基于误差最小的 SVM 最优分类面修正

蒋觉义 何玉珠 李建宏

(北京航空航天大学 仪器科学与光电工程学院, 北京 100191) (北京航空航天大学 可靠性与系统工程学院, 北京 100191)

摘 要: 针对 *C*-支持向量机(*C*-SVM, *C*-Support Vector Machine)中惩罚系数 *C* 可能导致最优分类面不合理的问题,提出基于误差最小的 SVM 最优分类面修正方法. 通过调整正负类分类间隔的约束条件,求解使训练样本总误差最小的偏置系数,并兼顾与正负类误差之差的绝对值的平衡,得到误差最小的更优分类面. 实验证明该修正方法与 *C*-SVM 及其它修正方法相比,具有较高的分类精度和较强的抗噪声与野值数据干扰能力.

关 键 词: 支持向量机; 最优分类面; 修正

中图分类号: TP 206^{+}. 3

文献标识码: A **文 章 编 号**: 1001-5965(2012)11-1483-04

Modification of SVM's optimal hyperplane based on minimal mistake

Jiang Jueyi He Yuzhu

(School of Instrument Science and Opto-electronics Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Li Jianhong

(School of Reliability and Systems Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Since some value of error penalties *C* in *C*-support vector machine (*C*-SVM) may cause extreme and irrational optimal separating hyperplanes, a new modification of SVM's optimal hyperplane was proposed. By modifying the distance restriction of separating hyperplane between positive and negative classes, the bias coefficient was calculated with minimal training samples' total error, while the absolute value of the error difference between positive and negative classes was balanced considered, a better separating hyperplane with minimal mistake was obtained. The experimental results show that this algorithm has improved the classified precision and enhanced the ability of reducing the outliers and noises' effect, compared to *C*-SVM and other modification algorithm.

Key words: support vector machine; optimal separating hyperplane; modification

标准的支持向量机(SVM[1], Support Vector Machine)为基于线性可分情况下提出,其最优分类面将两类点完全正确地分开,且距两类点间隔最大. 在线性不可分情况下,这种分类面是不存在的. *C*-支持向量机(*C*-SVM, *C*-Support Vector Machine)为解决不可分问题,引入惩罚系数 *C* 作为分类间隔和训练集被错划的程度这个两目标之间的权重,从而得到平分最近点和最大间隔之间的折衷. 这种方法得到的分类面依赖于样本点和 *C*,当 *C* 取值较小时,很可能出现当所有训练样本划分为一类的极端情况,这显然是不合理的. 关于分类面问题的研究,一些学者在不同程度上跳出 SVM 思想框架,如文献[2-5],以丧失 SVM 的一些优良性质和增加很多运算量为代价设计分类面,具有较低性价比;也有学者基于 SVM 思想框架下的方差修正法[6],分类精度得到一定提高,但是存在稳定性不高、对噪声更为敏感等问题.

本文结合 SVM 的数学模型,提出基于误差最小的 SVM 最优分类面的修正方法,通过对传统 *C*-SVM 中求得的分类面进行适当修正,使新得到

收稿日期: 2011-06-21; **网络出版时间**: 2012-11-09 17:20

网络出版地址: www.cnki.net/kcms/detail/11.2625.V.20121109.1720.012.html

作者简介: 蒋觉义(1987-),男,江西余干人,硕士生,ijueyi@aspe.buaa.edu.cn.

的分类面对训练集的错划程度最小,在一定程度上降低 C 取值的困难程度.

1 支持向量机

标准的支持向量机核心思想是构建一个最大间隔的分类超平面,使得已知类别的两类样本分布于分类超平面的两边.

对于线性可分的两类样本:

$$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_i,y_i),\cdots,(\boldsymbol{x}_l,y_l)$$

$$\boldsymbol{x}_i \in \mathbf{R}^n \quad y_i \in \{-1,1\} \quad i=1,2,\cdots,l$$

其中,l 为训练样本总数;n 为样本空间的维数;$\boldsymbol{x}_i$ 为样本空间向量;y_i 为样本的类别标志. 分类超平面为,$\boldsymbol{w}\cdot\boldsymbol{x}+b=0$ 如图 1 所示,其中 $\boldsymbol{w}$ 为最优超平面的法向量;b 为偏置;R_1 和 R_2 分别为 $\boldsymbol{x}$ 的第一维和第二维的取值.

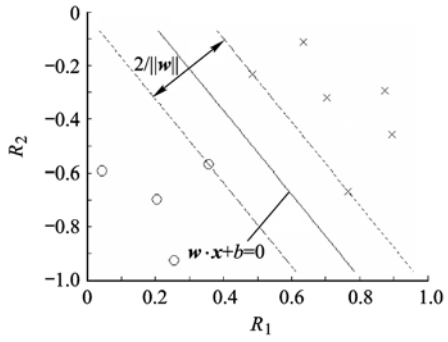

图 1 最优超平面

对于线性可分的情况,最优超平面的求解问题可归结为如下二次规划问题:

$$\left.\begin{aligned}&\min_{\boldsymbol{w},b}\varphi(\boldsymbol{w})=\frac{1}{2}\|\boldsymbol{w}\|^2\\&\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1 \quad i=1,2,\cdots,l\end{aligned}\right\} \tag{1}$$

对于线性不可分情况下,满足 $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1, i=1,2,\cdots,l$ 的分类超平面并不存在,文献[7]提出 C-SVM. C-SVM 引入 ξ_i 作为样本 $\boldsymbol{x}_i$ 的松弛变量,以 $\sum_{i=1}^{l}\xi_i$ 作为训练集被错划的程度的度量,并以参数 C 作为分类间隔 $\|\boldsymbol{w}\|^2/2$ 和训练集被错划的程度这个两目标之间的权重,极小化新的目标函数 $\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i$,求得最优解 $\boldsymbol{w}^*$,b^*,$\boldsymbol{\xi}^*$,进而得到决策函数 $f(\boldsymbol{x})=\mathrm{sgn}[(\boldsymbol{w}^*\cdot\boldsymbol{x})+b^*]$. 但是这种方法得到的分类面极大依赖于参数 C,并且当 C 取值较小时,可能得极端情况. 对于相同训练样本,C 取不同值时得到的分类面如图 2 所示.

从图 2 中可以看出 C 取不同值时,得到的分类面相差很大,这使得 C 的取值相当困难. 当 C 取 0.007 时的分类面将所有样本全部划分为一类中,这种极端的分类面是没有意义的.

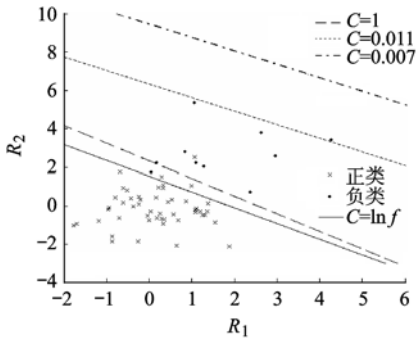

图 2 不同 C 值对应的分类面

2 C-SVM 分类面的修正

简单地将标准 SVM 中分类面的确定方法照搬到 C-SVM 中并不理想,分类精度不好控制,分类性能很不稳定,容易出现极端情况.

2.1 修正方法

从 Vapnik 最优分类面出发,将各类函数间隔为 1 的约束条件调整为保持两类的函数间隔之和为 2,令正负类的分类间隔分别为 η^+,η^-,则 $\eta^++\eta^-=2$,此时优化问题为

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

其中,N^+,N^- 分别为正负样本个数;l 为样本总个数. 式(2)对应的 Lagrange 函数为

$$\begin{aligned}L(\boldsymbol{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{r})=&\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}r_i\xi_i-\\&\sum_{y_i=1}\{\alpha_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-\eta^++\xi_i\}-\\&\sum_{y_i=-1}\{\alpha_i[-(\boldsymbol{w}\cdot\boldsymbol{x}_i)-b]-\eta^-+\xi_i\}\end{aligned} \tag{3}$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)$ 为 Lagrange 乘子;$\boldsymbol{r}=(r_1,r_2,\cdots,r_l)$ 为松弛项.

由鞍点条件得

$$\frac{\partial L}{\partial\boldsymbol{w}}=\boldsymbol{w}-\sum_{y_i=1}\alpha_i\boldsymbol{x}_i+\sum_{y_i=-1}\alpha_i x_i=\boldsymbol{w}-\sum_{i=1}^{l}\alpha_i\boldsymbol{x}_i=0 \tag{4}$$

$$\frac{\partial L}{\partial\xi_i}=C-\alpha_i-r_i=0 \tag{5}$$

$$\frac{\partial L}{\partial b}=\sum_{y_i=1}\alpha_i-\sum_{y_i=-1}\alpha_i=\sum_{i=1}^{l}\alpha_i y_i=0 \tag{6}$$

结合式(6)和 $\eta^++\eta^-=2$,得

$$\sum_{y_i=1}\alpha_i\eta^+ + \sum_{y_i=-1}\alpha_i\eta^- = \eta^+\sum_{y_i=-1}\alpha_i + \eta^-\sum_{y_i=-1}\alpha_i = 2\sum_{y_i=-1}\alpha_i = \sum_{i=1}^{l}\alpha_i \tag{7}$$

将式(4)~式(7)代入式(3)中得式(2)的对偶问题为

$$\left.\begin{aligned}&\max\ L(\boldsymbol{\alpha}) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i}^{l}\sum_{j}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\\&\text{s.t. } 0\leqslant\alpha_i\leqslant C\\&\sum_{i=1}^{l}\alpha_i y_i = 0\\&i = 1,2,\cdots,l\end{aligned}\right\} \tag{8}$$

从式(8)可得出修正后分类超平面与标准 SVM 具有相同的法向量 $\boldsymbol{w}$,只是二者的偏置 b 有所不同.

2.2　偏置 b 的求解

本文修正方法取 $\eta^+ = 1+b_m$, $\eta^- = 1-b_m$,得修正后的偏置 $b = b^* + b_m$,其中 b^* 为传统 C-SVM 的最优分类面的偏置.

将训练样本分为 4 类:正类正确划分样本 p_c、负类正确划分样本 n_c、正类错分样本 p_e 和负类错分样本 n_e,则

$$N_{p_e} + N_{n_e} = \frac{1}{2}\sum_{i=1}^{l}[1 - y_i\mathrm{sgn}(\boldsymbol{w}^*\cdot\boldsymbol{x}_i + b^* + b_m)] \tag{9}$$

$$|N_{p_e} - N_{n_e}| = \frac{1}{2}\sum_{i=1}^{l}|[y_i - \mathrm{sgn}(\boldsymbol{w}^*\cdot\boldsymbol{x}_i + b^* + b_m)]| \tag{10}$$

其中,N_{p_c},N_{n_c},N_{p_e},N_{n_e}分别为 4 类样本的个数.

考虑下列优化问题:

$$\min_{b_m}\sum_{i=1}^{l}[1 - y_i\mathrm{sgn}(\boldsymbol{w}^*\cdot\boldsymbol{x}_i + b^* + b_m)] + U\sum_{i=1}^{l}|[y_i - \mathrm{sgn}(\boldsymbol{w}^*\cdot\boldsymbol{x}_i + b^* + b_m)]| \tag{11}$$

其中,$\boldsymbol{w}^*$,b^* 为式(8)的最优解;$U>0$ 为总训练误差 $N_{p_e}+N_{n_e}$ 和正负类训练误差之差 $|N_{p_e}-N_{n_e}|$ 之间的折衷系数.

当 U 取值小时,式(11)求得的 b_m 将使最终的分类面具有最小的总训练误差,而正负类训练误差之差不一定取得最小,这适用于正负类训练样本个数相差不大,而且追求最小的总误差的情况;当 U 取值大时,b_m 将使得最终的分类面具有最小正负类训练误差之差,这适用于正负类训练样本个数不平衡的情况,保证个数少的一类也能有较高分类精度.

式(11)的目标函数最多只能取得 l+1 个值,因此可用穷举法求解. 解得最优解为 b_m.

由于 b 使得 $(N_{p_e}+N_{n_e})+U|N_{p_e}-N_{n_e}|$ 最小,所以分类面 $(\boldsymbol{w}^*, b^*+b_m)$ 对训练样本有最小的分类误差,具有最高的训练精度.

样本可分时,式(11)的解为 0,分类面未作修正,具有和标准 SVM 有相同的分类面.

3　实验仿真

3.1　稳 定 性

与图 2 对应的样本修正后的分类面如图 3 所示,其中正类样本 50 个,负类样本 10 个.

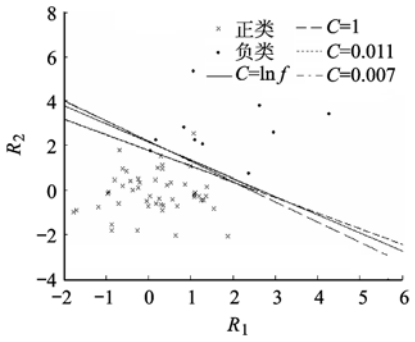

图 3　修正后的分类面

从图 3 可以看出,经修正后的分类面对于不同 C 的取值相差不大,具有很好的稳定性,且均有较高分类精度(尤其对于 C 取值小时),在一定程度上降低了 C 取值的困难程度.

3.2　抗干扰性

在二维空间生成 77 个正类样本和 10 个负类样本进行训练,引入 2 个噪声和野值数据.

从图 4 可知,传统 C-SVM 在干扰数据存在时,C 取不同值时,均得不到一个较好的分类面,对噪声和野值数据较为敏感.

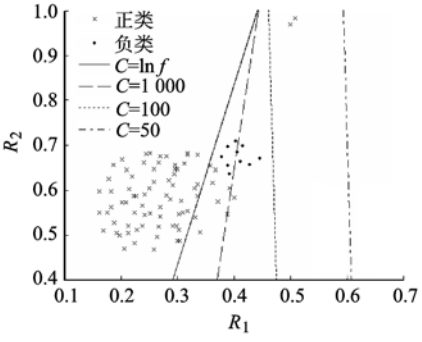

图 4　有噪声时 C-SVM 的分类面

从图 5 中可得,经过本文方法进行分类面修正后,当 C 为 100 和 50 时均有相对较好的分类

面. 这说明本文的分类面修正方法较传统 C-SVM 具有更优的抗噪声和野值数据干扰能力.

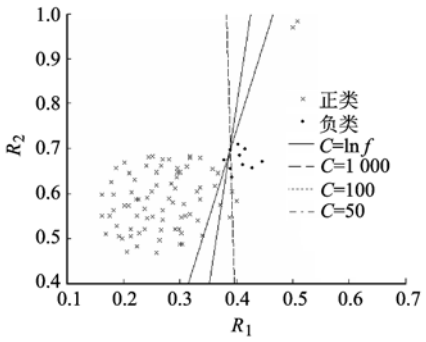

图 5 有噪声时修正后的分类面

3.3 分类精度

以 50 个中心点为[0,0]、协方差阵为 [1,0;0,1]的正类点;10 个中心点为[2,2]、协方差阵为[1,0;0,1]的负类点作为训练样本. 以正负类各 1 000个(共 2 000 个点)作为测试样本,比较 C 在[0.001,1 000]区间取不同值时,文献[5]中方差修正法和本文修正方法的分类误差,实验结果如图 6 所示. 加入正类噪声点(6,7)后,实验结果如图 7 所示. 从图 6 和图 7 中可得,本文修正方法的分类精度、分类性能的稳定性均较方差修正法好,特别是当样本中存在噪声时本文修正方法仍然具有很高的分类精度.

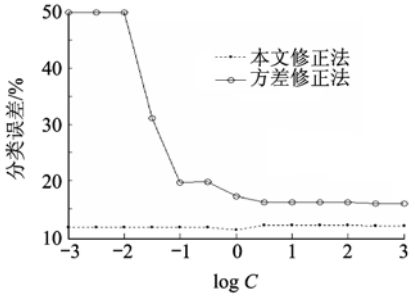

图 6 无噪声不同 C 值分类误差比较

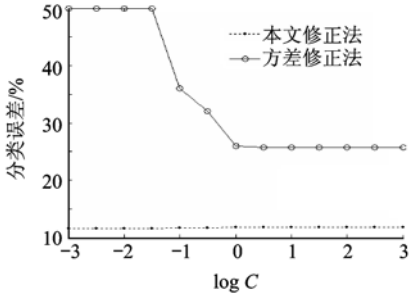

图 7 有噪声不同 C 值分类误差比较

4 结 束 语

基于SVM的原始模型,同时考虑最小化总体分类误差和正负类误差之差,提出一种新的分类面的修正方法,并指出其理论依据.

该修正方法在仅增加很少运算量的前提下得到误差最小的分类面,解决了传统 C-SVM 中的分类面依赖于 C 的取值、分类性能很不稳定及对噪声和野值点干扰敏感等问题. 实验证明该修正方法相比于其他修正方法也具有较高的分类精度和抗干扰能力.

本文的修正方法仅对训练样本不可分时具有修正作用,而不影响线性可分时标准 SVM 的最优分类面. 另外通过调整参数 U 可以在最小化总体分类误差和正负类误差之差两者之间进行取舍,以适用于不同场合.

参考文献(References)

[1] Vapnik V N. Statistical learning theory [M]. New York: Wiley, 1998

[2] Yang Chanyun, Yang J Y, Wang Jianjun. Margin calibration in SVM class-imbalanced learning [J]. Neurocomputing, 2009, 73(13): 397-411

[3] Yeom H G, Jang I H, Sim K B. Variance considered machines: modification of optimal hyperplanes in support vector machines [C]// IEEE International Symposium on Industrial Electronics. Seoul: IEEE, 2009: 1144-1147

[4] Enrico F, Marco M. Implementing reliable learning through reliable support vector machines [C]//Proceedings 2011 IEEE Symposium on Foundations of Computational Intelligence. Paris: IEEE, 2011: 100-106

[5] Zhao Xing, Zhou Yanquan, He Huacan. Researches on algorithm for confidence evaluation and decision modification of SVM [C]// International Conference on Natural Language Processing and Knowledge Engineering. Dalian: NLPKE, 2009: 1-7

[6] 周皓,李少洪. SVM 最优分类面相对位置的修正[J]. 北京航空航天大学学报,2009,35(11):1302-1305
Zhou Hao, Li Shaohong. Relative position modification of SVM's optimal hyperplane [J]. Journal of Beijing University of Aeronautics and Astronautics, 2009, 35(11): 1302-1305 (in Chinese)

[7] Fung G, Mangasarian O L. Proximal support vector machine classifiers [C] // Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2001: 77-86

[8] 刘芸,唐发根,林广艳. 一种改进的近似支持向量机算法[J]. 北京航空航天大学学报,2007,33(9):1090-1093
Liu Yun, Tang Fagen, Lin Guangyan. Robust proximal support vector machine [J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, 33(9): 1090-1093 (in Chinese)

[9] Hao Peiyi. New support vector algorithms with parametric insensitive/margin model [J]. Neural Networks, 2010, 23(1): 60-73

(编 辑:刘登敏)