

混合策略超球支持向量机算法

刘 爽^{1,2}, 陈 鹏³, 李克秋¹

(1 大连理工大学 计算机科学与技术学院, 辽宁 大连 116024; 2 大连民族学院 计算机科学与工程学院, 辽宁 大连 116605;
3 大连东软信息学院 计算机科学与技术系, 辽宁 大连 116023)

摘 要: 针对多类别分类超球支持向量机算法的重叠区域数据分类问题, 提出了一种混合策略决策算法. 首先对超球相交区域的数据分布情况分析得到数据分布的特点, 然后根据数据分布特点采用不同的决策策略. 如果用两球相交面直接可以把两类数据分开, 则直接用相交面作为分类平面. 如果两类数据近似线性可分, 构造最优二分超平面作为分类平面. 如果两类数据非线性可分, 则引入核函数构造最优二分超平面为分类球面. 如果相交区域只包含一个类别的数据, 则采用排它法作为测试样本的决策规则. 实验结果表明所提出的算法性能优于单一决策策略的超球支持向量机算法, 在提高分类精度的同时, 降低了决策规则求解的复杂度.

关键词: 超球支持向量机; 重叠区域; 混合策略; 多类别分类

中图分类号: TP389.1

文献标识码: A

文章编号: 1000-7180(2014)07-0001-05

An Hyper-sphere Support Vector Machine with Hybrid Decision Strategy

LIU Shuang^{1,2}, CHEN Peng³, LI Ke-qiu¹

(1 School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;

2 College of Computer Science & Engineering, Dalian Nationalities University, Dalian 116605, China;

3 Department of Computer Science & Technology, Dalian Neusoft University of Information, Dalian 116023, China)

Abstract: In order to solve classification problem of the intersections for multi-class classification based on hyper-sphere support vector machines, a hybrid decision strategy is put forward in this paper. First, characteristics of data distribution in the intersections are analyzed and then decision class is decided by different strategies. If training samples of two classes in the intersection can be classified by intersection hyper-plane for two hyper-spheres, then new test samples can be decided by this plane. If training samples of two classes in the intersection can be approximately linearly classified, new test samples can be classified by standard optimal binary-SVM hyper-plane. If training samples of two classes in the intersection can not be linearly classified, new test samples can be decided by introducing kernel function to get optimal classification hyper-plane. If training examples belong to only one class, then new test samples can be classified by exclusion method. Experimental results show performance of our algorithm is more optimal than hyper-sphere support vector machines with only one decision strategy. And this algorithm improves the performance and decreases computation complexity.

Key words: hyper-sphere support vector machine; intersection region; hybrid decision strategy; multi-class classification

1 引言

Vapnik 于 1995 年在统计学习理论基础上提出了支持向量机算法^[1], 目前已经成功应用到很多领域. 最初的各种支持向量机算法均是解决两类分类

问题的, 但实际应用中最常见的是多类别分类问题, 因此需要把两类分类问题推广到多类分类. 目前的支持向量机多类分类问题^[2]分为四大类: (1) 第一大类采用组合二分分类器的方法, 计算复杂度较高; (2) 第二大类是把多分类面的所有参数求解合并

收稿日期: 2013-10-08; 修回日期: 2013-11-23

基金项目: 国家自然科学基金项目(71303031)

为一个复杂的优化问题,导致计算的复杂度很高,计算结果也不一定能保证;(3) 第三大类结合树形数据结构优化决策过程,如有向无环图^[3]建立的二叉树可以逐层缩减决策所需的计算量;(4) 第四大类为超球支持向量机算法,降低了计算的复杂度.文献[4-6]中提出了各自的超球支持向量机算法,但尚存在不足之处.

本文根据测试样本在超球划分空间的位置,采用不同的决策规则.根据样本分布特点,提出了新的测试样本决策分类算法.

2 超球支持向量机描述

假设训练数据维数为 n ,训练样本类别数为 m ,第 k 个类别的数据集合样本个数为 l_k .给定 m 个 n 维空间的集合 $X_k(k=1, \dots, m)$,对每个集合寻找一个超球 (c_k, R_k) ,其中 c_k 为超球球心, R_k 为超球半径,使其尽量小,使该最小超球尽可能包含所有同类样本.考虑到存在一些孤立点,允许这些点落入球面外,引入松弛因子,该原始的优化问题可以表示为

$$\begin{aligned} \min_{c_k, R_k} & R_k^2 + C \sum_{i=1}^m \xi_i \\ \text{subject to} & \\ & \| \varphi(\mathbf{x}_i) - c_k \|^2 \leq R_k^2 + \xi_i \\ & \xi_i \geq 0, i = 1, \dots, l_k \end{aligned} \quad (1)$$

式中,第 k 类样本的最小超球 S_k 由球心 c_k 和半径 R_k 确定, C 为惩罚因子, $\xi_i \geq 0$ 为松弛变量.对非线性的训练集,引入非线性映射 $\varphi(\mathbf{x}_i)$ 把训练数据 $\mathbf{x}_i$ 映射到高维线性特征空间.引入拉格朗日乘子 α_i, β_i 求解该二次规划问题,拉格朗日多项式为如下形式:

$$\begin{aligned} L(R_k, c_k, \xi_i, \alpha_i, \beta_i) = & R_k^2 + C \sum_{i=1}^{l_k} \xi_i - \\ & \sum_{i=1}^{l_k} \alpha_i (R_k^2 + \xi_i - \| \varphi(\mathbf{x}_i) - c_k \|^2) - \sum_{i=1}^{l_k} \beta_i \xi_i \end{aligned} \quad (2)$$

对该多项式分别对 R_k, c_k, ξ_i 求偏导并令其为 0, 可得

$$\frac{\partial L}{\partial R_k} = 2R_k - 2R_k \sum_{i=1}^{l_k} \alpha_i = 0 \Rightarrow \sum_{i=1}^{l_k} \alpha_i = 1 \quad (3)$$

$$\begin{aligned} \frac{\partial L}{\partial c_k} = & 2 \sum_{i=1}^{l_k} \alpha_i \varphi(\mathbf{x}_i) - 2c_k \sum_{i=1}^{l_k} \alpha_i = 0 \Rightarrow \\ c_k = & \frac{\sum_{i=1}^{l_k} \alpha_i \varphi(\mathbf{x}_i)}{\sum_{i=1}^{l_k} \alpha_i} = \sum_{i=1}^{l_k} \alpha_i \varphi(\mathbf{x}_i) \end{aligned} \quad (4)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i = C - \beta_i \quad (5)$$

$$\Rightarrow 0 \leq \alpha_i \leq C$$

把式(3)、(4)、(5) 带入拉格朗日多项式,可以得到对偶的优化问题:

$$\begin{aligned} \min_{\alpha_i} & \sum_{i,j=1}^{l_k} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{l_k} \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) \\ \text{s. t.} & \sum_{i=1}^{l_k} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C, i = 1, \dots, l_k \end{aligned} \quad (6)$$

对非线性函数,无需直接计算,在求解过程中引入核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$ 即可计算.

求解完对偶的二次规划问题,可以计算出拉格朗日乘子,非零的拉格朗日乘子,即 $\alpha_i > 0$ 所对应的样本点 $\mathbf{x}_i$ 为支持向量.根据支持向量可以求解各超球的球心和半径.

本文中选用常见的高斯 RBF 核函数,其计算公式为 $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\| \mathbf{x}_i - \mathbf{x}_j \|^2 / (2\sigma^2))$.

3 测试样本分类分析

给定训练样本,计算完所有类别的超球后,新的测试样本或者属于两超球的相交区域,或者不属于任何超球.只有二维、三维超球计算具有直观效果,因此这里以二维为例加以说明.

图 1 为两个超球相离的示意图.假设有一个新的测试样本点 x ,该点到两球心的距离相等.

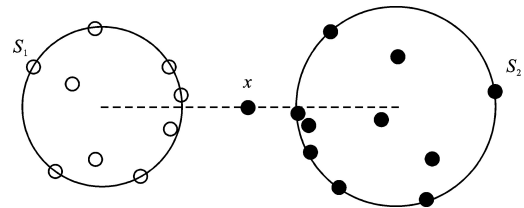


图 1 两球相离,点 x 到两球心距离相等

判断测试样本的类别时不仅应该考虑数据的中心(即球心),还应该考虑数据的分布(即半径),因此采用如下计算公式:

$$\min_j \left(\frac{d_{xj}^2}{R_j^2} - 1 \right) \quad (7)$$

其中

$$\begin{aligned} d_{xj}^2 = & \| \varphi(x) - c_j \|^2 \\ = & K(x, x) - 2 \sum_{i=1}^{l_k} \alpha_i K(x, x_i) + \\ & \sum_{i,j=1}^{l_k} \alpha_i \alpha_j K(x_i, x_j) \end{aligned}$$

对图1中情况,如果单纯考虑测试样本与超球的球心距离,而不考虑相对距离,则测试样本可以属于两个类别,很显然不符合训练样本的分布规律.对两个分类超球而言,显然 $R_1 < R_2$,则按该公式计算可得 $\frac{D^2}{R_1^2} - 1 > \frac{D^2}{R_2^2} - 1$,因此测试样本 x 属于类别2,符合训练样本的数据分布规律.或者用更直观的解释,判别公式的计算结果越大,说明测试样本与超球的球心距离相对比值越远,测试样本属于该类别的可能性越小.

图2所示为二维情况下两个超球相交的情况.假定类别1的训练样本分类超球为 S_1 ,类别2的训练样本分类超球为 S_2 , S_1' 为 S_1 在超球 S_2 中的相交部分, S_2' 为 S_2 在超球 S_1 中的相交部分.这里首先对文献[4]的思想简单加以描述.该文提出了提出子超球支持向量机算法进行重叠区域的样本分类.该算法首先求出相交的超球,把重叠区域的样本点划分为同类错误样本点(与母超球所代表的样本类别相同的样本点)和异类错误样本点(与母超球所代表的样本类别不同的样本点),然后计算同类错误样本点子超球和异类错误样本点子超球.测试样本如果属于重叠区域,则采用同类错误样本点子超球和异类错误样本点子超球来判定其正确的类别.其算法描述如图2所示.

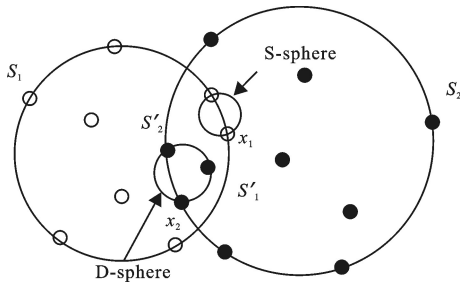


图2 子超球支持向量机示意图

由于建立母子超球求解二次规划问题计算代价较高,本文为了降低计算的复杂度,对两球相交区域数据分布及决策分为如下三种情况讨论.

第一种情况,两球相交平面直接可以把两类数据分开,如图3所示.此时,直接用相交超平面作为测试样本的分类超平面即可实现正确分类.

第二种情况,两球相交平面无法直接分开两类数据,则引入松弛因子,数据线性可分情况下构造的分类超平面如图4所示,该超平面即为测试样本的决策超平面.

数据线性不可分情况下,则引入核函数和松弛因子,求解非线性的二分最优超平面,构造的超平面

如图5所示.

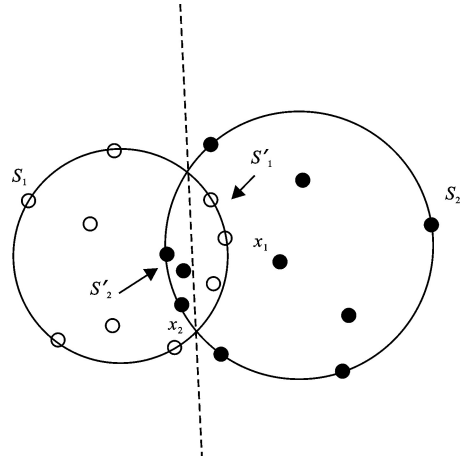


图3 相交平面分类示意图

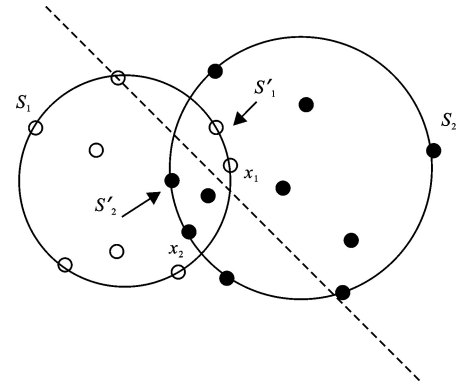


图4 二分最优分类超平面分类示意图

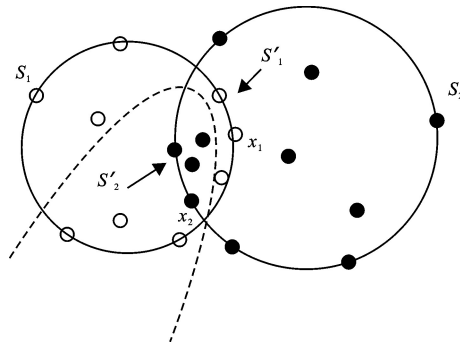


图5 引入核函数线性不可分的分类超平面示意图

第三种情况,相交区域只包含一类数据,如图6所示,相交区域只包含类别2的样本,则认定新的测试样本落入相交区域时只属于该类别,反之亦然.

4 算法的具体实现

这里假设训练数据维数为 n ,训练样本类别数为 m ,第 k 个类别的数据集合样本个数为 l_k ,测试样本个数为 M 个.算法实现步骤如下:

步骤1 读取各类别的训练样本数据;
对第 k 类训练样本

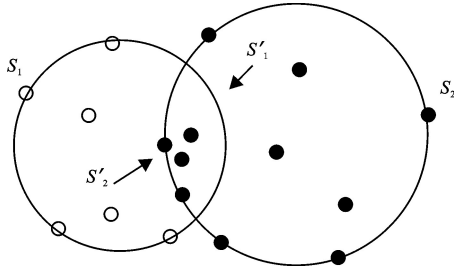


图6 相交区域只有一类数据

for $k = 1$ to m do{

利用 SMO 算法求解二次规划问题, 得到拉格朗日乘子解向量 α ;

由 α 计算出 R 半径的平方, 得到球模型;

}//end of loop

步骤 2 记第 k 个超球半径为 R_k , 圆心为 c_k , 计算所有超球两两之间的位置关系:

for $i = 1$ to m do

for $j = i + 1$ to m do

{

/(第 i 个与第 j 个超球球心之间的欧氏距离)

$D_{i,j} = \|c_i - c_j\|^2$;

if $|R_i - R_j| \leq D_{i,j} \leq |R_i + R_j|$, 超球 S_i, S_j 相交, 记入 SR_i ;

if $D_{i,j} > |R_i + R_j|$, 超球 S_i, S_j 外离, 记入 SR_d ;

}

步骤 3 对 SR_i 中每对超球 S_i, S_j , 即 S_i, S_j 相交, 记第 j 类超球 S_j , 记 S_j' 为超球 S_j 与 S_i 的重叠区域, 集合 C_j 为 S_j' 的异类错误样本点, A_j 为相交部分 S_j' 的同类样本点。

步骤 3.1 如果 $|C_j| = 0$ 或 $|A_j| = 0$, 则相交区域只有一种类别样本, 当测试样本落入相交区域时其类别判断直接属于样本个数不为 0 的样本类别, 记入排它法分类文件;

步骤 3.2 如果 C_j 与集合 A_j 个数均不为 0.

步骤 3.2.1 如果 $|C_j| > 1$ 且 $|A_j| > 1$, 计算 S_i, S_j 相交的交线超平面, 这里直接用超球方程相减即可得到相交的超平面; 如果超平面直接把两类数据分开, 具体实现过程为计算测试样本点到球心 c_j 的向量与相交平面法矢量的点积结果, 根据点积结果的符号判定相交区域点在相交平面的哪一侧. 如果两类数据在相交平面的两侧, 存入分类超平面文件;

步骤 3.2.2 如果步骤 3.2.1 计算的相交平面不能直接区分相交区域的训练样本, 则引入核函数

与松弛因子, 构造二分最优超平面, 记入超平面模型文件;

步骤 3.2.3 如果 $|C_j| = 1$ 且 $|A_j| = 1$, 则两个超球为相切关系, 测试样本类别判定属于两类别均可。

步骤 4

for $i = 1$ to M do{// 对第 i 个测试样本

for $j = 1$ to m do{// 对第 j 个球模型

计算样本 i 到球 S_j 距离的平方 d_{ij}^2 ;

如果 $d_{ij}^2 - R_j^2 \leq 0$, 则样本 i 落入 S_j 内, 否则落入球外

}// end of j loop

统计样本 i 落入的超球个数 N ;

如果 $N = 1$, 则 i 落入唯一的超球, 该测试样本属于该类别;

如果 $N = 0$, 则 i 落入所有球外, 对所有超球计

算 $\min(\frac{d_{ij}^2}{R_j^2} - 1)$ 确定其所属的类别;

如果 $N > 1$, 设样本 i 落入超球 $S_j, j = 1, \dots, r$

for $j = 1$ to r do

for $e = j + 1$ to r do

{

if $S_j, S_e \in SR_i$ {

if 存在分类超平面, 直接根据步骤 3.2.2 或步骤 3.2.3 得到的超平面方程计算测试样本 i 所属类别;

if 不存在分类超平面, 则按步骤 3.1 计算的结果确定类别归属分类结果;

if $i \in$ spherical surface of S_j, S_e , 由 i 到圆心的距离是否等于半径判断, 则 i 可以属于两个类别, 此时 i 是两个超球相交区域的交点(或是相切点)

}//end of if

}//end of intersection computation

}//end of outer i loop

5 实验

为了验证本文提出的混合策略决策超球支持向量机算法的性能, 选取三类数据. 第一类选取人工生成的数据, 即二维数据, 因其计算结果更具有直观性; 第二类选取 UCI 机器学习数据库中的 Iris 数据集^[7]; 第三类选取实际应用项目中视频监控数据的异常行为分类数据。

第一类别人工生成数据训练数据如图 7 所示. 人工生成的数据分为两类, 黑色圆点代表第一类数

据,灰色圆点代表第二类数据,很显然该数据集非线性可分,直接用超平面无法正确分类,引入 RBF 核函数,计算的分类结果如图 7 所示。

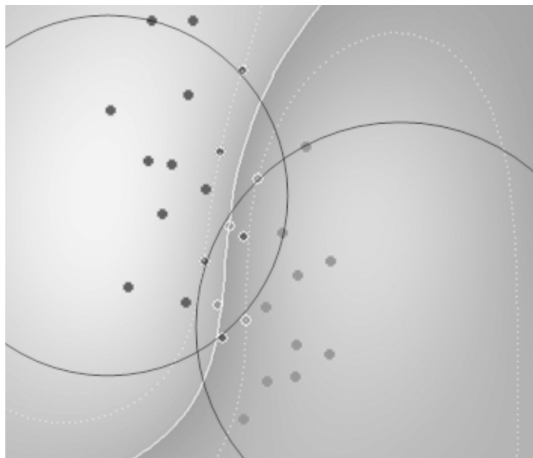


图 7 超球分类与非线性核函数分类对比

在图 7 中,左侧的分类平面为第一类数据的分类面;右侧的分类面为第二类数据的分类面。这两类数据对应的超球如图 7 所示,两个超球所包含的数据范围及分类面的界定与非线性核函数界定的分类平面具有相似的范围。新的测试样本如果落入单一超球,其类别很容易确定;如果落入右侧超球外部,按本文中的类别判定公式计算,其类别归属是第二类;如果落入相交区域,不考虑孤立点的情况下,相交区域中样本属于第二个类别。因此测试样本的归属也与非线性决策函数类似。

第二类别 Iris 数据集共有三个类别的 150 个样本,实验中随机把数据集分为三部分,每个类别选取 45 个样本做训练数据,5 个样本做测试数据。重复 10 次取实验结果的平均值。实验中原有的 Iris 数据集经过数据标准化处理、参数寻优训练后得到优化的参数值,RBF 核函数宽度 $\sigma=3$,惩罚因子 $C=0.1$,对测试样本的测试准确率平均达到 99.86%,训练时间平均为 3.566 s。实验结果如表 1 所示。而对相同的数据集合,文献[4]中的分类平均准确率为 98%,训练时间平均为 4.235 s。

表 1 Iris 数据集实验结果

类别	超球半径	支持向量个数/ 训练样本总数	准确率/%
1	0.418 26	12/45	100
2	0.473 10	11/45	99.6
3	0.575 19	14/45	99.98

第三类别为实际应用数据,选取视频监控系统

中异常行为识别的分类数据,该分类数据的特征提取维数为 20 维,类别数为 6 类,数据集大小为 9 000 个样本,随机选取其中的 7 800 个样本每个类别 1 300 个样本作为训练数据集合,1 200 个样本每个类别 200 个样本作为测试数据集合,重复五次随机选取求平均值作为实验结果。表 2 是本文算法与文献[4]子超球算法、文献[3]的超球支持向量机算法的分类时间、准确率和支持向量个数的对比结果。

表 2 本文算法与子超球算法实验结果对比

算法	训练时间/s	支持向量个数/ 训练样本总数	准确率/%
本文算法	936.35	536/7 800	90.26
子超球算法	1 525.60	729/7 800	88.45
超球算法	1 120.16	635/7 800	82.98

上述三类实验数据实验结果表明新的决策方法性能优于单一决策算法,且计算代价小、训练时间较快。

6 结束语

本文首先介绍了多分类支持向量机算法的分类,选用计算代价相对较低的超球支持向量机算法讨论。为了准确的对测试样本分类,有效降低计算二次规划的复杂度,分为四种情况讨论测试样本的分类规则,并给出了详细的算法实现步骤。对基准数据集及实际视频监控异常行为数据的分类进行测试,该混合策略决策算法不仅分类性能有所提高,而且模型的泛化性能较好,计算复杂度相对较小。

参考文献:

- [1] Vapnik V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Networks (S1045-9227), 1999,10(5):988-999.
- [2] Hsu Chihwei, Lin Chihjen. A comparison of methods for multi-class support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13 (2): 415-425.
- [3] Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification [C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000(12):547-553.
- [4] Zhu M L, Chen S F, Liu X D. Sphere-structured support vector machines for multi-class pattern recognition [J]. Lecture notes in computing science, 2003(2369): 589-593.

(下转第 9 页)

240 的图像来说,完成高斯差分尺度空间的构造仅需要 0.59 ms.

表1 消耗 FPGA 资源的比较

	Bonato ^[2]	提出的结构
存储器容量/Mbit	0.91	0.085
寄存器	7 256	6 487
查找表单元	15 137	16 182
最大工作频率/MHz	149	131

5 结束语

DoG 模块是 SIFT 算法硬件加速中的关键模块.该模块中延迟线数量多,需要的存储容量大.为了降低延迟线的数量,本文提出了一种基于并行平滑机制的并行滤波结构.测试结果表明,在相似的误匹配率下,本文提出的结构延迟线数量降低了 84%,进而使得存储容量降低了 90%左右.在 131 MHz 的工作频率下,对于 320×240 的图像来说,完成高斯差分尺度空间的构造仅需要 0.59 ms.

参考文献:

- [1] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004,60(2):91-110.
- [2] Bonato V, Marques E, Constantinides G A. A parallel hardware architecture for scale and rotation invariant

feature detection[J]. IEEE Trans. Circuits System Video Technology, 2008,18(12):1703-1712.

- [3] Chang L, Hernandez-Palancar J. A Hardware architecture for SIFT candidate keypoints detection [C]// CIARP. Berlin, Germany, 2009:95-102.
- [4] Yao L, Feng H, Zhu Y, et al. An architecture of optimised SIFT feature detection for an FPGA implementation of an image matcher[C]//FPT. Sydney, Australia, 2009:30-37.
- [5] Huang F C, Huang S Y, Ker J W, et al. High-Performance SIFT hardware accelerator for real-time image feature extraction[J]. IEEE Trans. Circuits System Video Technology, 2012, 22(3):340-351.
- [6] Lindeberg T. Scale-space theory: a basic tool for analysing structures at different scales[J]. Journal of Applied Statistics, 1994,21(2):224-270.
- [7] Lindeberg T. Feature detection with automatic scale selection[J]. International Journal of Computer Vision, 1998,30(2):79-116.
- [8] Mikolajczyk K. Affine covariant features[EB/OL]. (2007-07-15). [2014-04-20]. <http://www.robots.ox.ac.uk/~vgg/research/affine/>.

作者简介:

张 静 女,(1980-),博士研究生.研究方向为数字图像处理和数字集成电路设计.

桑红石 女,(1969-),副教授.研究方向为数字图像处理和数字集成电路设计.

(上接第5页)

- [5] 吴强,贾传茨. 球结构支持向量机的改进算法及仿真研究[J]. 系统仿真学报, 2008, 20(2):345-348.
- [6] 刘爽,陈鹏. 改进的超球支持向量机算法[J]. 计算机工程与应用,2009(16):149-151.
- [7] Fisher R A. Iris plants database[DB/OL]. (1988-07-15). [2013-10-11]. <http://www.ics.uci.edu/~ml-earn/MLRepository.html>.

作者简介:

刘 爽 女,(1977-),博士,副教授.研究方向为智能信息处理.

陈 鹏 男,(1977-),硕士,副教授.研究方向为 Web 技术.

李克秋 男,(1971-),博士,教授,博士生导师.研究方向为云计算、大数据、数据中心网络.