不等距超球体支持向量机

张慧敏[1,2],柴 毅[1]
ZHANG Huimin[1,2],CHAI Yi[1]
1.重庆大学 自动化学院,重庆 400044
2.重庆电子工程职业学院 通信系,重庆 401331
1.College of Automation,Chongqing University,Chongqing 400044,China
2.Department of Communication,Chongqing College of Electronic Engineering,Chongqing 401331,China

ZHANG Huimin,CHAI Yi.Non-equidistant margin hypersphere SVM.Computer Engineering and Applications,2011,47(11):19-22.

Abstract: For different kinds of samples,the following cases are often confronted:The scopes of sample-distribution differ largely,the degrees of harm for misjudgment of sample class are different,or,the numbers of different kinds of samples are various,pointed on which,SVM(NMS-SVM) based on the non-equidistant margin hypersphere is proposed.Aim at optimizing the maximal margin,SVM introduces distance ratio parameter and adjusts the distance between the optimized classing surface and the two classes.Through the simulation experiment of classification of sample collection by UCI corpus,the validity of SVM is proved by comparing the precision of classification with the normal hypersphere arithmetic as well as the maximal margin hypersphere arithmetic.

Key words: Support Vector Machine(SVM);hypersphere;non-equidistant margin;maximal margin

摘 要:针对现实中经常遇到的各类样本分布范围相差很多、将各类样本误判的危害程度不同、或者各类样本数量差异悬殊等情况,提出了一种基于不等距超球体的SVM(NMS-SVM)算法。该算法以最大间隔为优化目标建立分类模型,同时引入距离比例参数λ,调整最优分类面到两类之间的距离。通过UCI数据库中数据集的分类仿真实验,比较了该算法与普通超球体算法以及最大间隔超球体算法的分类精度,证明了该算法的有效性。

关键词:支持向量机;超球体;不等距;最大间隔
DOI:10.3778/j.issn.1002-8331.2011.11.006 文章编号:1002-8331(2011)11-0019-04 文献标识码:A 中图分类号:TP18

1 引言

传统的支持向量机都是基于超平面的[1],虽然基于超平面的支持向量机在很多领域中都得到了广泛的使用,但也存在一定的局限性。Tax和Duin提出了一种基于超球体的支持向量机数据描述(SVDD)[2]。其主要思想是在高维空间计算包含样本映射的最小超球体,并权衡超球体半径和它所覆盖的样本数。在此基础上,文献[3]提出了一种最大间隔最小体积球形支持向量机算法,它是以最大化R_2和最小化R_1为优化目标的,这是个间接过程。本文将采用直接最大化间隔方式来建立优化方程。

Vapnik定义的最优分界面都是等距的,这对两类分布相当的样本效果是好的。但实际中经常遇到的分类数据往往是非平衡样本,少数类别的数据有可能有很大的分类代价,如各类样本分布范围相差很多、将各类样本误判的危害程度不同、或者各类样本数量差异悬殊等情况。此时分类性能不仅要考虑分类精度,还要考虑分类代价。如果还使用传统的支持向量机最优分类面来分类,将会造成误判率升高或者因为误判而带来比较大的危害。针对这些特殊情况,提出了一种基于不等距的超球体支持向量机(NMS-SVM)算法。最后通过仿真实验证明了该算法的有效性。

2 SVDD算法

假设有训练样本:$x_i\in R^N, i=1,2,\cdots,l$,$l$为样本数。算法要求以$a$为中心,$R$为半径的圆可以包含所有的样本点,并且要求这个圆尽可能的小[4-6]。类似于超平面算法,超球体算法为了解决非线性问题,也采用了核函数思想。同时,为了解决数据存在噪声的问题,引入了松弛变量ξ。

原始的优化问题为:

$$\min R^2+C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.}\begin{cases}(\varphi(x_i)-a)(\varphi(x_i)-a)^{\mathrm{T}}\le R^2+\xi_i\\ \xi_i\ge 0\\ i=1,2,\cdots,l\end{cases}\tag{1}$$

写出式(1)的对偶形式为:

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60974090);教育部博士点基金(No.200806110016)。
作者简介:张慧敏(1981—),女,博士生,讲师,主要研究方向:信息处理、融合与控制等;柴毅(1962—),男,教授,博士生导师。E-mail:zhuomi99@126.com
收稿日期:2010-12-14;修回日期:2011-03-03

$$\max\sum_{i=1}^{l}\alpha_i K(x_i,x_i)-\sum_{i,j=1}^{l}\alpha_i\alpha_j K(x_i,x_j)$$
$$\text{s.t.}\begin{cases}0\le\alpha_i\le C\\ \sum_{i=1}^{l}\alpha_i=1\\ i=1,2,\cdots,l\end{cases}\tag{2}$$

其中 $K(x_i,x_j)=\langle\varphi(x_i)\cdot\varphi(x_j)\rangle$，$K(x_i,x_i)=\langle\varphi(x_i)\cdot\varphi(x_i)\rangle$。通过QP优化方法解这个对偶问题可以得到最终判决函数为：

$$f(x)=\text{sgn}(R^2-\sum_{i,j=1}^{l}\alpha_i\alpha_j K(x_i,x_j)+2\sum_{i=1}^{l}\alpha_i K(x_i,x)-K(x,x))\tag{3}$$

其中 $K(x_i,x_j)=\langle\varphi(x_i)\cdot\varphi(x_j)\rangle$，$K(x_i,x)=\langle\varphi(x_i)\cdot\varphi(x)\rangle$，$K(x,x)=\langle\varphi(x)\cdot\varphi(x)\rangle$。

3 NMS-SVM算法

上章介绍的SVDD算法在构造超球体时，是以构造最小超球体为目标的，而非以最大间隔为目标，这样导致算法的推广性能有限。因此，这里来构造一个以最大间隔为目标、分类面到两类之间距离不相等的超球体SVM算法，通过引入比例系数 λ 可以根据实际需要调整最优超球面与两类之间的距离，如图1所示。“○”表示正类样本，记为 x_i^+；“□”表示负类样本，记为 x_i^-；带阴影的样本代表支持向量。令输入空间经过 $\varphi(x)$ 映射到特征空间后为超球可分，即存在图1所示的两个同心圆S1，S2，S1将正类样本包裹其中，S2将负类样本排除其外，S1半径为 R，S2半径为 $R+\rho$。

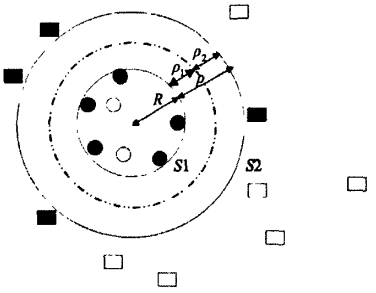


图1 不等距超球体

引入距离比例参数 $\lambda(0<\lambda<1)$，令 $\rho_1=\lambda\rho$，$\rho_1+\rho_2=\rho$，则 $\rho_2=(1-\lambda)\rho$。这里为了提高类内聚类性和增大类间间隔，要求 R 最小，ρ 最大，也等价于 R^2 最小，ρ 最大。优化问题可以描述为：

$$\min R^2-\rho+C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t.}\begin{cases}R^2-(\varphi(x_i^+)-a)^{\mathrm{T}}(\varphi(x_i^+)-a)\ge-\xi_i\\ (\varphi(x_i^-)-a)^{\mathrm{T}}(\varphi(x_i^-)-a)-R^2\ge\rho-\xi_i\\ \xi_i\ge0\\ i=1,2,\cdots,l\end{cases}\tag{4}$$

其中 $y_i=\{-1,+1\}$。采用Lagrange乘子算法求解问题(4)。其Lagrange函数为：

$$L(R,a,\rho,\xi)=R^2-\rho+C\sum_{i=1}^{l}\xi_i-$$
$$\sum_{i=1}^{l}\alpha_i[(R^2-(\varphi(x_i^+)-a)^{\mathrm{T}}(\varphi(x_i^+)-a))+\xi_i]-$$
$$\sum_{i=1}^{l}\beta_i[((\varphi(x_i^-)-a)^{\mathrm{T}}(\varphi(x_i^-)-a)-R^2)-\rho+\xi_i]-\sum_{i=1}^{l}\gamma_i\xi_i\tag{5}$$

对式(5)求导并令其为0，分别得到：

$$\begin{cases}\frac{\partial L}{\partial R}=2R-2\sum_{i=1}^{l}\alpha_i R+2\sum_{i=1}^{l}\beta_i R=0\\ \frac{\partial L}{\partial a}=-2\sum_{i=1}^{l}\alpha_i(\varphi(x_i^+)-a)+2\sum_{i=1}^{l}\beta_i(\varphi(x_i^-)-a)=0\\ \frac{\partial L}{\partial\rho}=-1+\sum_{i=1}^{l}\beta_i=0\\ \frac{\partial L}{\partial\xi}=C-\alpha_i-\beta_i-\gamma_i=0\end{cases}\tag{6}$$

由式(6)可得：

$$a=\sum_{i=1}^{l}\alpha_i\varphi(x_i^+)-\sum_{i=1}^{l}\beta_i\varphi(x_i^-)\tag{7}$$

将式(6)以及式(7)代入式(5)中可以写出其对偶形式为：

$$\max\sum_{i=1}^{l}\alpha_i(\varphi(x_i^+)\cdot\varphi(x_i^+))-\sum_{i=1}^{l}\beta_i(\varphi(x_i^-)\cdot\varphi(x_i^-))-$$
$$\sum_{i=1}^{l}\alpha_i\alpha_j(\varphi(x_i^+)\cdot\varphi(x_j^+))+2\sum_{i=1}^{l}\alpha_i\beta_j(\varphi(x_i^+)\cdot\varphi(x_j^-))-$$
$$\sum_{i=1}^{l}\beta_i\beta_j(\varphi(x_i^-)\cdot\varphi(x_j^-))$$
$$\text{s.t.}\begin{cases}0\le\alpha_i\le C\\ 0\le\beta_i\le C\\ \sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}\beta_i=1\\ \sum_{i=1}^{l}\beta_i=1\\ i=1,2,\cdots,l\end{cases}\tag{8}$$

求解式(8)可以得到其最优解 α_i,β_i，代入式(7)可以得到 a 的最优解。

由KKT条件可知：当 $0<\alpha_i<C$，$0<\beta_i<C$ 时，$\xi=0$，$R^2=\|x_i^+-a\|^2$，$(R+\rho)^2=\|x_i^--a\|^2$，则 $\rho=(\|x_i^--a\|^2-\|x_i^+-a\|^2)^{\frac{1}{2}}$。

引入核函数，令 $K(x_i,x_j)=\langle\varphi(x_i)\cdot\varphi(x_j)\rangle$，则原问题的判决规则为：

令 $R_1=R+\lambda\rho$，对于测试样本，如果 $\|x-a\|\le R_1$，则判断成正类，否则判断成负类。或者最终判决函数可以描述为：

$$f(x)=\text{sgn}[(R+\lambda\rho)^2-K(x,x)+2\sum_{i=1}^{l}\alpha_i K(x_i,x_i^+)-$$
$$2\sum_{i=1}^{l}\beta_i K(x_i,x_i^-)-\sum_{i,j=1}^{l}\alpha_i\alpha_j K(x_i^+,x_j^+)+$$
$$2\sum_{i,j=1}^{l}\alpha_i\beta_j K(x_i^+,x_j^-)+\sum_{i,j=1}^{l}\beta_i\beta_j K(x_i^-,x_j^-)]\tag{9}$$

式中的参数 λ 的选择要视具体的情况而定。

当 $\lambda=0.5$ 时，该算法等同于最大间隔超球体SVM算法。

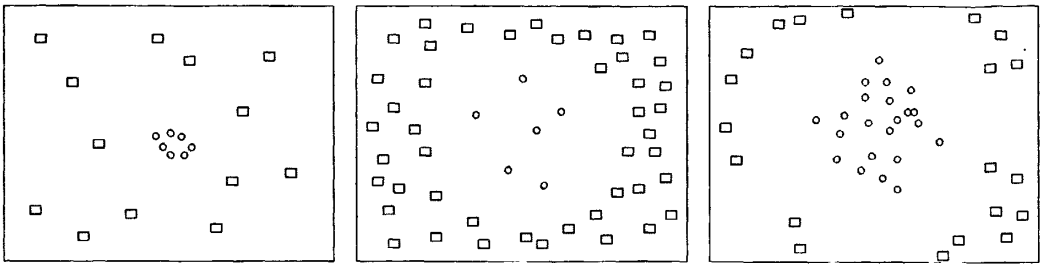
4 NMS-SVM的有效性分析

为了检验NMS-SVM的分类性能，简单起见，假设正负类样本超球可分。

实际中经常遇到以下三种情况：

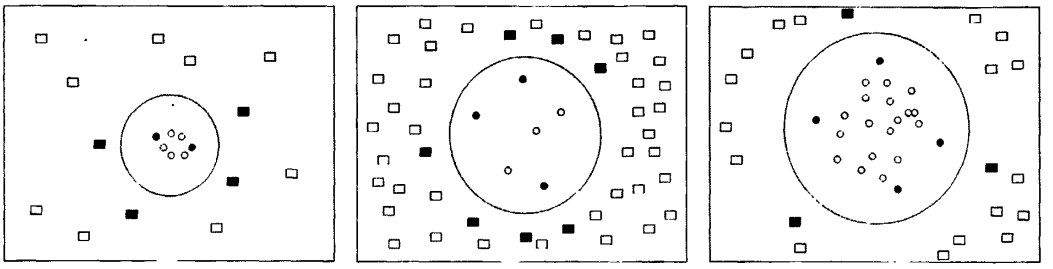
(1)两类样本数据分布范围相差很大(如图2(a))，为提高分类器推广性能，应使超球面尽量靠近分布较集中的那类，从而使分布广的那类获得更大的分类区域。

(2)两类样本数量悬殊很大(如图2(b))，样本数越多越反映真实分布情况，当正负类样本数量悬殊很大时，分类超球面与样本少的那类间隔要大，即分类超球面要向样本多的那类



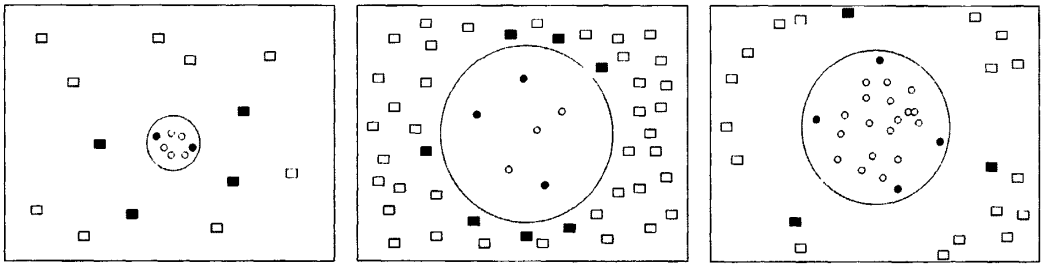
(a)两类样本数据分布范围相差很大　(b)两类样本数量悬殊很大　(c)错分正负类样本

图2　三种不平衡数据的样本分布



(a)正类样本分布集中　(b)正类样本少、负类样本多　(c)两类样本被公平对待

图3　MS-SVM的分割球面(C=100)



(a) $\lambda=0.1$　(b) $\lambda=0.9$　(c) $\lambda=0.2$

图4　NMS-SVM的分割球面

靠近。

(3)错分正负类样本所造成的危害程度不同(如图2(c)),如故障诊断,医疗诊断、网络安全性等情况,把正类样本错判为负类样本所带来的危害远小于把负类样本错判为正类样本。这时,为减少危害,分类超球面要靠近正类样本,远离负类样本。

图3是采用MS-SVM算法建立的分割超球面,"○"表示正类样本,"□"表示负类样本。从图3(a)中可以看出,样本分布较集中的正类获得了大片空白区域,这样会增加将负类样本错判为正类样本的可能性。当正负类样本数量悬殊很大时,样本数越多越反映真实分布情况,而图3(b)中可以看出,数量少的正类样本与数量多的负类样本是"平等"的,这样会增加将负类样本错判为正类样本的概率,从而降低分类器的推广性能。当错分正负类样本所造成的危害程度不同时,超球面在分割正负类样本时就不能平等对待这两类样本,而图3(c)中,两者是被公平对待的,也就是正类样本被误判定成负类样本和负类样本被误判成正类样本的概率是相等的,这在现实中会带来比较严重的危害。

因此,这里重新采用NMS-SVM建立分割超球面,对于第一种情况,可以根据样本的离散度来确定 λ,第二种情况,可以根据正负类样本数量的比值来确定 λ,第三种情况,可以根据错分正负类样本的危害程度来确定 λ,如图4所示。

从图3与图4的效果对比来看,NMS-SVM算法可以克服MS-SVM算法在遇到特殊分类问题时的缺陷,提高分类器的推广能力。

5　实验仿真分析

通过两组实验来说明这种算法与普通超球体SVM算法及最大间隔超球体SVM算法的区别。第一组实验比较了样本分布均衡时,NMS-SVM算法与普通超球体算法的区别;第二组实验分析了样本分布非均衡时,不同 λ 值对NMS-SVM算法精度的影响。

第一组实验数据来自UCI机器学习数据库的经典数据集Iris(http://archive.ics.uci.edu/ml/datasets.html),Iris数据集包含3类样本,每类各有50个样本,每个样本有4个属性。因为本文研究的是二类分类,所以将第一类与第二类合并为正类,第三类为负类,这样就变成了一个二类分类问题。

实验开发与调试工具为Microsoft Visual C++6.0。从各个类别的样本中分别提取60%的样本数作为训练,剩余40%的样本作为测试使用。分类器所选的核函数为高斯径向基核函数:

$$K(x_i,y_i)=\exp\left(-\frac{\|x_i-y_i\|^2}{\sigma^2}\right) \tag{10}$$

采用5-折交叉验证方法确定核函数参数,令C的取值分别为:1,5,10,50,100,200,容许误差$\varepsilon=0.005$。这里因为样本分布均衡,因此取$\lambda=0.5$。表1给出了NMS-SVM算法与普通超球体算法的分类精度。

表1 NMS-SVM算法与普通超球体算法的分类性能比较

C	样本集	普通超球体SVM算法		NMS-SVM算法$\lambda=0.5$	
		分类精度/(%)	训练时间/s	分类精度/(%)	训练时间/s
1	正类	93.7	0.082	94.5	0.092
	负类	78.9		90.1	
5	正类	93.7	0.077	94.6	0.088
	负类	79.0		90.2	
10	正类	93.8	0.072	94.8	0.082
	负类	79.1		90.4	
50	正类	94.0	0.063	94.9	0.078
	负类	79.3		90.8	
100	正类	94.1	0.055	95.1	0.069
	负类	79.5		91.0	
200	正类	94.2	0.047	95.1	0.055
	负类	79.7		91.1	

从表1可以看出,对于正、负类样本,NMS-SVM算法的分类精度都比普通超球体SVM算法的分类精度高。普通超球体SVM算法对负类的分类精度比较低,提出的NMS-SVM算法对负类的分类精度有了很大的提高。

为了验证NMS-SVM算法在处理非均衡数据时的有效性,这里在样本分布数量相差悬殊的情况下进行实验。第二组实验数据来自UCI机器学习数据库(http://archive.ics.uci.edu/ml/datasets.html)的Ablaone数据集,Glass数据集,Breast Cancer数据库(正类样本只取恶性乳腺癌中的30个样本),Yeast数据库,Car数据库。从这些数据集的多类数据中分别选取一类数据作为正类,其他数据则作为负类。表2中列出了本次实验所用的数据。取各个数据集中的60%的样本数作为训练,剩余40%的样本作为测试使用。

表2 实验中所使用的数据

数据集	被选取的正类标号	正类样本数量	负类样本数量
Ablaone	4	57	4 120
Glass	5	13	201
Breast Cancer	-	30	357
Yeast	3	244	1 484
Car	3	69	1 659

为了方便,这里的实验数据都是正样本数相对于负样本数极其稀少的情况。因此进行仿真实验时,分别取$\lambda=0.01$和$\lambda=0.5$进行分析比较。普通的分类评价标准对非均衡数据的分类并不适用,这里采用文献[7]的方法确定分类评价标准。本次实验引入正负查全率(Recall)和v均值方法来评价实验结果:

$$Recall^{+}=\frac{T^{+}}{T^{+}+F^{-}} \tag{11}$$

$$Recall^{-}=\frac{T^{-}}{T^{-}+F^{+}} \tag{12}$$

$$v=\sqrt{Recall^{+}\cdot Recall^{-}} \tag{13}$$

其中T^{+}、T^{-}表示正确分类的正类和负类,F^{+}、F^{-}表示错误分类的正类和负类。$Recall^{+}$、$Recall^{-}$表示正、负类的查全率。

表3显示了不同λ值时NMS-SVM算法的正负查全率。表4显示了不同λ值时算法的v均值及其平均值。

表3 不同λ值时算法的正、负查全率

数据集	NMS-SVM(λ=0.5)		NMS-SVM(λ=0.01)	
	$Recall^{+}$	$Recall^{-}$	$Recall^{+}$	$Recall^{-}$
Ablaone	0	1	0.831	0.855
Glass	0.827	1	0.912	0.901
Breast Cancer	0.812	1	0.932	0.941
Yeast	0.856	1	0.928	0.905
Car	0	1	0.925	0.920

表4 不同λ值时算法的v均值及其平均值

数据集	NMS-SVM(λ=0.5)	NMS-SVM(λ=0.01)
Ablaone	0	0.843
Glass	0.909	0.906
Breast Cancer	0.901	0.936
Yeast	0.925	0.916
Car	0	0.922
平均值	0.547	0.905

从表3可以看出$\lambda=0.5$时,NMS-SVM算法对负类(数量多)样本的查全率很高,但对正类(数量少)样本的查全率比较低;$\lambda=0.01$时,NMS-SVM算法对正类样本的查全率有比较明显的提高。表4显示了$\lambda=0.01$时NMS-SVM算法的v均值平均值明显比$\lambda=0.5$时NMS-SVM算法的v均值平均值高。

从上面分析可看出在数据分布不均衡时,不同λ对NMS-SVM算法的分类精度影响是不同的。具体的λ值要根据实际样本的分类情况或者错分处罚机制情况来决定。

6 结论

基于SVDD的最大间隔超球体SVM算法既能提高类内聚类性,也能保证正类与负类之间的距离最大,从而提高分类器性能。在此基础上,针对现实中经常遇到的各类样本分布范围相差很多、将各类样本误判的危害程度不同、或者各类样本数量差异悬殊等情况,提出一种NMS-SVM算法。最后通过仿真实验,比较了样本均衡分布时NMS-SVM算法与普通超球体算法的区别,以及样本分布非均衡时,不同λ值对NMS-SVM算法精度的影响,从而证明了NMS-SVM算法的有效性。

参考文献:

[1] Cristianini N,Shawe-Taylor O.支持向量机导论[M].英文版.北京:机械工业出版社,2005.

[2] Tax D M J.Support vector data description[J].Machine Learning,2004,54:45-66.

[3] 文传军,詹永照,陈长军.最大间隔最小体积球形支持向量机[J].控制与决策,2010,25(1).

[4] 朱孝开,杨德贵.基于推广能力测度的多类SVDD模式识别方法[J].电子学报,2009,37(3).

[5] Tao Ban.Implementing multi-class classifiers by one-class classification methods[C]//2006 International Joint Conference on Neural Networks.Vancouver:IEEE Press,2006:327-332.

[6] Jin Hongliang,Liu Qingshan,Lu Hanqing.Face detection using one-class based support vectors[C]//Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition,FGR'04.Seoul:IEEE Press,2004.

[7] 姚全珠,田元,王季,等.基于最小二乘支持向量机的非平衡分布数据分类[J].计算机工程与应用,2008,44(4):166-169.

不等距超球体支持向量机

作者： 张慧敏， 柴毅， ZHANG Huimin， CHAI Yi
作者单位： 张慧敏,ZHANG Huimin(重庆大学,自动化学院,重庆,400044;重庆电子工程职业学院,通信系,重庆,401331)， 柴毅,CHAI Yi(重庆大学,自动化学院,重庆,400044)
刊名： 计算机工程与应用 ISTIC PKU
英文刊名： COMPUTER ENGINEERING AND APPLICATIONS
年，卷(期)： 2011,47(11)
被引用次数： 3次

参考文献(7条)

1. Cristianini N;Shawe-Taylor O 支持向量机导论 2005
2. Tax D M J Support vector data description[外文期刊] 2004
3. 文传军;詹永照;陈长军 最大间隔最小体积球形支持向量机[期刊论文]-控制与决策 2010(01)
4. 朱孝开;杨德贵 基于推广能力测度的多类SVDD模式识别方法[期刊论文]-电子学报 2009(03)
5. Tao Ban Implementing multi-class classifiers by one-class classification methods 2006
6. Jin Hongliang;Liu Qingshan;Lu Hanqing Face detection using one-class based support vectors 2004
7. 姚全珠;田元;王季 基于最小二乘支持向量机的非平衡分布数据分类[期刊论文]-计算机工程与应用 2008(04)

本文读者也读过(6条)

1. 武小红.周建江.WU Xiao-hong.ZHOU Jian-jiang 一种基于类中心最大间隔的支持向量机[期刊论文]-信息与控制2007,36(1)
2. 徐世六.武俊齐 A/D转换器发展动态[会议论文]-2001
3. 蔡小萍 论顾准的政治思想[学位论文]2007
4. 方海兰.杨意.黄懿珍.赵晓艺.奚有为.张菊芳 上海几块新建绿地的土壤现状调查[会议论文]-2004
5. 陈森平.陈启买.游才文.彭利宁.CHEN Senping.CHEN Qimai.YOU Caiwen.PENG lining 基于最大间隔的支持向量机特征选取算法研究[期刊论文]-华南师范大学学报（自然科学版）2010(4)
6. 江文涛 中国农业发展银行发展战略研究[学位论文]2005

引证文献(3条)

1. 李新 基于SVM的高校微机室入侵检测技术研究[期刊论文]-科技通报 2012(10)
2. 文传军.柯佳 广义最大间隔球形支持向量机[期刊论文]-计算机工程与应用 2012(29)
3. 李秋林 基于ν-最大间隔超球体支持向量机的非平衡数据分类[期刊论文]-重庆理工大学学报：自然科学 2012(12)