一种多项式光滑的半监督支持向量机分类算法

刘叶青[1,2] 刘三阳[1] 谷明涛[3]
(西安电子科技大学数学科学系 西安 710071)[1] (河南科技大学理学院 洛阳 471003)[2]
(解放军 96251 部队 洛阳 471003)[3]

摘 要 为了处理半监督支持向量机优化中的非凸非光滑问题,引入一个多项式光滑函数来逼近非凸的目标函数,给出的多项式函数在样本的高密度区逼近精度高,逼近精度低时出现在样本的低密度区。采用共轭梯度法求解模型。在人工数据和UCI数据库中的4个数据集上的实验结果显示,算法不仅能保证标号数据很少时的分类精度,而且不因标号数据的增多而明显提高分类性能,因此给出的分类器性能是稳定的。

关键词 半监督学习,支持向量机,分类
中图法分类号 TP18 **文献标识码** A

Polynomial Smooth Classification Algorithm of Semi-supervised Support Vector Machines

LIU Ye-qing[1,2] LIU San-yang[1] GU Ming-tao[3]
(Department of Mathematical Sciences, Xidian University, Xi'an 710071, China)[1]
(School of Science, Henan University of Science & Technology, Luoyang 471003, China)[2] (PLA Unit 96251, Luoyang 471003, China)[3]

Abstract In order to solve the nonconvex and nonsmooth problem of semi-supervised support vector classification, a polynomial smooth function was introduced in this paper which was used to approach the nonconvex objective function. The introduced polynomial function has a high approximation accuracy in high density regions of samples and poor approximation performance appear in low density regions of samples. The model was solved by the method of conjugate gradient. Experimental results on artificial and real data support that the proposed algorithm can guarantee the accuracy when the percentage of labeled sample is very low and the accuracy is not improved obviously as the number of labeled data increasing. The performance of the proposed classifier is stable.

Keywords Semi-supervised learning, Support vector machine(SVM), Classification

1 引言

支持向量机[1](support vector machine-SVM)由于其出色的泛化性能已在很多领域引起人们的广泛关注,如模式识别和回归估计。然而为了得到精确的分类器,支持向量机训练过程中需要大量的标号数据。虽然随着信息量的增大,大量未标号数据的获得是廉价的,然而人工标号的过程很浪费时间,同时也经常产生错误。当有小部分的标号数据和大部分的未标号数据时,半监督学习一般能给我们提供一个性能良好的分类器。在训练过程中,半监督学习同时运用了标号数据的信息和未标号数据的信息以提高泛化性能,这就优于只在标号数据上训练,产生的分类器对训练数据过拟合的监督学习。半监督学习的实用性使其近年来受到了越来越多的关注。将半监督学习和支持向量机这一机器学习的新工具相结合已成为机器学习新的研究热点[2-6]。

半监督支持向量机(S^3VMs)把间隔最大化原则应用到标号和未标号样本,因此不像SVM最后得到的是一个凸优化问题,S^3VMs得到的是一个非凸优化问题。文献中利用各种不同的优化技术来求解S^3VMs[7-12]。2005年Chapelle和Zien[13]用指数函数逼近的方法来建立一个无约束的半监督支持向量机模型。光滑函数的应用使原来不可微的模型变成可微模型,从而能用优化中的优秀算法来求解。文献[13]中使用的指数函数虽然具有任意阶光滑,但是它的逼近精度却不高。本文用一个多项式光滑函数来逼近,并用共轭梯度法求解模型,实验结果显示本文给出的分类器不仅能保证标号数据很少时的分类精度,而且不因标号数据的增多而明显提高分类性能,与用文献[13]中给出的指数函数的分类器相比,性能稳定且精度有明显提高。

2 半监督支持向量机(S^3VMs)

考虑两类分类问题。训练集包个标号样本$\{(x_i,y_i)\}_{i=1}^{l}$,$y_i=\pm1$,和u个未标号样本$\{x_i\}_{i=l+1}^{l+u}$。目的是找到一标号向量$y_u=[y_{l+1},\cdots,y_{l+u}]$,使得SVM在$\{(x_i,y_i)\}_{i=1}^{l+u}$上的训练能得到最大间隔。在线性情况下,$S^3$VMs可写为下面的最小化问题:

$$\min \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{i=l+1}^{l+u}\xi_i$$

到稿日期:2008-08-22 返修日期:2008-11-19 本文受国家自然科学基金(60574075)资助。
刘叶青(1979-),女,博士生,研究方向为模式识别、机器学习、最优化理论及其应用,E-mail:xiangshui15@163.com;刘三阳(1959-),博士,教授,博士生导师,主要研究方向为最优化理论与算法、最优化技术在机器学习中的应用。

$$s.t.\ y_i(w^Tx_i+b)\geqslant 1-\xi_i, i=1,\cdots,l \quad (1)$$
$$|w^Tx_i+b|\geqslant 1-\xi_i, i=l+1,\cdots,l+u$$

对非线性的情况，可以使用核函数来构造[1]。

上面的最小化问题可以等价地写为下面的无约束最小化问题：

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}w^2 + C\sum_{i=1}^{l} L(y_i(w^Tx_i+b)) + C^*\sum_{i=l+1}^{l+u} L(|w^Tx_i+b|) \quad (2)$$

其中 L 为损失函数，$L(t)=\max(0,1-t)$ 。

式(2)中的最后一项使得问题非凸很难求解，文献[13]中用 $L^*(t)=\exp(-3t^2)$ 来代替它。这个光滑函数虽然可以任意阶光滑，但逼近精度却不高，如图 1 所示。为此本文用一个多项式光滑函数来逼近。

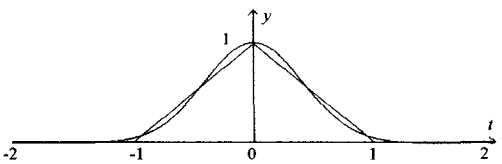

图 1　指数函数 e^{-3t^2} 对 $\max(0,1-|t|)$ 的逼近情况

3　多项式光滑的半监督支持向量机

本文要用多项式光滑函数来逼近 $L(|w^Tx_i+b|)$，由图 1 中 $\max(0,1-|t|)$ 的图像，只用在 $[-1,1]$ 上逼近它即可。给出多项式光滑函数 $P(t)$：

$$P(t)=\frac{1-t^2}{2}+\frac{1}{8}(1-t^2)^2+\frac{1}{16}(1-t^2)^3+\frac{5}{128}(1-t^2)^4+\frac{7}{256}(1-t^2)^5, t\in[-1,1]。$$

用 $P(t)$ 来逼近 $\max(0,1-|t|)$，逼近图像如图 2 所示。同时本文想用共轭梯度法来求解式(2)，因为第二项不光滑，所以把式(2)写为：

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}\|w\|^2+\frac{1}{2}C\sum_{i=1}^{l}L^2(y_i(w^Tx_i+b))+C^*\sum_{i=l+1}^{l+u}P(w^Tx_i+b) \quad (3)$$

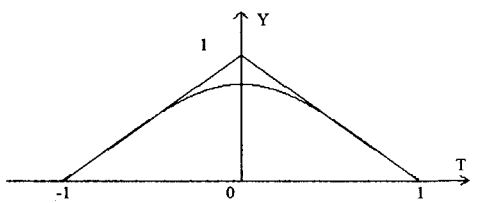

图 2　$P(t)$ 在 $[-1,1]$ 上对 $1-|t|$ 的逼近情况

比较图 1 和图 2 知道，在 $t=\pm1$ 附近，$P(t)$ 的逼近误差远小于 e^{-3t^2} 的逼近误差。而在 $t=\pm1$ 附近正是样本的高密度区，$P(t)$ 的最大逼近误差出现在 $t=0$ 样本的低密度区，应首先保证样本高密度区的逼近精度。这样 $P(t)$ 比 e^{-3t^2} 逼近的优势在于样本的高密度区逼近精度高，逼近精度低时出现在样本的低密度区。

4　算法

令

$$\min_{(w,b)\in R^{n+1}}\varphi(w,b)=\min_{(w,b)\in R^{n+1}}\frac{1}{2}\|w\|^2+\frac{1}{2}C\sum_{i=1}^{l}L^2(y_i(w^Tx_i+b))+C^*\sum_{i=l+1}^{l+u}P(w^Tx_i+b) \quad (4)$$

采用共轭梯度法[14]求解式(4)。常用的共轭梯度法像 PRP 算法及改进的 PRP 算法需要精确地实现一维搜索才能保证构造出共轭方向。而在实际计算中，不能期望精确地实现一维搜索，而且在离极小点很远时也没有这种必要。为克服这个缺点，采用文献[15]中的方法构造搜索方向。即使不进行一维搜索，这样构造的方向也是互相共轭的。算法如下：

Step1　算法初始化：$(w^1,b^1)=P^1\in R^{n+1}$，$\varepsilon>0$，$d^0=0$，$i=2$

Step2　计算 $g^1=\nabla\varphi(P^1)$

Step3　若 $\|g^1\|_2^2\leqslant\varepsilon$，停，$P^1=(w^1,b^1)$ 为最优解；否则 $d^1=-g^1$

Step4　沿方向 d^1 做一维搜索，得到步长因子 $\lambda_1>0$。令 $P^2=P^1+\lambda_1d^1$

Step5　计算 $g^i=\nabla\varphi(P^i)$

Step6　若 $\|g^i\|_2^2\leqslant\varepsilon$，停，$P^i=(w^i,b^i)$ 为最优解；否则

$y_{i-1}=g^i-g^{i-1}$

$\beta_{i-1}=\langle y_{i-1},y_{i-1}\rangle/\langle y_{i-1},d^{i-1}\rangle$

$\gamma_{i-2}=\langle y_{i-2},y_{i-1}\rangle/\langle y_{i-2},d^{i-2}\rangle$

$d^i=-y_{i-1}+\beta_{i-1}d^{i-1}+\gamma_{i-2}d^{i-2}$

Step7　沿方向 d^i 做一维搜索，得到步长因子 $\lambda_i>0$。令 $P^{i+1}=P^i+\lambda_id^i$

Step8　置 $i=i+1$，转 step5

Step4 和 Step7 中求步长时采用不精确一维搜索 Armijo-Goldstein 准则[14]即

$\varphi(P^i+\lambda_id^i)\leqslant\varphi(P^i)+\rho(g^i)^Tu_i$

$\varphi(P^i+\lambda_id^i)\geqslant\varphi(P^i)+(1-\rho)(g^i)^Tu_i$

$u_i=\lambda_id^i$，$0<\rho<\frac{1}{2}$

5　实验

实验分两部分，在人工数据上的实验和在 UCI 数据库中 4 个数据集上的实验。所有实验都是在 AMDsempron(tm) 2400+，512M 内存，MATLAB7.0.1 环境中进行的。

5.1　在人工数据上的实验

此实验的目的是在标号数据较少，信息不充分时，检验本文算法的分类性能。使用的人工数据 artificial 是二维可分的。只标号其中的任意两个(保证一个正类一个负类)，其余的 98 个用两种方法来标号。一是同时考虑了标号数据信息和未标号数据信息的本文算法，另一个是只运用标号数据信息的支持向量机算法即在本文算法中令 $C^*=0$，实验中随机选取一个正类点和一个负类点，实验进行 10 次，结果如图 3 所示。图中 x 轴是实验次数，y 轴是分类精度。实线是本文算法结果，虚线是本文算法中令 $C^*=0$ 的实验结果。

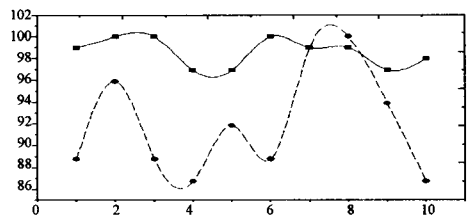

图 3　多项式光滑的 S³VM 和 SVM 分类性能比较

由图 3 知道，在标号数据少信息不充分时，同时运用标号数据信息和未标号数据信息的半监督支持向量机的分类器的性能比只考虑标号数据信息的支持向量机的分类器的性能明显趋于稳定，且分类精度有了明显提高。

5.2 在 UCI 数据集上的实验

评价半监督学习算法好坏的一个重要准则是算法能保证标号数据很少时的分类精度，同时算法不能因标号数据的增多而明显提高分类性能，好的分类器的性能应该是稳定的。为此，从 UCI 数据库中 4 个数据集上分别随机提取 1%，3%，5%，10%作为标号数据，其余的作为未标号数据，用本文算法来标号未标号数据，如表 1 所列。

表 1　UCI 4 个数据集实验的相关数据

Data set	m×n	类别	C	σ=1	σ=3	σ=5	σ=10
Heart	270×14	2	0.1	3	8	14	27
Breast	428×10	2	0.1	4	13	21	43
Diabetes	766×9	2	10	8	23	38	77
Monks	429×7	2	1	4	13	21	43

对给定的数据规模 $m\times n$，m 为样本个数，n 为样本的属性个数。σ 为标号数据占全部数据的比例(%)。按 Armijo-Goldstein 准则求步长时，参数 ρ 取 0.05。比较了 C 和 C^* 取 10^{-3}，10^{-2}，10^{-1}，10^{0}，10^{1}，10^{2} 值的情况，发现 C 和 C^* 的取值对分类精度不敏感，因此取 $C=C^*$。所有数据均经过标准化处理。核函数采用高斯核。表 1 列出了相关的数据。其中 σ 下的数值为各数据集取相应百分比时具体的标号个数。表 2 列出了采用 $P(t)$逼近时在 4 个数据集上的 10 次精度平均值。

表 2　本文算法在 4 个数据集上的平均实验结果

Data set	σ=1	σ=3	σ=5	σ=10
Heart	83.146	83.969	83.594	83.539
Breast	95.047	95.422	95.086	96.364
Diabetes	69.567	69.852	69.78	69.956
Monks	67.765	68.029	67.892	69.171

用 e^{-3t^2} 代替 $P(t)$逼近，再用本文方法求解，在以上 4 个数据集上的实验结果却很不稳定。图 4 给出了在 breast 数据集上标号 4 个数据时，$P(t)$逼近和 e^{-3t^2} 逼近都用本文算法连续 10 次分类的精度。图中 x 轴是实验次数，y 轴是分类精度，实线和虚线分别为 $P(t)$逼近和 e^{-3t^2} 逼近时的分类结果。很明显，$P(t)$逼近时精度提高，而且分类器性能稳定。

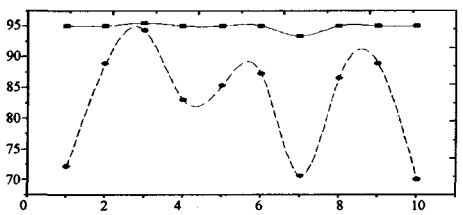

图 4　在 breast 数据集上 $P(t)$逼近和 e^{-3t^2} 逼近分类精度

结束语　目前文献中对 S^3VMs 的研究方法主要有两种：组合优化和连续优化。组合优化是对一个给定的 $y_u=[y_{l+1},\cdots,y_{l+u}]$，在$(w,b)$上优化问题，这样变成一个标准 SVM 训练。根据某一策略，交替进行上面两步。文献[7,9,10]属于这种方法。连续优化是对一个固定(w,b)，$y_u=[y_{l+1},\cdots,y_{l+u}]$的最优由 w^Tx_i+b 的符号给出，用这种方式消除 y_u，给出一个连续的目标函数，如(2)。文献[12,13]属于这种方法。本文的方法属于后者。和文献[13]中方法类似，本文也是引入了光滑函数来逼近非凸的目标函数，但本文的多项式函数的优势是在样本的高密度区函数逼近精度高，逼近精度低时出现在样本的低密度区。

因为目标函数的第二项为一阶光滑，所以对模型采用共轭梯度法求解。在人工数据和 UCI 数据库中 4 个数据集上的实验结果显示，算法不仅能保证标号数据很少时的分类精度，而且不因标号数据的增多而明显提高分类性能，因此本文给出的分类器性能是稳定的。

参 考 文 献

[1] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 127-167

[2] Bennett K P, Demiriz A. Semi-supervised support vector machines[M]. Cambridge: MIT Press, 1998: 368-374

[3] Bengio Y, Grandvalet Y. Semi-supervised learning by entropy minimization[J]. Neural Information Processing Systems, 2004, 17

[4] Li Yuanqing, Li Huiqi, Guan Cuntai, et al. A Self-training Semi-supervised Support Vector Machine Algorithm and its Applications in Brain Computer Interface[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Hawaii, USA, 2007

[5] Qin Jianzhao, Li Yuanqing. An Improved Semi-supervised Support Vector Machine Based Translation Algorithm for BCI Systems[C]// Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China, 2006

[6] Chapelle O, Sindhwani V, Keerthi S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research, 2008, 9: 203-233

[7] Chapelle O, Sindhwani V, Keerthi S. Branch and bound for semi-supervised support vector machines[J]. Neural Information Processing Systems, 2006

[8] Astorino A, Fuduli A. Nonsmooth optimization techniques for semi-supervised classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2135-2142

[9] Joachims T. Transductive inference for text classification using support vector machines[C]// International Conference on Machine Learning. Pittsburgh, 1999

[10] Sindhwani V, Keerthi S, Chapelle O. Deterministic annealing for semi-supervised kernel machines[C]// International Conference on Machine Learning. Pennsylvania, 2006

[11] Collobert R, Sinz F, Weston J, et al. Large scale transductive SVMs[J]. Journal of Machine Learning Research, 2006, 7: 1687-1712

[12] Chapelle O, Chi M, Zien A. A continuation method for semi-supervised SVMs[C]// International Conference on Machine Learning. Pennsylvania, 2006

[13] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]// Tenth International Workshop on Artificial Intelligence and Statistics. Barbados, 2005

[14] 袁亚湘，孙文瑜. 最优化理论与方法[M]. 北京：科学出版社，1997

[15] Nazareth L. A conjugate direction algorithm without line searches[J]. Journal of Optimization Theory and Application, 1977, 23(3): 373-387