密度加权近似支持向量机

王熙照 崔芳芳 鲁淑霞
(河北大学数学与计算机学院 河北省机器学习与计算智能重点实验室 保定 071002)

摘 要 标准的近似支持向量机(PSVM)用求解正则化最小二乘问题代替了求解二次规划问题,它可以得到一个解析解,从而减少训练时间。但是标准的PSVM没有考虑数据集中正、负样本的分布情况,对所有的样本都赋予了相同的惩罚因子。而在实际问题中,数据集中样本的分布是不平衡的。针对此问题,在PSVM的基础上提出了一种基于密度加权的近似支持向量机(DPSVM),其先计算样本的密度指标,不同的样例有不同的密度信息,因此对不同的样例给予不同的惩罚因子,并将原始优化问题中的惩罚因子由数值变为一个对角矩阵。在UCI数据集上用这种方法进行了实验,并与SVM和PSVM方法进行了比较,结果表明,DPSVM在正负类样本分布不平衡的数据集上有较好的分类性能。

关键词 支持向量机,近似支持向量机,密度加权,不平衡数据
中图法分类号 TP181 **文献标识码** A

Density Weighted Proximal Support Vector Machine

WANG Xi-zhao CUI Fang-fang LU Shu-xia
(Machine Learning and Computational Intelligence Laboratory, College of Mathematics and Computer, Hebei University, Baoding 071002, China)

Abstract The regularized least-squares problem replaces the quadratic programming problem in the standard proximal support vector machines (PSVM). The proximal support vector machines has an analytic solution, so it reduces the training time. But the unbalanced data of positive and negative class is disregarded in the standard proximal support vector machines, and the same penalty factors are assigned to the all training samples. In practical problem, the distribution of positive and negative class is unbalance. Aiming at this problem, a density weighted proximal support vector machines (DPSVM) based on the support vector machines was presented, it is a modified proximal support vector machines algorithm. First calculated the density information of the different data, then according to the density information of the different sample, the different penalty factors were assigned to different training sample which has different density. The penalty values in the original problem of proximal support vector machines were transformed into a diagonal matrix. This method was used on UCI dataset, and compared with support vector machines and proximal support vector Machines methods. The experiment results indicate that the density weighted proximal support vector machines have a better efficiency classified performance in the unbalance data sets of positive and negative samples.

Keywords Support vector machines, Proximal support vector machines, Density weight, Unbalance data sets

1 引言

支持向量机(SVM)[1]是由Vapnik提出的,它建立在统计学习理论中结构风险最小化原则的基础上,是目前运用广泛、功能强大的分类算法,尤其对解决小样本的问题,具有很好的泛化能力。标准SVM的求解属于二次规划问题,对于大样本集的问题,SVM的训练时间较长。为了减少运行时间复杂度[2],许多学者在标准SVM基础上提出了关于它的变形,如LS_SVM (Least Square Support Vector Machines)[3]、Least squares twin support vector machines[4]、近似支持向量机PSVM (Proximal Support Vector Machines)[5]等。

其中标准的PSVM是一种新型的求解分类问题的方法,它不是二次规划问题,而是一个正则最小二乘问题[6],能求出解析解,因此训练速度大大提高。但是PSVM算法在分类时并没有考虑到数据集中正、负样本的分布情况,对每一个样例都给予了相同的惩罚因子,而在实际分类问题中,有很多数据正类和负类的分布是不平衡的,这时就会影响分类效果。针对这个问题,H. G. Chew[7]提出了根据两类样本的数目不同分别给予不同的惩罚因子,不同的样本对分类的贡献也不相同的思想。本文提出一种基于密度加权的PSVM方法,为每一个样本分配不同的隶属度[8,9],首先计算每一个样本的密度信息,根据样本的密度信息对样本的误差项进行密度加权。

到稿日期:2011-02-10 返修日期:2011-04-05 本文受国家自然科学基金项目(60903088,60903089),河北省自然科学基金项目(F2010000323,F2011201063)资助。
王熙照(1963-),男,博士,教授,主要研究方向为模糊测度与模糊积分、模糊神经系统与遗传算法,E-mail:xizhaowang@ieee.org;崔芳芳(1984-),女,硕士,主要研究方向为支持向量机;鲁淑霞(1966-),女,博士,教授,CCF会员,主要研究方向为机器学习与计算智能、支持向量机。

由于每一个样本的密度指标值不同，因此样本对模型的贡献也是不同的，DPSVM给不同的样本赋予了不同的惩罚因子。将该方法用于UCI[10]数据集中，与其他方法分别进行比较，实验结果表明，DPSVM具有较好的分类性能。

2 近似支持向量机

PSVM是Mangasarian等人在优化理论的基础上提出的一种对SVM的变形。PSVM将SVM的二次规划问题转化为求解线性方程的问题，达到了降低时间复杂度的目的，使得求分类器的过程简单、快速，适合解决大样本问题。标准的PSVM的分类面是拟合数据点，得到两个平行的间隔面，这两个间隔面通过类中心，使得样本的聚类误差尽可能小的同时，两个间隔面的间隔最大。

标准的PSVM的原始问题为：

$$\min_{(\omega,b,\xi)\in R^{n+1+m}} \upsilon \| \xi \|^2+\frac{1}{2}(\omega'\omega+b^2) \quad (1)$$
$$s.t.\ D(A\omega-eb)+\xi=e$$

式中，矩阵$A^{m\times n}$表示训练集合，A_+表示训练集合中正类样本组成的集合，A_-表示训练集合中负类样本组成的集合，ξ是误差项，ω是权向量，e为单位向量，D是一个$m\times m$的对角矩阵，对角线元素取$+1$时，对应的点为点集A_+中的元素，对角线元素取-1时，对应的点为点集A_-中的元素。

3 密度加权近似支持向量机

标准的PSVM对数据进行分类时，没有考虑数据分布的不平衡问题，而给每一个样例赋予了相同的惩罚因子υ。但通常情况下两类分类问题中正类和负类的样本个数总是不平衡的，如果给两类样本的惩罚因子相同，则得到的分类面就会偏向于两类中的一类，在一定程度上影响了分类效果。为了处理这种正、负样本的非对称问题，本文引入了密度(Density)加权M到目标函数的二范数误差项中，构成了密度加权近似支持向量机(DPSVM)，其根据样本点的密度信息，给样本赋予不同的惩罚因子。定义样本$\{x_k,y_k\}$处的密度指标为：

$$M_k=\sum_{j=1}^{k}\exp\left(-\frac{\| x_k-x_j \|^2}{\gamma/2}\right) \quad (2)$$

式中，γ表示其邻居半径，一般取值为0.05～0.2之间。两个样本点$\{x_k,y_k\}$和$\{x_j,y_j\}$之间的距离越近，则对密度指标M_k的贡献就越大。如果一个样本相邻的其他样本点很多，则该样本具有较高的密度指标值。对样本$\{x_k,y_k\}$的误差ξ_k进行密度加权为$R_k=M_k\xi_k$，R_k显示了训练样本集的分布信息，同时也显示了样本$\{x_k,y_k\}$对分类器的误差贡献信息。

DPSVM的目标函数为：

$$\min_{\omega,b,\xi\in R^{n+1+m}} \| M^{1/2}\xi \|^2+\frac{1}{2}(\omega'\omega+b^2) \quad (3)$$
$$s.t.\ D(A\omega-eb)=e-\xi$$

M是一个对角矩阵，即

$$\begin{cases} M_{ij}=M_i, & i=j \\ M_{ij}=0, & i\neq j \end{cases} \quad (4)$$

3.1 线性密度加权近似支持向量机

DPSVM对数据集合线性分类时，根据KKT条件，可以将目标函数转换为如下的Lagrange函数。

$$L(\omega,b,\xi,\alpha)=\| M^{1/2}\xi \|^2+\frac{1}{2}(\omega'\omega+b^2)-\alpha[D(A\omega-eb)+\xi-e]$$

式中，$\alpha\in R^m$是Lagrange乘子，根据Karush-Kuhu-Tucher(KKT)条件，Lagrange函数关于ω,b,ξ,α求导可以得到：

$$\begin{aligned} &\omega=A'D\alpha \\ &b=-e'D\alpha \\ &\xi=\frac{1}{2}(M^{1/2})^{-1}\alpha \end{aligned} \quad (5)$$

由此可以得到α的表达式为：

$$\begin{aligned} \alpha&=((M^{1/2})^{-1}+D(AA'+ee')D)^{-1}e \\ &=((M^{1/2})^{-1}+HH')^{-1}e \end{aligned} \quad (6)$$

式中，$H=D[A-e]$。

由式(5)、式(6)确定(ω,b)之后，就可以对一个未知样例x进行预测：

$$x'\omega-b\begin{cases} >0, x\in A_+ \\ <0, x\in A_- \\ =0, x\in A_+ \text{或} A_- \end{cases} \quad (7)$$

3.2 非线性密度加权近似支持向量机

当非线性可分时，可以引入一个核函数(例如高斯核函数$K,K_{ij}=K(A_i,A_j')=\varphi(A_i)\varphi(A_j)'$)将数据映射到高维空间中，这里$AA'$可以用核函数$K(A,A')$来表示。用$\omega=A'D\alpha$代替目标函数中的$\omega$，可以得到：

$$\min_{\omega,b,\xi\in R^{n+1+m}} \| M^{1/2}\xi \|^2+\frac{1}{2}(a'a+b^2) \quad (8)$$
$$s.t.\ D(AA'Da-eb)=e-\xi$$

用非线性核$K(A,A')$来代替AA'，于是得到非线性DPSVM的目标函数：

$$\min_{\omega,b,\xi\in R^{n+1+m}} \| M^{1/2}\xi \|^2+\frac{1}{2}(\alpha'\alpha+b^2)$$
$$s.t.\ D(K(A,A')D\alpha-eb)=e-\xi$$

令$K=K(A,A')$，利用KKT条件，可以得到拉格朗日函数：

$$L(s,b,\xi,\eta)=\| M^{1/2}\xi \|^2+\frac{1}{2}(\alpha'\alpha+b^2)+\eta(D(KD\alpha-eb)-e+\xi)$$

从而得到相应的拉格朗日乘子η：

$$\begin{aligned} \eta&=[(M^{1/2})^{-1}+D(KK'+ee')D]^{-1}e \\ &=[(M^{1/2})^{-1}+HH']^{-1}e \end{aligned} \quad (9)$$

式中，$H=D[K-e]$。由于$K=K(A,A')$是一个方阵，转置以后维数不会降低，因此式(9)不再适用于用Sherman-Morrison-Woodbury[11]求解。由于数据集的规模很大，核矩阵K的规模也很大，当求逆矩阵时降低了训练速度，因此这里使用约简支持向量机(Reduced Support Vector Machine，RSVM)[12]，从数据集合$A^{m\times n}$中任意选取一个子集合$\overline{A}^{\overline{m}\times n}$，其中$\overline{m}\ll m$，用$\overline{A}'$代替$A'$，核矩阵就由$K(A,A')$变为$K(A,\overline{A}')$，从而减少了矩阵的运算量，缩短了运行时间。

非线性DPSVM的决策函数为：

$$(K(x',A')K(A,A')'+e')D\eta\begin{cases} >0, x\in A_+ \\ <0, x\in A_- \\ =0, x\in A_+ \text{或} A_- \end{cases} \quad (10)$$

4 实验结果与分析

为了验证本文所提出的基于密度加权的PSVM的性能，在UCI机器学习数据库中的5个数据集上进行了实验，并与其他的方法做了比较。这里所选取的Pima数据集为两类数

据，因 Cmc，Yeast，Sat，Spambase 数据集都是多类数据，故人为地将多类数据分为两类。表 1 给出了所选数据集的样例个数、维数，以及分成两类以后的正、负样例的个数，正类用 P 表示，负类用 N 表示。

表 1　试验数据

数据集合	样例个数	维数	正负样例的个数	
Pima	768	9	P:268	N:500
Cmc	1473	10	P:629	N:844
Yeast	1484	9	P:1299	N:185
Sat	4435	37	P:3397	N:1038
Spambase	4601	58	P:1813	N:2788

由于在求解 DPSVM 的优化问题中，涉及到矩阵的求逆问题，而有些数据集合中存在奇异样本数据，因此在进行实验前需先对所有的数据进行归一化，即把数据归一到[0，1]之间。在对数据集进行非线性分类时，引用的核函数是高斯核函数，核参数的选择采用的是网格搜索法，将参数值的可行区间划分为一系列的小区间，然后求出对应各参数值的组合，从而求得该区间内最小目标值及其对应的参数值。表 2 给出了用网格搜索法得到的各个数据集所对应的高斯核函数的参数。在求解优化问题的过程中，涉及到核矩阵的求逆，由于数据集的规模较大，在核矩阵的求逆过程中消耗时间很多，因此降低了训练速度。这里采用了约简支持向量机的方法来减少运行时间。然后从数据集合中选取 70%的样例作为训练样例来训练分类器，30%的样例作为测试样例来测试分类器的性能。

表 2　高斯核参数 σ

数据集合	Pima	Cmc	Yeast	Sat	Spambase
σ	50	10	45	8.5	0.5

表 3 给出了标准 SVM、标准 PSVM、DPSVM 的线性分类结果。从表中可以看出，密度加权近似支持向量机在 5 个数据集上的精度都比 PSVM 要高，因为在求解过程中 DPSVM 需要求一个密度指标，运行时间上比标准的 PSVM 要多，但是与标准的 SVM 运行时间相比还是很低的。

表 3　线性分类结果

数据集合		SVM	PSVM	DPSVM
Pima	时间(s)	4.220	0.015	0.750
	精度(%)	81.34	67.91	67.95
Cmc	时间(s)	53.09	0.063	5.094
	精度(%)	68.08	66.81	67.23
Yeast	时间(s)	6.190	0.078	3.250
	精度(%)	94.01	85.54	87.67
Sat	时间(s)	308.95	1.500	80.50
	精度(%)	93.94	91.76	95.79
Spambase	时间(s)	144.43	1.750	139.78
	精度(%)	74.95	80.51	82.73

表 4 列出了 SVM、标准 PSVM、DPSVM 的非线性分类结果，核函数引用的是高斯核函数，核参数的选择采用的是网格搜索法。在求解优化问题的过程中，核矩阵的规模较大，导致训练速度降低，这里采用了约简支持向量机的方法，减少了运行时间。从表 4 中的实验结果可以看出 3 种分类方法的分类性能，DPSVM 比 PSVM 的精度稍高，或者是与 PSVM 的精度相等，但是训练时间要比标准的 SVM 少很多。因此本文所提出的密度加权的近似支持向量机具有较好的分类性能。

表 4　非线性分类结果

数据集合		SVM	PSVM	DPSVM
Pima	时间(s)	3.460	0.046	0.343
	精度(%)	67.54	80.97	81.72
Cmc	时间(s)	33.49	3.563	3.500
	精度(%)	68.50	66.81	82.53
Yeast	时间(s)	5.806	0.141	3.359
	精度(%)	85.545	85.54	86.57
Sat	时间(s)	124.92	4.390	83.109
	精度(%)	97.22	96.79	96.79
Spambase	时间(s)	211.69	4.891	91.781
	精度(%)	80.87	99.93	99.95

结束语　本文针对数据集中正类点和负类点分布不平衡的特点，提出了一种基于密度加权的近似支持向量机。该方法充分考虑了样本分布的情况，先计算每一个样本的密度信息，再根据密度信息对样本的误差项进行密度加权，从而给样本赋予了不同的惩罚因子，使得分类面不会偏向于某一类，提高了分类性能。通过在 UCI 数据集上的实验结果及 3 种方法的比较和分析可以看出，DPSVM 能得到较好的测试精度，具有更好的预测能力。

参 考 文 献

[1] Vapnik V N. The Nature lf Statistical Learning Theory [M]. New York, USA: Springer-Verlag, 1995: 273-297

[2] Osuna E, Girosi F. Reducing the Run-Time Complexity of Support Vector Machines [C] // Proceedings of the International Conference on Pattern Recognition. 1998

[3] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. Neural Process, 1999, 9(3): 293-300

[4] Fung G, Mangasarian O. Proximal Support Vector Machines [C] // Proceedings of the Knowledge Discovery and Data Mining Conference. San Francisco. 2001: 77-86

[5] Kumar M A, Gopal M. Least Squares Twin Support Vector Machines for Pattern Classification [J]. Expert Systems with Applications. 2008

[6] Tikhonov A N, Arsenin V Y. Solutions of Ill-Posed Problems [J]. Jhon Wiley & Sons, New York, 1997

[7] Chew H G, Bogner R E, Lim C C. Dual v-Support Vector Machines with Error Rate and Training Size Biasing[C] // IEEE Int Conf Acoust Speech Signal Process Proc. Piscataway. 2001: 1269-1272

[8] Carlos R. Improving Radial Basis Function Kernel Classification Through Incremental Learning and Automatic Parameters Selection [J]. Neurocomputing, 2008, 72: 3-14

[9] Lin C F, Wang S D. Fuzzy Support Vector Machines [J]. Transsctions on Neural Network, 2002, 13(2): 464-471

[10] Murphy M. UCI-Benchmark Repository lf Artificial and Real Data Set[OB/OL]. http: // www. ics. uci. edu

[11] Golub G H, van Loan C C. Matrix Computations [M]. Baltimore, Maryland: The John Hopkins University Press, 1996

[12] Lee Y-J, Manfasarian O L. RSVM: Reduced Support Vector Machines [C] // Proceedings of the First SLAM International Conference on Data Mining Chicago. 2001: 5-7