文章编号:1000 - 582X(2005)02 - 0081 - 04

一种新型支持向量机*

杨　强[1],吴中福[1],余　萍[2],钟　将[1]
(1. 重庆大学 计算机学院,重庆　400030; 2. 盐城师范学院 计算机系,江苏 盐城　224002)

摘　要:讨论了现有的用于分类的支持向量机(SVM)所确定的边界在抗干扰方面的局限性。在此基础之上提出了一种新型支持向量机,即基于边界调节的支持向量机,并利用 K - T 条件得到了这种支持向量机的对偶目标函数。通过对人工数据集和真实数据集的仿真实验表明,相对于 L1 - SVM 而言,基于边界调节的支持向量机具有更少的支持向量和更好的推广性能。

关键词:支持向量机 SVM;最优超平面;核函数;二次规划
中图分类号:TP181　　**文献标识码**:A

由于引入了结构风险最小化原则,使得支持向量机比传统的统计学习方法具有更为坚实的理论基础。支持向量机在数学上所具有的二次规划形式及其在分类、回归分析和密度估计等方面所表现出来的良好性能,使它越来越受到从事神经网络、模式识别和数据挖掘等领域的学者们的关注。现在,支持向量机已经成为神经网络的一个重要分支,也是从事模式识别和数据挖掘等领域的重要研究手段。

但是,根据目前通用的方法所得到的支持向量机存在抗干扰能力差,对噪声信号敏感等问题[1-2]。对野点给予适当的处理一直是与支持向量机密切相关的一个重要论题。人们为此也提出了不同形式的支持向量机,例如 L1 - SVM、L2 - SVM、μ - SVM 等。近年来还提出了模糊支持向量机[3]和经验风险最小化的 SVM[4]。各种 SVM 在不同的样本集中,有不同的推广性能,但目前还没有一种 SVM,对所有数据集,均好于其它 SVM。因此对 SVM 进行研究是很有必要的。

笔者对现有的 SVM 在最优超平面的构造上给予改进,提出了一种基于边界调节的支持向量机。实验表明,该种 SVM 相对于 L1 - SVM 具有更强的抗干扰能力,对真实数据集具有更好的推广性能。

1　支持向量机的一般性原理

支持向量机是一种基于结构风险最小化原则,以构造最优超平面为目标的统计学习机器。对于在输入空间中不能正确分类的数据集就利用非线性变换 $\Phi(\boldsymbol{x}_i)$, $i=1,2,\cdots,l$,将样本映射到某一更高维的特征空间中,使样本在这个高维的特征空间中可实现正确分类。在特征空间中样本之间的内积用核函数 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 表示。在学习中,为了防止过学习现象,允许一定的经验风险的存在,为此在目标函数中引入松弛变量。这种引入松弛变量的方法是目前用支持向量机进行学习的最主要方法。由于 L1 - SVM 具有广泛的代表性,在这里对 L1 - SVM 进行概略的介绍。L1 - SVM 的目标函数为[5](以下论述中,采用非核函数形式):

$$\begin{aligned} &\min \quad \frac{1}{2}\boldsymbol{\omega}\cdot\boldsymbol{\omega} + C\sum_{i=1}^{l}\xi_i \\ &\text{s.t.} \quad \xi_i \geqslant 0, \quad i = 1,2,\cdots,l \\ &\qquad y_i(\boldsymbol{\omega}\cdot\boldsymbol{x}_i - b) \geqslant 1 - \xi_i \end{aligned} \tag{1}$$

其中 · 表示内积运算,C 为一个大于 0 的常数。根据 K - T 条件,得到对偶目标函数:

$$\begin{aligned} &\max \quad W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \\ &\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,\cdots,l \\ &\qquad \sum_{i=1}^{l} y_i\alpha_i = 0 \end{aligned} \tag{2}$$

通过求解这个二次规划问题得到相应的系数 $\alpha_i^0 \geqslant 0$, $i=1,2,\cdots,l$,其中大于 0 的系数所对应的样本向量就是支持向量。进而得到关于最优超平面的权值 $\boldsymbol{\omega}_0$ 和

* 收稿日期:2004 - 09 - 15
作者简介:杨强(1972 -),男,重庆人,博士,主要研究方向:支持向量机和图像处理。

阈值 b_0。

$$\omega_0 = \sum_{\text{支持向量}} y_i \alpha_i^0 x_i \tag{3}$$

$$b_0 = \frac{1}{2}[(\omega_0 \cdot x^*(1)) + (\omega_0 \cdot x^*(-1))] \tag{4}$$

其中 $x^*(1)$,$x^*(-1)$分别为属于第 1 类和第 2 类且对应的 ξ 等于 0 的支持向量。因此,得到基于最优超平面的分类规则的指示函数:

$$f(x) = \operatorname{sgn}(\sum_{\text{支持向量}} y_i \alpha_i^0 (x_i \cdot x) - b_0) \tag{5}$$

对于需要进行非线性变换,而引入核函数的情况只需将上述的 x 替换成 $\Phi(x)$,$(x_i \cdot x_j)$替换成 $k(x_i \cdot x_j)$即可[5]。

2 基于边界调节的支持向量机

从上一节的分析可知,L1 - SVM 是在约束条件:

$$\xi_i \geqslant 0, \quad i = 1,2,\cdots,l \tag{6}$$

和 $$y_i((\omega \cdot x_i) - b) \geqslant 1 - \xi_i, \quad i = 1,2,\cdots,l \tag{7}$$

的基础上得到的。从一般意义上讲,这 2 个约束条件是合理的,也是必须的。但是,式(7)表示 SVM 能够对 $0 < \xi_i < 1$ 的样本给予正确的判断,只是由于该样本距离最优超平面太近,而给予一定的惩罚;而将 $\xi_i > 1$ 的样本作为野点,即 SVM 不能对该样本给予正确的判断。L2 - SVM、μ - SVM 也是基于这样的假设建立起来的。但是,对不同的数据集和不同的核函数,都将 $\xi_i > 1$ 的样本作为野点并不是十分合理的。

为有效地解决这个问题,对边界约束条件给予改进。将 $\xi_i > d$(d 是一个大于 0 的常数)的样本作为野点,得到如下目标函数(以下论述中,采用非核函数形式):

$$\begin{aligned} \min \quad & \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{l} \xi_i \\ \text{s.t.} \quad & \xi_i \geqslant 0, \quad i = 1,2,\cdots,l \\ & y_i(\omega \cdot x_i - b) \geqslant d - \xi_i \end{aligned} \tag{8}$$

其中,d 为边界调节所对应的常数。根据 K - T 条件,得到对偶目标函数:

$$\begin{aligned} \max \quad & W(\alpha) = d^2\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant Cd \quad i = 1,2,\cdots,l \\ & \sum_{i=1}^{l} y_i \alpha_i = 0 \end{aligned} \tag{9}$$

显然,式(9)所确定的 Hessian 矩阵与式(2)所确定 Hessian 矩阵是完全相等的,满足正定性条件。因此,可以确保对偶目标函数式(9)与原目标函数式(8)的等价性。

通过求解,得到关于最优超平面的权值$\frac{\omega_0}{d}$和阈值$\frac{b_0}{d}$:

$$\frac{\omega_0}{d} = \frac{1}{d^2}\sum_{\text{支持向量}} y_i \alpha_i^0 x_i \tag{10}$$

$$\frac{b_0}{d} = \frac{1}{2d^2}[(\sum_{\text{支持向量}} y_i \alpha_i (x_i \cdot x^*(1))) + (\sum_{\text{支持向量}} y_i \alpha_i (x_i \cdot x^*(-1)))] \tag{11}$$

因此,得到基于边界调节的支持向量机的分类规则的指示函数:

$$f(x) = \operatorname{sgn}((\sum_{\text{支持向量}} y_i \alpha_i^0 (x_i \cdot x) - b_0)) \tag{12}$$

由式(9)可知,所需求解的目标函数是一个典型的二次规划问题。且式(9)相对于式(2)而言,只是在一次项系数和约束条件多了一个常数。因此,在处理较小的数据集时,可以利用现有的二次规划软件,在对比较大的数据集进行处理时,只需对已有的序列算法[6]或改进算法[7]作简单的修改即可。

3 实验与分析

一种新理论或新方法是否正确和具有一定意义,就应看它是否满足以下 2 个条件:

1) 能够解决现有理论或方法能够解决的问题;

2) 能够更好地解决现有理论或方法不能很好解决的问题。

因此,在这里主要做了以下 3 个方面的实验:

1) 无野点的人工数据对比实验;

2) 有野点的人工数据对比实验;

3) 真实数据的对比实验。

在所有实验中,都使用高斯核函数:

$$k(x_i - x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}) \tag{13}$$

无野点的人工数据是随机产生的 $x_i, y_j \in (0,1)$ 的均匀分布的 200 个二维数据点 $x_i = (x_i, y_i)$,$i = 1,2,\cdots,l$,并在此基础上去掉 $\| x_i - x_j \| < 0.001$,$i,j = 1,2,\cdots,200$ 的点所得到的。其中,如果 $x_i < y_i$,则样本属于第 1 类,如果 $x_i > y_i$,则样本属于第 2 类,如图 1 所示。利用已有的 L1 - SVM 的实验结果和基于边界调节的支持向量机所得到的分类超平面如图 2、图 3 所示,所得到的实验结果如表 1 所示。

表 1 中的 Cd 对于 L1 - SVM 而言,$Cd = C$,$\omega_0/d = \omega_0$,$b_0/d = b_0$,即 $d = 1$。

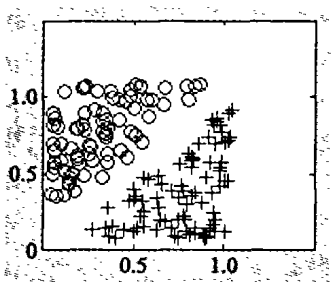

图 1　无野点的人工数据

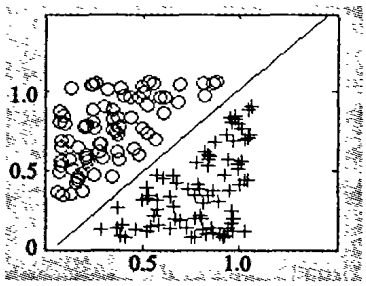

图 2　L1 - SVM 的分类超平面

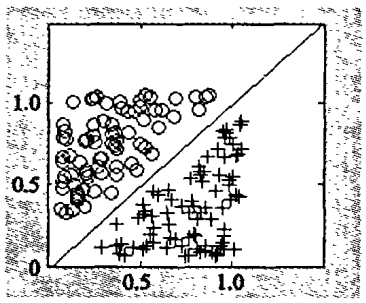

图 3　$d^2 = 0.2$ 的基于边界调节的 SVM 的分类超平面

表 1　无野点的人工数据集实验结果

支持向量机类型	常数 Cd	支持向量个数 m	权值(ω_0/d)	阈值(b_0/d)
L1 - SVM	1.0	17	(5.775 5, -5.698 5)	-0.087 9
边界调节 SVM	0.5	11	(5.774 8, -5.757 6)	-0.056 4
边界调节 SVM	0.2	7	(5.589 5, -5.513 8)	-0.050 9

实验结果说明，L1 - SVM 与基于边界调节的 SVM 都能够对该数据集实现正确分类。但是，基于边界调节的 SVM 的支持向量个数小于 L1 - SVM 的支持向量个数。这说明基于边界调节的 SVM 的结构复杂度低于 L1 - SVM 的结构复杂度。由推广错误率的上界[5]：

$$EP_{\text{error}} \leqslant E\left(\frac{m}{l}\right) \tag{14}$$

得知，基于边界调节的 SVM 的推广性能比 L1 - SVM 的推广性能好。

上面讲述了无野点人工数据的实验结果，下面讲述带野点的人工数据的实验情况。对于本实验而言所谓带野点的人工数据就是将前面的无野点人工数据中的一个处于边缘的数据所属的类改变为另一相反的类所得。如图 4 所示。利用已有的 L1 - SVM 的实验结果和基于边界调节的支持向量机所得到的分类超平面如图 5、图 6 所示，所得到的实验结果如表 2 所示。表 2中的 d - SVM 表示基于边界调节的支持向量机。

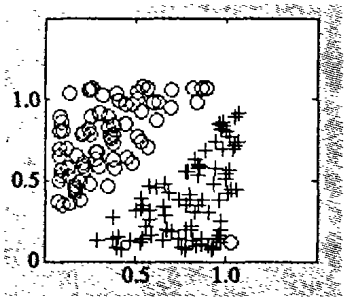

图 4　有野点的人工数据

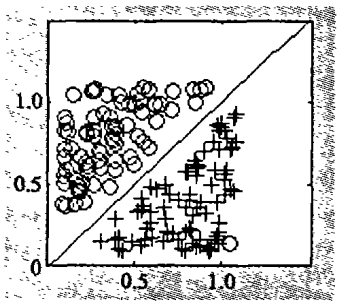

图 5　L1 - SVM 的分类超平面

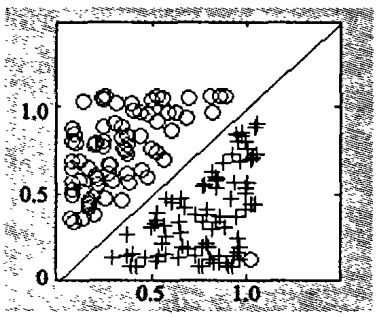

图 6　$d^2 = 0.2$ 的基于边界调节的 SVM 的分类超平面

表 2　有野点的人工数据集实验结果

支持向量机类型	常数 Cd	支持向量个数 m	权值(ω_0/d)	阈值(b_0/d)
L1 - SVM	1.0	22	(5.339 4, -5.086 6)	-0.042 4
d - SVM	0.5	17	(5.319 6, -5.280 7)	-0.038 9
d - SVM	0.2	11	(5.229 7, -5.224 6)	-0.031 9

与无野点信号的情况相似，该实验表明，在带有野点信号的情况下，基于边界调节的 SVM 仍然具有比 L1 - SVM 更好的推广性能。

在讲述了基于边界调节的 SVM 在人工数据集方面的表现之后，在下面讨论基于边界调节的 SVM 在真实数据集中的表现。实验数据为 Titanic 数据集和 breast cancer 数据集[8]。实验结果如表 3 所示。表 3 中的 d - SVM 表示基于边界调节的支持向量机。实验结果数据表明，基于边界调节的 SVM 相对于 L1 - SVM 而言，具有更少的支持向量个数及更好的推广性能。

表 3 对真实数据集的实验结果

数据集	方法	支持向量个数	推广错误率
Titanic	L1 - SVM	9	23.56
Titanic	d - SVM	8	22.47
Breast cancer	L1 - SVM	180	31.27
Breast cancer	d - SVM	162	29.32

以上的实验结果说明,对一般数据集而言,基于边界调节的 SVM 的性能比 L1 - SVM 的性能要好。因此,基于边界调节的 SVM 是具有理论和现实意义的。

4 结 语

提出了基于边界调节的支持向量机,通过理论推导给出了相应的对偶目标函数,值得注意的是该目标函数与 L1 - SVM 的对偶目标函数具有相似的形式。因此,只须将现有的一些关于支持向量机的算法作略微的修改即可使用。通过实验证明基于边界调节的支持向量机具有更好的抗干扰能力和具有更好的推广性能。

进一步的研究方向:由于基于边界调节的支持向量机的目标函数相对于分类错误程度的惩罚因子而言仍然是线性的。因此,为了进一步提高支持向量机的推广能力,笔者将研究关于分类错误的非线性惩罚问题。

参考文献:

[1] GUYON I, MATIC N, VLADIMIRNVAPNIK. Discovering Information Patterns and Data Cleaning[M]. Cambridge: MA MIT Press, 1996, 181 - 203.

[2] ZHANG X. Using Class-center Vectors to Build Support Vector Machines[Z]. Proc IEEE NNSP'99, Wisconsin, USA, 1999.3 - 11.

[3] LIN CHUN FU, WANG SHENG DE. Fuzzy Support Vector Machines[J]. IEEE Transac Neural Networks, 2002, 13(2):464 - 471.

[4] FERNANDO. Emprical Risk Minimization for Support Vector Classifiers[J]. IEEE Transac Neural Networks, 2003, 14(2):296 - 303.

[5] VLADIMIRNVAPNIK. Statistical Learning Theory[M]. New York: Wiley, 1998.

[6] JOHN C PLATT. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines in Advances in Kernel Methods-Support Vector Learning[M]. Cambridge, MA:MIT Press, 1999. 185 - 208.

[7] KEERTHI S, SHEVADE S, BHATTACHARYYA C, et al. Improvements to SMO Algorithm for SVM Classifier[J]. Neural Comput, 2001, 13(3):637 - 649.

[8] BLAKE, CLMERZ. UCI Machine Learning Repository[EB/OL]. http://ida. first. gmd. de/ ~ raetsch/data, 2003 - 12 - 18.

A Sort of New SVM

YANG Qiang[1], *WU Zhong-fu*[1], *YU Ping*[2], *ZHONG Jiang*[1]

(1. College of Computer, Chongqing University, Chongqing 400030, China;
2. Dept of Computer, Yancheng Normal College, Yancheng 224002, China)

Abstract: The SVM 's general theorm and shortcoming in resistance desturbance and noise is discussed. The authors find that these shortcoming which is caused by the traditional separating hyperplane. They also present a define, which is called as adjustable separating hyperplane. Basing the adjustable separating hyperplane, a new sort of SVM is set up, and corresponding quadratic programming dual objective function is obtained as well. Simulation results of artificial and real data show that the sort of SVM based on adjustable margin has less number of support vectors and better generating ability than L1-SVM. So, the sort of SVM has some meaning of theory and realism.

Key words: support vectors; SVM; optimum hyperplane; kernel function; quadratic programming

(编辑 张 苹)