文章编号: 1001-0920(2009)04-0494-05

基于 k-最近邻的支持向量预选取方法

韩德强, 韩崇昭, 杨　艺
(西安交通大学 电信学院, 西安 710049)

摘　要: 在所有的训练样本中只有支持向量(SVs)能对支持向量机分界面优化结果产生显著影响. 基于 k-最近邻规则,提出了一种训练样本的预选取方法. 针对一些典型人工数据集、公用基准数据集以及 TM 遥感数据的实验结果表明,该方法能够有效减少训练样本数目,显著加快学习速度,并保证理想的分类精度.

关键词: 支持向量机; 样本预选取; k-最近邻; 模式分类
中图分类号: TP181　　**文献标识码**: A

Approach for pre-extracting support vectors based on k-NN

HAN De-qiang, HAN Chong-zhao, YANG Yi
(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China. Correspondent: HAN De-qiang, E-mail: digital_king@263.net)

Abstract: In support vector machine (SVM) only support vectors (SVs) have the significant influence on the optimization result. An approach for pre-extracting SVs based on k-NN is proposed. The experimental results based on some artificial datasets, some real-world datasets and TM remote sensing dataset show that the approach proposed can effectively reduce the size of training sets and accerlerate the learning speed. At same time, the classification accuracies are ensured.

Key words: SVM; Sample pre-extracting; k-NN; Pattern classification

1 引　　言

由 Vapnic[1] 提出的支持向量机(SVM)具有强大的非线性处理能力和良好的推广泛化能力. 标准的 SVM 学习问题可以归结成一个带约束条件的凸二次规划(QP)问题[2]. SVM 的优化问题仅与支持向量(SVs)有关,SVs 决定了最优分类面. 在诸多领域,如信号处理、模式识别(如人脸识别)、函数回归等[3-5],SVM 均得到了越来越广泛的应用. 目前 SVM 已经成为机器学习领域的研究热点之一.

SVM 在实际应用中也存在一些棘手的问题. 优化时采用全部的训练样本,遇到大规模数据集时,凸二次规划求解问题在求解时间和存储空间上都会遭遇瓶颈. 标准的 SVM 算法很难应用于大规模数据场合. 针对这些问题,出现了许多改进的方法,其中主要包括两个方面的思路: 1) 对优化算法本身进行改进,如最小二乘支持向量机(LS-SVM)[6],序贯最小化优化(SMO)[7],增量 SVM[8]等; 2) 对训练样本进行预处理,选取出最可能成为 SVs 部分训练样本,从而达到减少存储空间,加快求解速度的目的. 在文献[9]中,作者利用聚类分析,除去那些仅含有单一类别样本的聚类团,将剩余聚类团作为训练样本. 该方法虽然行之有效,但聚类团内部类别单一的要求过于严格,有可能误杀支持向量,易受噪声及孤立点影响,并且其性能与所选择的聚类方法密切相关. 其他一些方法包括中心距离比值法[10]、自适应投影算法[11]、两凸包相对边界向量法[12]等,虽然都能有效减少训练样本和加快训练速度,但同样存在易受噪声点及孤立点影响等问题,且在对训练样本进行预处理之前,要对训练样本集是否线性可分进行判断. 文献[13,14]提出了类似的方法,利用样本邻域信息作为判断依据进行训练样本预选取,能克服孤立点及噪声的影响,且无需判断样本集是否线性可分,但邻域信息提取的实现较为繁琐,其性能受到参数选择等问题的制约.

本文提出的方法基于 k-最近邻(k-NN)规则进行样本邻域信息提取,并以此作为样本预选取的判

收稿日期: 2008-03-18; **修回日期**: 2008-06-16.
基金项目: 国家自然科学基金项目(60574033); 国家 973 计划项目(2007CB311006).
作者简介: 韩德强(1980—),男,河南孟州人,博士生,从事信息融合、模式分类等研究; 韩崇昭(1943—),男,陕西乾县人,教授,博士生导师,从事信息融合、复杂系统调度与控制等研究.

断依据. 该方法抗噪声点及孤立点影响的能力强,无需事先判断样本集的线性可分性,实现简单方便,能在有效减少训练样本数目、提高训练速度的同时,保证较高分类精度.

2　SVM 概要

线性 SVM 的目的在于依据风险最小化准则构造一个最优分类超平面,使得分类间隔(余度)最大,如图 1 所示.

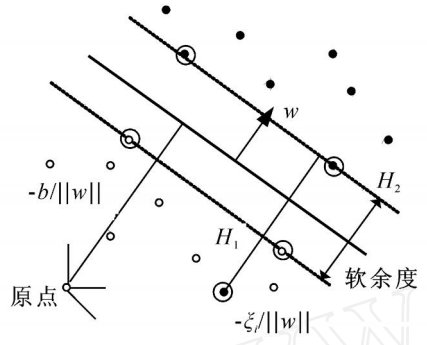


图 1　线性支持向量机最优分类面

SVM 的模型可描述如下:

$$\min_{w}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i,$$

$$\text{s.t. } y_i(x_i^{\mathrm{T}}w+b)-1+\xi_i\geq 0,$$

$$i=1,2,\cdots,l. \qquad (1)$$

最优分类超平面可根据式(1)确定. 其中:C 是对分类错误的惩罚因子,ξ 是松弛项,b 是偏移量,w 是用于决定最优分类面方向的权向量. 求解过程可由 Lagrange 优化方法实现,即

$$\max_{\alpha}L_D=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j,$$

$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i=0,$$

$$0\leq\alpha_i\leq C,\ i=1,2,\cdots,l, \qquad (2)$$

其中 α_i 是 Lagrange 乘子. 求解可得最优分类面方程为

$$f(x)=\mathrm{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i(x_i^{\mathrm{T}}x)+b^*\right\}, \qquad (3)$$

其中 α_i^* 和 b^* 为最优分类面对应的参数.

对于非线性分类问题,统计学习理论采用如下的方法:通过某种事先选择的非线性映射 $\Phi: x\rightarrow\Phi(x)$ 将样本空间映射到一个高维空间上. 定义一个"核函数"K,使得

$$K(x_i,x_j)=\langle\Phi(x_i),\Phi(x_j)\rangle. \qquad (4)$$

最终,超分类面也成为

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*\right). \qquad (5)$$

SVM 的 KKT(Karush-kuhn-tucker)条件[14]为:

1) $\alpha_i=0\Rightarrow y_i f_i=1$,普通样本;

2) $0<\alpha_i<C\Leftrightarrow y_i f_i=1$,边界处的 SV;

3) $\alpha_i=C\Leftrightarrow y_i f_i\leq 1$,间隔内的 SV.

对于约束优化问题,KKT 条件无论从理论上或实践上都是非常重要的. 由上述 KKT 条件可知,SVM 解将所有训练样本分为 3 个部分. 显而易见,仅有那些位于边界附近的支持向量对应的 Lagrange 乘子不为 0. 如果在训练 SVM 时,只采用这些支持向量应该能够取得与采用全部样本几乎相同的训练效果. 即除了支持向量之外的所有训练样本几乎对训练效果不产生任何影响. 如果能设法在训练样本中提取出有可能成为支持向量的样本,这样在保证分类性能的同时,便能够大大减少 SVM 训练学习的复杂度. 本文正是根据此思路提出一种基于 k-NN 规则的 SVM 分类预处理方法.

3　基于 k- 最近邻法的样本预处理方法

k- 最近邻法(k-NN)是一种非参数的分类方法,简单易行. 给定测试样本 x_q,在训练样本中寻找其对应的 k 个距离最近的样本. 针对不同的应用场合,距离的定义方式可以有多种形式,其中最为常用的是欧氏距离. 选出在 k 个最近邻样本中数量最多的类别,可以完成对测试样本的分类. 同样可以针对训练样本,利用 k-NN 规则得到该训练样本附近的类别分布密度. 对二分类问题,假设类别标号分别为 $\{0,1\}$. 训练样本的 k 个距离最近的样本中,属于类 0 的样本个数计为 k_0,属于类 1 的样本个数计为 k_1. 类 0 分布密度为 $\rho_0=k_0/k$,类 1 分布密度为 $\rho_1=k_1/k$,显然有 $\rho_0+\rho_1=1$. 选取合适的 k 值(一般可取训练样本总数的 1/10),获取 ρ_0 和 ρ_1 这样的邻域信息. 如果某样本远离边界,则其周围的样本隶属于同类的可能性更大,ρ_0 和 ρ_1 间的差异会相对较为明显;如果某样本位于边界附近,则其周围同时出现两类样本的可能性较大,ρ_0 和 ρ_1 会相对较为接近. 举例说明:如图 2 所示,区域 Ⅰ和区域 Ⅱ的中心样本点 $x_{\mathrm{Ⅰ}}$,$x_{\mathrm{Ⅱ}}$ 均远离分类边界. $x_{\mathrm{Ⅰ}}$ 及 $x_{\mathrm{Ⅱ}}$ 各自周围的 5 个最近邻样本点均属于同一类别. 在区域 Ⅰ中,$\rho_0=0$,$\rho_1=1$;Ⅱ中,$\rho_0=1$,$\rho_1=0$. 区域 Ⅰ和 Ⅱ中,ρ_0 和 ρ_1 差异均十分明显;区域 Ⅲ中的中心样本点 $x_{\mathrm{Ⅲ}}$ 在分类边界附近,其周围的 5 个最近邻样本点有 3 个属于类 0,2 个属于类 1,即区域 Ⅲ中 $\rho_0=0.6$,$\rho_1=0.4$,ρ_0 和 ρ_1 相对较为接近.

依据这一思路,选取合适的参数和阈值,训练样本数据预处理可实现如下:

针对每个训练样本 $x_i(i=1,2,\cdots,N)$,N 为训练样本总数,求取 x_i 的 k 个最近邻,求出相应的 ρ_0 及

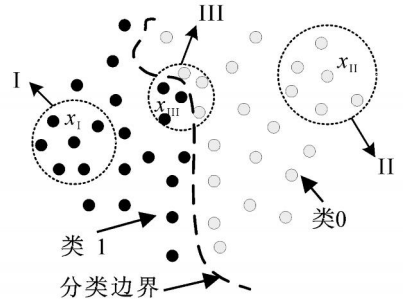


图 2 不同区域的样本邻域信息

ρ_i^j,

1) 如果 x_i 属于类 0,且 $\rho_i^j/(\rho_i^0+\rho_i^j) < \varepsilon$(选取参数 $0.5 \le \varepsilon \le 1$),则标记该训练样本;

2) 如果 x_i 属于类 1,且 $\rho_i^j/(\rho_i^0+\rho_i^j) < \varepsilon$(选取参数 $0.5 \le \varepsilon \le 1$),则标记该训练样本.

孤立点往往也会成为支持向量. 对于孤立点也同样能找到其 k 个最近邻,故无法根据样本 k 个近邻的类分布密度来判别其是否为孤立点. 所以,需要设法找出孤立点. 在获取每个样本 x_i 的 k 个近邻的同时,可得到 x_i 到每个近邻的距离和 d_i. 对所有 d_i ($i=1,2,\cdots,N$),求取其均值,记为 $\bar{d}$. 如果 $\bar{d}/d_i \le \delta$(δ 可选取为一个接近于 0 的小数,如 0.2),则训练样本 x_i 将被视为孤立点. 将所有发现的孤立点及步骤 1) 和 2) 中所有被标记点集求并集,即可获得用于最终训练的样本集.

对于采用非线性核函数 K 的 SVM 而言,欧氏距离定义如下:

$$d^{\Phi}(x_i,x_j) = \|\Phi(x_i),\Phi(x_j)\|_2 = \sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)}. \quad (6)$$

4 实验及讨论

本文实验针对一些典型的人工数据集及公用基准数据集,验证文中提出的训练样本预选取方法的有效性. 实验环境为:CPU 为 AMD Athlon 4000 + Dual Core, 2.11 GHz,内存为 2 GB DDR II,操作系统为 Windows XP -SP2, 软件平台为 Matlab R2007a.

4.1 人工数据集

在实验中采用 3 种人工数据集,分别是双螺旋线、Ripley 数据集和正态分布二分类数据集.

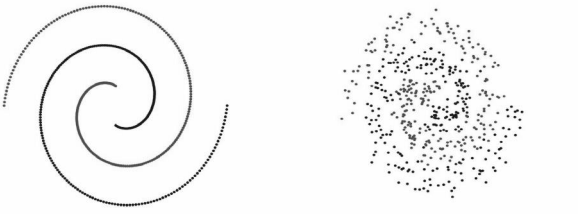
4.1.1 双螺旋线问题

双螺旋线是一个经典的二分类问题,数据可由下式产生[15]:

$$\text{spiral1}:\begin{cases} x = A\theta\cos(\theta), \\ y = A\theta\sin(\theta); \end{cases} \quad (7)$$

$$\text{spiral2}:\begin{cases} x = A\theta\cos(\theta+\pi), \\ y = A\theta\sin(\theta+\pi). \end{cases} \quad (8)$$

其中:参数 $A=3$, $\pi/2 \le \theta \le 3\pi$. 假定每个类别对应的样本在这两条曲线上均匀分布. 实验中所使用的双螺旋线增加了均值为零、方差为 1.5 的高斯噪声,如图 3 所示.



(a) 未加噪声污染的双螺旋线 (b) 高斯噪声污染的双螺旋线

图 3 两类双螺旋线

共产生 500 个样本点(每类 250 个). 每次实验,从每个类别中随机抽取 125 个样本作为训练样本(共 250 个),剩余样本作为测试样本(共 250 个). 实验共重复进行 10 次. 本实验采用 LS-SVM,核采用径向基函数

$$K(x_i,x_j) = \exp(-\|x_i - x_j\|/\sigma^2). \quad (9)$$

优选后的 LS-SVM 参数设置:惩罚系数为 2.5,径向基核参数为 $\sigma^2 = 0.5$. 依据本文提出的样本预选取方法,参数 $k=25$, $\varepsilon=0.75$, $\delta=0.2$. 测试样本集上的实验结果如表 1 所示.

表 1 双螺旋线数据集测试结果(平均值)

方 法	分类精度/ %	SV 数量	收敛时间/ s
LS-SVM	96.84	181.3	0.2564
本文 + LS-SVM	96.32	87.9	0.0269

4.1.2 Ripley 数据集

Ripley 数据集[16] 包含分属于两个类别的数据. 每个类别的数据都是基于混合高斯分布产生的. 训练样本集有 250 个样本(每类各 125 个样本),测试样本集有 1000 个样本(每类各 500 个样本). 本实验中使用 LS-SVM,采用线性核函数. 优选后的 SVM 参数设置:惩罚系数为 2. 依据本文提出的样本预选取方法,参数 $k=25$, $\varepsilon=0.75$, $\delta=0.2$. 测试样本集上的实验结果如表 2 所示.

表 2 Ripley 数据集测试结果

方 法	分类精度/ %	SV 数量	收敛时间/ s
LS-SVM	89.20	175	0.0175
本文 + LS-SVM	88.90	60	0.0038

4.1.3 正态分布二分类数据集

该数据集共两类数据,每类数据均服从正态分布. 每个样本都是二维的 (x,y), x 服从高斯分布, y 服从均匀分布,数据可依据下式生成:

$$p_{\text{类1}}(x,y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma(b-a)}\exp\left[-\dfrac{1}{2}\left(\dfrac{x}{\sigma}\right)^2\right], & a \le y \le b; \\ 0, \text{otherwise}; \end{cases}$$

$$p_{类2}(x,y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma(b-a)}\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right], & a \le y \le b; \\ 0, \text{ otherwise.} \end{cases} \tag{10}$$

这里选取 $\mu = 2.5$, $\sigma = 1$, $b - a = 3$. 本实验中,共生成 500 个样本点(每类各 250 个). 从每类中随机选取 125 个作为训练样本,其余留作测试样本. 实验共重复进行 10 次. 本实验使用 LS-SVM,采用线性核函数优选后的 LS-SVM 参数设置:惩罚系数为 3. 依据本文提出的样本预选取方法,参数 $k = 25$, $\varepsilon = 0.75$, $\delta = 0.2$. 测试样本集上的实验结果如表 3 所示.

表 3　正态二分类数据集测试结果(平均值)

方　法	分类精度 / %	SV 数量	收敛时间 / s
LS-SVM	90.28	171	0.0165
本文 + LS-SVM	90.40	53.6	0.0024

4.2　公用基准数据集

采用 UCI 数据库中的 Pima Indians Diabetes Database 来验证文中提出的方法. 在 Pima 数据集中,共有两个类别,分别包含 500 个及 268 个样本. 每个样本对应的特征维度为 8. 在进行实验之前,先对所有样本的各个特征维度进行如下的归一化处理:

$$a_i = (v_i - \min v_i)/(\max v_i - \min v_i), \tag{11}$$

其中 v_i 为属性 i ($i = 1, 2, \cdots, 8$) 的真实值,取最大及最小操作在所有样本上进行. 从每类随机抽取约 1/3 的样本作为训练样本,其余留作测试样本. 实验重复进行 10 次. 本实验中使用 LS-SVM,采用线性核函数. 优选后的 LS-SVM 参数设置:惩罚系数为 5. 依据本文提出的样本预选取方法,参数 $k = 38$, $\varepsilon = 0.75$, $\delta = 0.2$. 测试样本集上的实验结果如表 4 所示.

表 4　Pima 数据集测试结果(平均值)

方　法	分类精度 / %	SV 数量	收敛时间 / s
LS-SVM	76.86	178.6	0.0688
本文 + LS-SVM	75.32	59.2	0.0058

4.3　TM 遥感图像分类

实验数据为北京某地区 Landsat TM 影像,每个样本点包括 6 个波段的 TM 影像数据. 选取所有待分类地物中的两类(稻田和玉米田),每类各有 1600 个样本点. 从每类中随机选取 800 个样本作为训练样本,其余留作测试样本,实验重复进行 10 次. 采用线性核函数,优选后的 LS-SVM 参数设置:惩罚系数为 5. 依据本文提出的样本预选取方法,参数 $k = 160$, $\varepsilon = 2/3$, $\delta = 0.2$. 测试样本集上的实验结果如表 5 所示.

表 5　TM 影像数据集测试结果(平均值)

方　法	分类精度 / %	SV 数量	收敛时间 / s
LS-SVM	94.81	1101.6	6.3542
本文 + LS-SVM	94.63	205.2	0.2077

从以上所有实验结果可知,基于本文的样本预选取方法,可以在几乎不损失分类精度的情况下,显著缩减 SVM 的训练收敛时间,从而为 SVM 应用于大规模数据集场合创造了条件.

5　结　　论

本文为解决 SVM 在大数据量情况下的训练收敛慢等问题,提出了基于 k-最近邻规则对训练样本进行预选取的数据处理方法. 实验结果表明,本方法合理有效,能够筛选出有可能成为支持向量的训练样本点及孤立点. 在有效缩减收敛时间的同时,保证分类精度.

需要指出的是,本文方法的有效性,将在某种程度上受到参数选取的制约. 文中给出了一些经验性的参数选取方法. 未来会致力于研究和提出较为完善的参数优化选取方法,以及提出非参数的样本预处理方法,从而更为便捷地实现对 SVM 应用领域的拓展.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[2] Vapnik V N. An overview of statistical learning theory [J]. IEEE Trans on Neural Networks, 1999, 10(5): 988-999.

[3] 张宝昌, 陈熙霖, 山世光, 等. 基于支持向量的 Kernel 判别分析[J]. 计算机学报, 2006, 29(12): 2143-2150. (Zhang B C, Chen X L, Shan S G, et al. Kernel discriminant analysis based on support vectors [J]. Chinese J of Computers, 2006, 29(12): 2143-2150.)

[4] Smola A, Scholchopf B. On a kernel-based method for pattern recognition, regression, approximation and operator inversion[J]. Algorithmica, 1998, 22(1): 211-231.

[5] Suykens J A K, Van Gestel T, Vandewalle J, et al. A support vector machine formulation to PCA analysis and its kernel version[J]. IEEE Trans on Neural Networks, 2003, 14(2): 447-450.

[6] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers [J]. Neural Process Letter, 1999, 9(3): 293-300.

[7] Platt J C. Fast training of support vector machines using sequential minimal optimization[C]. Advances in Kernel

Methods-Support Vector Learning. Cambridge: MIT Press, 1998: 185-208.

[8] 李凯, 黄厚宽. 支持向量机增量学习算法研究[J]. 北方交通大学学报, 2003, 27(5): 34-37.
(Li K, Huang H K. Research on incremental learning algorithm of support vector machine[J]. J of Northern Jiaotong University, 2003, 27(5): 34-37.)

[9] Almeida M B, Braga A P, Braga J P. SVM-KM: Speeding SVMs learning with a priori cluster selection and *k*-means[C]. Proc of the 6th Brazilian Symposium on Neural Networks. Brazil, 2000: 162-167.

[10] 焦李成, 张莉, 周伟达. 支撑矢量预选取的中心距离比值法[J]. 电子学报, 2001, 29(3): 383-386.
(Jiao L C, Zhang L, Zhou W D. Pre-extracting support vectors for support vector machine[J]. Acta Electronica Sinica, 2001, 29(3): 383-386.)

[11] 丁爱玲, 刘芳, 曹伟. 支撑矢量预选取的自适应投影算法[J]. 计算机工程与应用, 2002, 38(19): 116-118.
(Ding A L, Liu F, Cao W. Adaptive projective algorithm for selecting support vector beforehand[J]. Computer Engineering and Applications, 2002, 38(19): 116-118.)

[12] 安金龙, 王正欧. 预抽取支持向量机的支持向量[J]. 计算机工程, 2004, 20(30): 10-11, 48.
(An J L, Wang Z O. Pre-extracting support vectors for support vector machine[J]. Computer Engineering, 2004, 20(30): 10-11, 48.)

[13] 廖东平, 魏玺章, 黎湘, 等. 一种新的支持向量机快速训练算法[J]. 系统工程与电子技术, 2007, 29(11): 1954-1957.
(Liao D P, Wei X Z, Li X, et al. New fast training algorithm of support vector machine[J]. Systems Engineering and Electronics, 2007, 29(11): 1954-1957.)

[14] Meng D Y, Xu Z B, Jing W F. A more efficient preprocessing method for support vector classification[C]. Proc of Int Conf on Neural Networks and Brain 2005. Beijing: IEEE Press, 2005: 1173-1177.

[15] Du H, Chen Y Q. Rectified nearest feature line segment for pattern classification[J]. Pattern Recognition, 2007, 40(5): 1486-1497.

[16] Ripley B D. Pattern recognition and neural networks[M]. Cambridge: Cambridge University Press, 1996.

(上接第 493 页)

参考文献(References)

[1] Alanis A Y, Sanchez E N, Loukianov A G. Discrete-time adaptive backstepping nonlinear control via high-order neural networks[J]. IEEE Trans on Neural Networks, 2007, 18(4): 1185-1195.

[2] Yeh P C, Kokotovic P V. Adaptive control of a class of nonlinear discrete-time systems[J]. Int J of Control, 1995, 62(2): 303-324.

[3] Zhang Y, Wen C Y, Soh Y C. Discrete-time robust backstepping adaptive control for nonlinear time-varying systems[J]. IEEE Trans on Automatic Control, 2000, 45(9): 1749-1755.

[4] Zhang Y, Wen C Y, Soh Y C. Robust adaptive control of nonlinear discrete-time systems by backstepping without overparameterization[J]. Automatica, 2001, 37(4): 551-558.

[5] Ge S S, Li G Y, Lee T H. Adaptive NN control for a class of strict-feedback discrete-time nonlinear systems[J]. Automatica, 2003, 39(5): 807-819.

[6] Ge S S, Li G Y, Lee T H. Correction to "daptive NN control for a class of strict-feedback discrete-time nonlinear systems"[J]. Automatica, 2008, 44(7): 1930-1931.

[7] Ge S S, Zhang J, Lee T H. Adaptive neural networks control for a class of MIMO nonlinear systems with disturbances in discrete-time[J]. IEEE Trans on System, Man and Cybernetics, 2004, 34(4): 1630-1645.

[8] Zhang J, Ge S S, Lee T H. Output feedback control of a class of discrete MIMO nonlinear systems with triangular form inputs[J]. IEEE Trans on Neural Networks, 2005, 16(6): 1491-1503.

[9] He P, Jagannathan S. Discrete-time neural network output feedback control of nonlinear systems in non-strict feedback form[C]. Proc of the 2004 American Control Conf. Boston, 2004: 2439-2444.

[10] Lin Z, Saberi A. Robust semi-global stabilization of minimum-phase input-output linearizable systems via partial state and output feedback[J]. IEEE Trans on Automatic Control, 1995, 40(6): 1029-1041.

[11] Ge S S, Zhang J, Lee T H. Adaptive MNN control for a class of non-affine NARMAX systems with disturbance[J]. Systems and Control Letters, 2004, 53(1): 1-12.