惩罚因子 $C$ 的挑选上,本文采用三种方法寻找最优参数,分别是交叉验证方法、遗传算法和粒子群算法。通过将这三种方法找到的最优 $\sigma^2$ 和 $C$ 作为 SVM 核函数的参数对不同样本按照不同组合方式进行分类比较,最终确定当 $\sigma^2$ 取 0.00012207,$C$ 取 0.5 时在绝大多数情况下都能得到最优解,因此选择其作为径向基核函数的参数 $\sigma^2$ 和惩罚因子 $C$。我们取 300、500、700、900、1100、1300、1427 个样本对不同的 SVM 多分类方法进行测试,测试时间均取重复 20 次实验的平均时间。用本文提出的方法产生的最优决策树如图 5 所示。

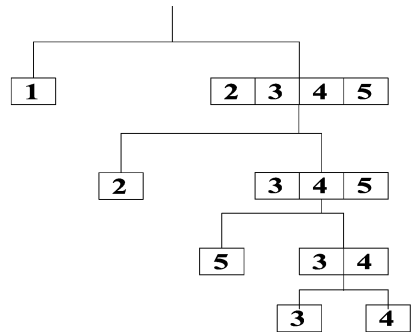



图 5 本文算法产生的最优决策树

各 SVM 多分类方法的分类精度见表 1:

表 1 各 SVM 多分类方法对于不同样本的分类精度

| | 300 | 500 | 700 | 900 | 1100 | 1300 | 1427 |
|---|---|---|---|---|---|---|---|
| 1-a-1 | 0.9967 | 0.9840 | 0.9957 | 0.9889 | 0.9891 | 0.9892 | 0.9967 |
| 1-a-r | 1 | 0.9740 | 0.9886 | 0.9833 | 0.9855 | 0.9877 | 0.9881 |
| SVM-DL | 1 | 0.9940 | 0.9986 | 0.9944 | 0.9955 | 0.9954 | 0.9958 |
| GADT-SVM | 0.9967 | 0.9880 | 0.9986 | 0.9922 | 0.9936 | 0.9938 | 0.9944 |
| 本文算法 | 1 | 0.9920 | 0.9986 | 0.9944 | 0.9967 | 0.9962 | 0.9965 |

各 SVM 多分类方法的分类时间( 秒 )见表 2:

表 2 各 SVM 多分类方法对于不同样本的平均分类时间

| | 300 | 500 | 700 | 900 | 1100 | 1300 | 1427 |
|---|---|---|---|---|---|---|---|
| 1-a-1 | 22.0839 | 37.9135 | 51.4615 | 66.8775 | 81.2385 | 96.3046 | 104.7944 |
| 1-a-r | 16.5536 | 31.5754 | 35.0895 | 55.8424 | 61.1586 | 80.4293 | 86.2418 |
| SVM-DL | 8.6950 | 15.7572 | 17.6132 | 27.6585 | 33.6683 | 40.0233 | 43.7721 |
| GADT-SVM | 10.6403 | 18.5672 | 17.4045 | 33.2303 | 40.0182 | 47.7589 | 52.0662 |
| 本文算法 | 8.0720 | 14.2135 | 16.2112 | 25.3593 | 30.8078 | 36.6488 | 40.0517 |

5 种方法在不同检测样本情况下的分类精度、平均分类时间比较如图 6 所示:

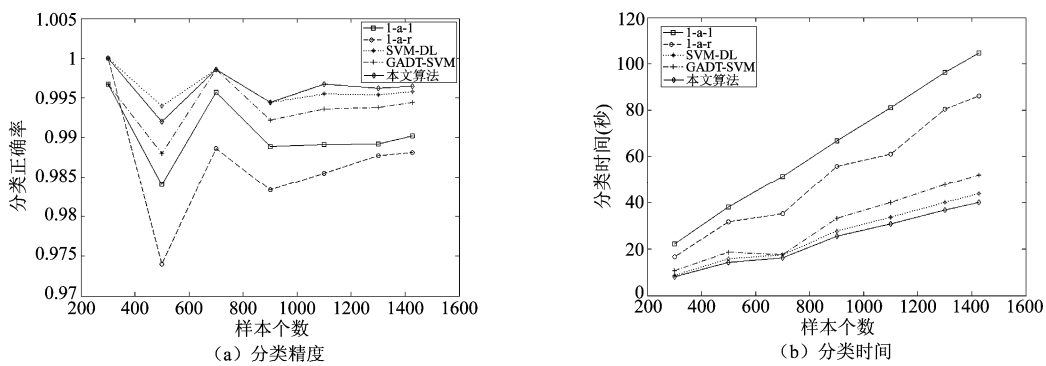



图 6 各 SVM 多分类方法分类精度、分类时间比较图

由于本文算法以全局最优分类为目标,克服了以往方法容易陷入局部最优的缺陷,因此分类精度较高,从图 6(a)中可看出本文算法在样本个数不同的情况下,分类精度较经典的 1-a-1,1-a-r 方法有了较大提高。并且随着样本个数增加,分类精度要好于 SVM-DL 和 GADT-SVM。

由于 GADT-SVM 与本文算法均采用了决策树的方法构建多级分类器,因此算法的运行时间是由构建最优决策树时间以及对样本识别时间两方面决定的。GADT-SVM 和本文算法构建针对不同样本最优决策树的平均时间(秒)见表 3:

表 3　GADT-SVM、本文算法构建针对不同样本最优决策树的平均时间

| | 300 | 500 | 700 | 900 | 1100 | 1300 | 1427 |
|---|---|---|---|---|---|---|---|
| GADT-SVM | 57.7405 | 61.8721 | 65.8674 | 70.1418 | 74.1289 | 77.3484 | 78.9458 |
| 本文算法 | 63.3567 | 67.9486 | 72.1254 | 76.0689 | 80.1890 | 82.9472 | 84.7567 |

将构建最优决策树时间与表 2 的分类时间相加, GADT-SVM 与本文算法的平均运行时间(秒)见下表 4:

表 4　GADT-SVM、本文算法平均运行时间

| | 300 | 500 | 700 | 900 | 1100 | 1300 | 1427 |
|---|---|---|---|---|---|---|---|
| GADT-SVM | 68.3808 | 80.4393 | 83.2719 | 103.3721 | 114.1471 | 125.1073 | 131.012 |
| 本文算法 | 71.4287 | 82.1621 | 88.3366 | 101.4282 | 110.9968 | 119.596 | 124.8084 |

从表 3 可以看出,本文算法在构建最优决策树时由于保留了较多的全局信息,因此较 GADT-SVM 耗费了更多的时间。但算法的运行时间是由决策树构建时间和分类时间共同决定的,正如表 4 所示,当将这两个时间一起加以考虑后本文算法能够取得比 GADT-SVM 更好的效率。另外,由于本文研究的低空飞行目标声识别系统在大多数情况下都可以通过先验信息提前构建出最优决策树,因此在用于目标分类时只需考虑分类时间,而不用将最优决策树构建时间计算在内,正如图 6(b)所示,本文算法在样本个数不同的情况下,分类所需的时间最少,这更体现出了本文算法的优越性。

## 6　结论

本文提出了一种基于遗传算法的 SVM 决策树多分类方法并将其应用于低空声目标的识别当中,利用遗传算法得到了对五种不同声目标分离效果最佳的决策树结构,用其构建 SVM 分类器对实际数据进行测试,结果表明该方法较以往的多分类方法在分类精度、分类时间上都有了一定的提高,在低空被动目标识别中具有良好的应用前景。

### 参考文献


[1] 刘辉,杨俊安,许学忠:基于 MFCC 参数和 HMM 的低空飞行目标声识别方法研究[J]. 弹箭与制导学报, 2007, 27(5):217-222.

[2] 柳回春,马树元:支持向量机的研究现状[J]. 中国图象图形学报,2002, 7(6):619-623.

[3] 李京华,许家栋,李红娟:支持向量机的战场直升机目标分类识别[J]. 火力与指挥控制, 2008, 33(1):31-34.

[4] C. Cortes, V. Vapnik: Support-vector network[J]. Machine Learning, 1995,20:273-297.

[5] 苟博,黄贤武:支持向量机多类分类方法[J]. 数据采集与处理,2006,21(3):335-339.

[6] J. Manikandan, B. Venkataramani, V. Amudha: A Novel Technique for Support Vector Machine based Multi-class Classifier[C]. TENCON 2008. Hyderabad India: TENCON 2008 IEEE region 10 conference publication,2008. 1-6.

[7] 连可,黄建国,王厚军,龙兵:一种基于遗传算法的 SVM 决策树多分类策略研究[J]. 电子学报, 2008, 36(8): 1502-1507.


### 作者简介


王　一(1985-),男,浙江萧山人,硕士研究生,研究方向为信号分析与识别技术。E-mail:wygggg@126.com



杨俊安(1965-),男,安徽巢湖人,教授,博士,博士生导师,研究方向为信号处理、智能计算等。



刘　辉(1983-),男,安徽阜阳人,博士研究生,研究方向为信号分析与识别技术。