

用于多类别分类的一种加权超球支持向量机算法

刘 爽¹, 陈 鹏², 李锡祚¹

(1 大连民族学院 计算机科学与工程学院, 辽宁 大连 116600;

2 大连东软信息学院 计算机科学与技术系, 辽宁 大连 116023)

摘 要: 在 One-Class 基础上发展起来的超球支持向量机算法能有效地解决多类别分类问题. 但是原始的超球支持向量机算法仍有很多需要改进的地方. 经过推导和实验, 得到如下结论, 即超球支持向量机算法过度依赖于每个训练样本, 即使该训练样本为噪音数据或是离群异常数据. 因此提出在训练之前加入预处理算法, 通过相似度计算删除噪音点和异常点. 在训练过程中, 根据公式计算每个样本的权值, 区别对待每个训练样本, 确保 SMO 求解过程迅速收敛. 在测试阶段, 根据测试点的位置合理选择分类规则进行正确分类. 实验结果表明提出的算法可以有效减少噪音数据和异常数据对分类结果的影响, 同时提高了分类精度.

关键词: 超球支持向量机; 加权; 异常点; 噪音; 多类别分类

中图分类号: TP389.1

文献标识码: A

文章编号: 1000-7180(2015)01-0019-05

A Weighted Hyper-Sphere Support Vector Machine for Multi-class Classification

LIU Shuang¹, CHEN Peng², LI Xi-zuo¹

(1 College of Computer Science & Engineering, Dalian Nationalities University, Dalian 116600, China;

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

Abstract: Based on One-Class support vector machine, Hyper-sphere support vector machine is an efficient tool to solve problems of multi-class classification. With theoretical deduction and experiments, we found that hyper-sphere support vector machine is over dependent on each training sample, even if the training sample belongs to noise or outlier. So a preprocessing method before training is proposed in this paper to remove noise and outlier by similarity computation from the original training data set. In the training process, a weight factor is added for each sample to differentiate its contribution to the resulting classifier, ensuring quick convergence of the algorithm based on SMO solving. In the testing process, an appropriate classification rule is selected by the test sample position. Experimental results compared to other hyper-sphere support vector machines, new proposed algorithm reduces influences of noises or outliers for the resulting classifier and improves classification accuracy.

Key words: Hyper-sphere support vector machine; weight; outliers; noise; multi-class classification

1 引言

支持向量机由 Cortes 和 Vapnik 于 1995 年首先提出, 它在解决小样本、非线性及高维模式识别中表现出许多特有的优势, 并已成功应用到图形图像

处理、数据挖掘、文本分类等众多领域^[1-3]. 最初的支持向量机算法以二分类为基础, 解决多分类问题时使用各种组合二分类支持向量机的方法实现. 由于每一个二分支支持向量机求解过程都是一次二次规划求解过程, 因此组合二分类支持向量机求解多分类

收稿日期: 2014-04-05; 修回日期: 2014-05-29

基金项目: 国家自然科学基金资助项目(71303031)

问题时计算代价较高. 为了降低二次规划求解的复杂度, 把二类分类的超平面二次规划求解问题转化成超球面的二次规划求解, 研究者提出了直接计算出各类别约束超球的超球支持向量机算法^[4]. 该算法在 One-class 支持向量机算法基础上发展而来^[5], 其思想是单独计算每一类别训练数据的最小包围球, 要求建立的最小包围球半径尽可能小, 尽量包含该类别的所有样本点^[6-7].

基本的超球支持向量机算法训练过程需要逐一扫描整个训练集中的每个训练样本, 即公平对待每个训练样本, 即便该样本是噪声数据点或离群异常数据点. 而且, 根据最后训练结果支持向量的比例及数量可知, 仅有小部分样本点对每个类别的分类超球有贡献, 以超球中心为聚类中心的大部分数据点不会影响分类结果.

本文提出一种引入预处理算法的加权超球支持向量机算法. 其主要思想是在训练之前加入预处理算法, 根据相似度公式计算训练样本的相似度值, 小于预设阈值的数据点记为噪音点和异常点, 直接从原始数据集中删除. 在训练过程中, 根据权值公式计算每个样本的权值, 根据权值大小优先选择对分类结果影响较大的训练样本, 确保 SMO 求解过程的迅速收敛. 在测试阶段, 根据测试点的位置合理选择分类规则进行正确分类.

2 加权超球支持向量机算法

为了解释权值计算规则, 这里先给出一个二维数据点的最小包围球示意图, 如图 1 所示. 球面上和球面外黑色填充的数据点为支持向量.

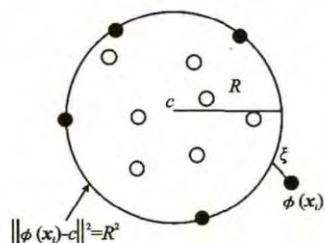


图 1 二维最小包围球示意图

加权因子 s_i 的计算公式为

$$s_i = \frac{\|x_i - \bar{x}\|^2}{\max_j \|x_j - \bar{x}\|^2}, i, j = 1, \dots, N \quad (1)$$

这里的 $\bar{x}$ 为所有训练样本 x 的平均值. 由于超球球心与样本均值接近, 因此根据该计算公式, 球面上的点权值近似为 1, 而靠近球心的点权值非常小, 在调用 SMO 算法求解时就可以直接忽略小于一定

阈值的训练样本, 而优先选择权值较大的样本点, 加快算法收敛速度.

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

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

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 $\varphi(x_i)$ 把训练数据 x_i 映射到高维线性特征空间. 对非线性映射无需直接计算, 在求解过程中引入核函数 $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ 即可计算.

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:

$$\begin{aligned} L(R_k, c_k, \xi_i, \alpha_i, \beta_i) &= R_k^2 + C \sum_{i=1}^{l_k} s_i \xi_i - \\ & \sum_{i=1}^{l_k} \alpha_i (R_k^2 + \xi_i - \|\varphi(x_i) - c_k\|^2) - \sum_{i=1}^{l_k} \beta_i \xi_i \end{aligned} \quad (3)$$

对该多项式分别对 R_k, c_k, ξ_i 求偏导并令其为 0, 可得

$$\frac{\partial L}{\partial R_k} = 2R_k - 2R_k \sum_{i=1}^{l_k} \alpha_i = 0 \Rightarrow \sum_{i=1}^{l_k} \alpha_i = 1 \quad (4)$$

$$\frac{\partial L}{\partial c_k} = 2 \sum_{i=1}^{l_k} \alpha_i \varphi(x_i) - 2c_k \sum_{i=1}^{l_k} \alpha_i = 0 \Rightarrow$$

$$c_k = \frac{\sum_{i=1}^{l_k} \alpha_i \varphi(x_i)}{\sum_{i=1}^{l_k} \alpha_i} = \sum_{i=1}^{l_k} \alpha_i \varphi(x_i) \quad (5)$$

$$\begin{aligned} \frac{\partial L}{\partial \xi_i} &= Cs_i - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i = Cs_i - \beta_i \\ &\Rightarrow 0 \leq \alpha_i \leq Cs_i \end{aligned} \quad (6)$$

把式(4)、(5)、(6) 带入拉格朗日多项式, 可以得到对偶的优化问题:

$$\begin{aligned} \min_{\alpha_i} & \sum_{i,j=1}^{l_k} \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{l_k} \alpha_i K(x_i, x_i) \\ s.t. & \sum_{i=1}^{l_k} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C s_i, i = 1, \dots, l_k \end{aligned} \quad (7)$$

由运算结果可知,加权超球支持向量机算法与原始的超球支持向量机算法目标函数的区别在于拉格朗日乘子的上限。原始的超球支持向量机算法拉格朗日乘子上限是固定的惩罚因子 C ,无法动态调整。根据惩罚因子的引入原则可知, C 越大,对错误分类惩罚越大; C 越小,对错误分类的惩罚越小,可能会导致错误分类数据点增多。如果要获得最佳的惩罚因子,需要使用不同的惩罚因子反复迭代训练,增加了计算量。而加权超球支持向量机算法的拉格朗日乘子上限是惩罚因子 C 与权值 s_i 的乘积,不仅可以动态调整,而且还与样本相关联。按照权值计算公式(2)的定义,可知所有的权值均小于等于 1。权值越大,说明该样本点接近球面的可能性越大,也就意味着该样本点可能是支持向量,因此 $C s_i$ 的值也大一些。权值越小,说明该样本点越靠近球心, $C s_i$ 的值较小。

由于无需求解非线性映射函数,因此求解超球半径时,不是直接按照式(5)求解,而是直接求球心的平方,导出如下计算公式:

$$c_k^2 = \sum_{i=1, j}^{l_k} \alpha_i \alpha_j K(x_i, x_j) \quad (8)$$

拉格朗日乘子大于 0 的样本为支持向量,因此满足公式(1)约束条件 1 的球面上的点即为支持向量。已知球心,可以由任一球面上的支持向量(此时,松弛因子为 0)求出半径的平方,即:

$$\begin{aligned} R_k^2 &= \|\varphi(x) - c_k\|^2 \\ &= K(x, x) - 2 \sum_{i=1}^{l_k} \alpha_i K(x, x_i) + \sum_{i=1, j}^{l_k} \alpha_i \alpha_j K(x_i, x_j) \end{aligned} \quad (9)$$

3 噪音点 / 离群点对分类结果的影响分析

这里以二维超球为例直观解释噪音点 / 离群点对分类结果的影响。假设类别 1 与类别 2 的训练超球结果分别为 S_1 和 S_2 ,如图 2 所示。那么此时的分类结果是最理想的,因为两个超球没有任何相交的部分,对测试样本的分类结果计算也相对简单。但是,如果第一个类别的训练数据集中包含一个噪音数据或离群点,如点 x ,那么该点对分类结果将产生巨大影响。这时,虽然仅仅加入了这一个点,但是

因为该点所处位置的特殊性,使第一个类别的超球半径增长 2 倍以上,分布范围迅速扩张。最后的分类超球就变为 S' 和 S_2 ,即两个分类超球有 90% 以上的重合区间,使测试样本的分类变得非常复杂。

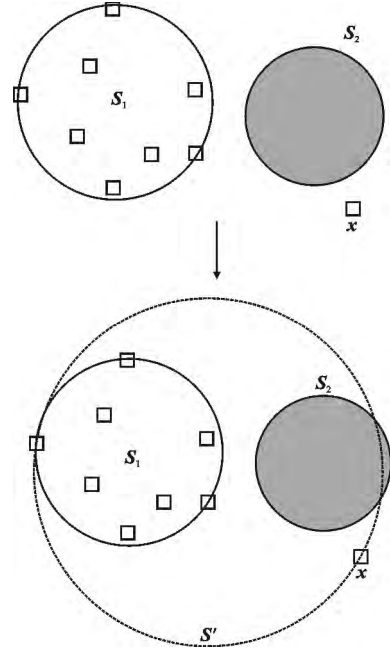


图 2 噪音 / 离群点对分类结果的影响

这里简单对测试样本分类规则加以说明。如果分类超球相互独立,采用综合超球球心、半径分布的分类规则计算公式计算测试样本的分类结果。如果一个分类超球位于另一个分类超球内部,那么采用排它法作为分类规则,即落入小超球的测试样本类别属于小类别,其余所有测试样本均属于大超球的类别。对图 3 所示的情况,假设有一个测试样本 y ,位于超球 S_2 附近。因为 S_2 在 S' 的内部,所以直接判定测试样本的类别为 1。但实际上,训练样本 x 为噪音数据,不考虑该训练样本时,按球心、半径分布规律 y 的正确分类结果应该是属于类别 2。

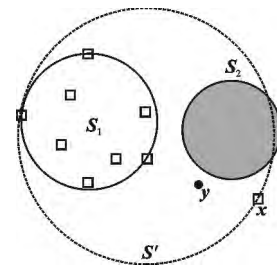


图 3 测试样本 y 的分类示意图

为了降低噪音 / 离群数据对分类结果的影响,采用文献[8]中提出的相似度判定方法,删除训练样本集合中的噪音 / 离群数据。其计算规则如下:计

算单个样本与整个样本的相似度,相似度大的点属于训练样本集合,相似度小于一定阈值的点即为噪音点/离群点,从原始训练数据集中删除.其计算方法如下:

$$D(x_i, X_k) = \sum_{j=1}^m \exp(-\frac{\|x_i - x_j\|^2}{\sigma})$$

$$\text{where } \sigma = \frac{1}{m} \sum_{x_j \in X_k} \|x_j - \bar{x}\|^2 \quad (10)$$

$$i, j = 1, \dots, N; k = 1, \dots, m$$

这里, $\|\cdot\|^2$ 表示欧氏距离的平方, $\bar{x}$ 为所有样本点的均值, m 为样本类别数, N 为样本总数. 噪音点/离群点应该与数据集合主体距离较远,且所占比例较小,这里根据样本分布规律设定一个阈值.

在对测试样本进行分类时,分类规则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).

本文分如下三种情况考虑. 如果测试样本位于多个超球内(超球个数大于等于 2) 时, 根据如下公式计算该测试样本所属的类别:

$$\text{sim}(x, S_k) = \frac{\|x - c_k\|^2}{R_k^2} (k = 1, \dots, m) \quad (11)$$

新的测试样本类别归属于相似度大的类别.

如果测试样本位于所有超球外部时,使用如下分类决策公式确定其类别归属:

$$\min(\frac{d_{xj}^2}{R_j^2} - 1) \quad (12)$$

其中

$$d_{xj}^2 = \|\varphi(x) - c_j\|^2$$

$$= K(x, x) - 2 \sum_{i=1}^{l_k} \alpha_i K(x, x_i) + \sum_{i,j=1}^{l_k} \alpha_i \alpha_j K(x_i, x_j) \quad (13)$$

如果测试样本仅位于单一超球内,则测试样本就属于该类别.

4 加权超球支持向量机算法及实验

4.1 算法描述

图 4 所示为本文加权超球支持向量机的训练算法描述,图 5 所示为本文加权超球支持向量机的测试分类算法描述.

4.2 实验描述

为了验证本文算法的有效性,取三组数据进行验证.

第一组数据人工生成的二维数据为三个类别数

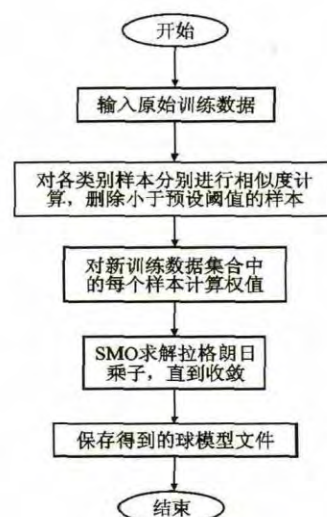


图 4 加权超球支持向量机训练算法

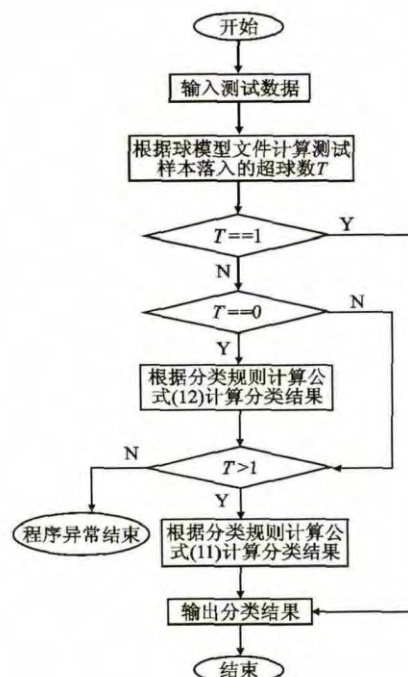


图 5 加权超球支持向量机测试算法

据,如图 6 所示. 横纵坐标分别表示二维数据的 x 坐标值, y 坐标值.

根据三个类别数据的分布求得其各自的最小半径约束超球如图 6 所示. 此时分类超球分布属于最理想状态,对新的测试样本可以直接根据测试样本的位置选择分类规则公式计算其类别. 但是,如果存在第二个类别的噪音数据 x ,由 SMO 算法求得的分类超球结果如图 7 所示. 横纵坐标分别表示二维数据的 x 坐标值, y 坐标值. 这时,第二个类别的分类超球与第一类别、第三类别超球都相交,使落入相交区域的测试样本分类变得较为复杂,而且分类结果也不一定准确.

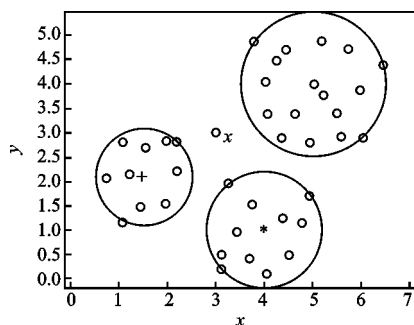
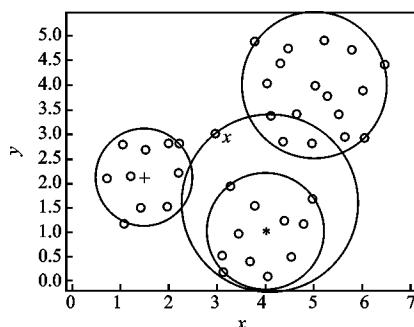


图6 人工生成三个类别超球例子

图7 噪声数据 x 对分类结果的影响

按照本文的算法,在输入原始训练数据集后,首先做数据预处理,删除相似度小于设定阈值的数据,就可以剔除掉噪音数据 x ,从而保证分类超球的准确性。

第二组数据取自 UCI 机器学习数据库^[9]中的 iris 数据集,iris 数据集包括三类数据 150 个,每个类别 50 个,分别加入人工噪声数据 2 个,随机选择 90% 作为训练样本,10% 作为测试样本,重复 10 次实验,取计算结果的平均值。这里的核函数均选取高斯 RBF 核函数,本文的算法经过参数寻优找到最优的核函数宽度和惩罚因子。文献^[4]的核函数宽度取为 1.25,惩罚因子取为 50。实验结果如表 1 所示。这里的训练时间是指 SMO 算法求解拉格朗日乘子的时间,不包含去除噪音的预处理时间。

表1 iris 数据集分类结果对比

	训练时间/s	支持向量个数	测试精度/%
文献 ^[4] 算法	4.31	54	89.15
本文算法	3.65	45	98.96

由表中实验数据可以看出,标准的超球支持向量机算法训练时间与本文算法相当,但是由于受噪音数据影响其分类器性能明显低于本文算法。

第三组数据选择一个目标图像分类数据,也是比较本文算法与文献^[4]算法的分类性能。目标图像共分为 6 类,包括花、树、草、动物、人、房屋几个类

别,提取轮廓、小波、纹理等 25 维特征,采集的样本数量为 1 000 个,随机选取 85% 的样本作为训练数据,其余 15% 作为测试。因此采集的图像数据本身就包含噪音数据,所以这里不再人为加入噪音数据。文献^[4]的算法核函数宽度取为 1.56,惩罚因子取为 100。实验重复 5 次,实验结果取其平均值。实验数据如表 2 所示。

表2 目标图像分类结果对比

	训练时间/s	支持向量个数	测试精度/%
文献 ^[6] 算法	556	289	80.14
本文算法	425	212	89.77

由表 2 中的数据可知,对相同的数据集,由于本文算法对噪音数据进行了预处理,因此支持向量机数量小于标准的超球支持向量机算法,训练时间较短而且测试精度较高。

5 结束语

本文给出了加权因子的计算方法,用于区别对待每一个训练样本,并加快 SMO 求解算法的收敛性。在此基础上推导了加权超球支持向量机算法的目标函数及对偶目标函数,并分析了加权因子动态调整拉格朗日乘子上界的功能。本文分析了噪音数据/离群数据对分类结果的影响,给出了预处理算法,实验结果验证了算法的有效性及可行性。虽然本文算法在数据预处理、加权因子的计算、分类精度上均有所改进,但是对噪音数据的判断、考虑类别不平衡信息的加权因子计算等仍需深入研究。

参考文献:

- [1] 徐冉冉, 据昊霖, 李朝锋. 非平衡二叉树主动学习支持向量机[J]. 微电子学与计算机, 2013, 30(5): 55-58.
- [2] 于静, 王辉. 改进支持向量机在网络入侵检测中的应用[J]. 微电子学与计算机, 2012, 29(3): 10-13.
- [3] 彭辉. 小波变换和支持向量机相融合的 ECG 身份识别算法[J]. 微电子学与计算机, 2013, 30(3): 152-155.
- [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.
- [5] Shin H, Eom D, Kim S. One-class support vector machines - an application in machine fault detection and classification[J]. Computers & Industrial Engineering, 2005, 48(2): 395-408.

(下转第 28 页)

- 2007, 42(7): 1472-1480.
- [4] Lu L, Chen J, Yuan L, et al. An 18-mW 1.175 - 2-GHz frequency synthesizer with constant bandwidth for DVB-T Tuners [J]. IEEE Trans. Microw. Theory Tech., 2009, 57(4): 928-937.
- [5] Kim J, Shin J, Kim S, et al. A Wide-Band CMOS LC VCO with linearized coarse tuning characteristics [J]. IEEE Trans. Circuits Syst. II, Exp. Briefs, 2008, 55(5): 399-403.
- [6] Mira J, Divel T, Ramet S, et al. Distributed MOS varactor biasing for VCO gain equalization in 0.13 μm CMOS technology [J]. IEEE RFIC Symp. Dig., 2004(6): 131-134.
- [7] Moon Y J, Roh Y S, Jeong C Y, et al. A 4.39-5.26GHz LC-tank CMOS voltage-controlled oscillator with small VCO-gain variation [J]. IEEE Microw. Wireless Compon. Lett., 2009, 19(8): 524-526.
- [8] 谢连波, 桑红石, 方海涛, 等. 低功耗 CMOS 差分环形压控振荡器设计 [J]. 微电子学与计算机, 2013, 30(5): 104-107.
- [9] 吕荫学, 刘梦新, 罗家俊, 等. 一种 1GHz 多频带压控振荡器的设计 [J]. 微电子学与计算机, 2013, 30(2): 1-4.
- [10] Kim J, Plouchart J, Zamdmer N, et al. A 44GHz differentially tuned VCO with 4GHz tuning range in 0.12 μm SOI CMOS [C] // ISSCC. China: Beijing, 2005, 22: 416-607.
- [11] Deng Z, Niknejad A M. A 4-port-inductor-based VCO coupling method for phase noise reduction [J]. IEEE J. Solid-State Circuits, 2011, 46(8): 1772-1781.
- [12] Hsu M, Chiu C. A low-power 10-GHz current reused VCO using negative resistance enhancement technique [C] // Asia Pacific Microwave Conf. Singapore, 2010: 2276-2279.
- [13] Wang T, Li C. A 0.4-V 1.08mW 12 GHz high performance VCO in 0.18- μm CMOS [C] // Radio and Wireless Symp. USA, Santa Clara, CA, 2012: 207-210.

作者简介:



陈 华 男, (1984-), 硕士研究生. 研究方向为模拟/射频集成电路设计. E-mail: chenhu@imc.ac.cn.

郭桂良 男, (1981-), 博士, 副研究员. 研究方向为模拟电路及射频前端电路设计.

阎跃鹏 男, (1963-), 研究员, 博士生导师. 研究方向为通信与导航系统、模拟射频集成电路设计.

(上接第 23 页)

- [6] 刘爽, 史国友. 基于加权超球支持向量机算法的超文本分类研究 [J]. 大连海事大学学报, 2009, 35(1): 71-74.
- [7] Guo L, Xiao H T, Fu Q. New support vector data description fuzzy recognition method [J]. Journal of System Simulation, 2009, 21(7): 1882-1886.
- [8] Lee G H, Taur J S, Tao C W. A robust fuzzy support vector machine for two-class pattern recognition [J]. International Journal of Fuzzy Systems, 2006, 8(2): 76-86.
- [9] Blake C L, Merz C J. UCI repository of machine learning databases [EB/OL]. (1988-07-15). [2014-03-25]. <http://www.ics.uci.edu/mllearn/>.

作者简介:

刘 爽 女, (1977-), 博士, 副教授. 研究方向为智能信息处理. E-mail: dlnuliushuang@126.com.

陈 鹏 男, (1977-), 硕士, 副教授. 研究方向为信息安全.

李锡祚 男, (1963-), 博士, 教授, 硕士生导师. 研究方向为电子商务安全.