复杂分类问题支持向量机的简化

方景龙[1], 陈 铄[1], 潘志庚[2], 梁荣华[3]
(1. 杭州电子科技大学图形图象研究所, 浙江杭州 310018; 2. 浙江大学 CAD & CG 国家重点实验室, 浙江杭州 310027;
3. 浙江工业大学信息工程学院, 浙江杭州 310014)

摘 要: 对于复杂分类问题, 不可避免的会有错分情况, 此时支持向量机的支持向量较多, 影响了识别速度. 为了解决这个问题, 我们提出了基于最小错分间隔的分类思想, 并在此基础上得出了一种新的简化支持向量机. 与普通支持向量机相比, 这种简化支持向量机有较少的支持向量、较高的识别速度, 而且实验结果表明, 它的识别精度完全可以与普通支持向量机的识别精度相媲美, 甚至更优.

关键词: 支持向量机; 模式识别; 支持向量缩减
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2007) 05-0858-04

A Simplification to Support Vector Machine for Complicated Recognition Problem

FANG Jing-long[1], CHEN Shuo[1], PAN Zhi-geng[2], LIANG Rong-hua[3]
(1. Institute of Graphics and Image, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China;
2. State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou, Zhejiang 310027, China;
3. College of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, China)

Abstract: For complicated recognition problem, the number of support vectors is large and recognition speed is low, because some sample were divided into section by error this time. To solve this problem, a method is bought to simplify the support vector machines based the minimal misestimate margin idea. Experiments show that this new support vector machine not only reduces the number of support vectors and recognition time but also has the same accuracy as(even better than) traditional support vector machine.

Key words: support vector machine; pattern recognition; support vector pruning

1 引言

因为支持向量机(Support Vector Machine, 简称 SVM)的复杂度取决于支持向量数目, 因此降低支持向量数目、简化支持向量机是一项非常有意义的工作. 例如, Burges[1]通过在给定精度损失下生成支持向量缩减集合的方式得到简化支持向量机, 它的计算量很大且在简化的同时牺牲了识别精度. Lee 等[2] 提出 RSVM (Reduced Support Vector Machine), Lin 等[3] 对这种方法作了进一步的研究, 该方法人为地将支持向量限制在一个训练样本子集中, 它只有在训练样本很多且支持向量所占比例极高的情况下能减少支持向量, 在一般情况下支持向量个数反而可能会增加. Scholkopf 等人[4,5] 提出了所谓的 ν-SVM, 证明了参数 ν与支持向量数目及误差之间的关系, 但这种方法在参数 ν过小时将降低机器泛化能力.

在国内, 刘向东和陈兆乾[6] 提出了一种快速支持向量机分类算法 FCSVM, 实验表明在几乎不损失识别精度的情况下识别速度有不同程度的提高. 该方法通过变换, 用少量的支持向量代替全部支持向量进行分类计算, 但在求解变换矩阵时需要求解复杂优化问题. 李红莲等[7] 提出了一种大规模样本集的支持向量机学习策略: 首先用小规模的样本集训练得到初始分类器, 然后用这个分类器对大规模训练集进行修剪, 得到一个规模很小的约减集, 再用这个约减集进行训练得到最终的分类器. 实验表明这种方法既减少了训练时间, 也减少了识别速度, 且在最优阈值时识别率还可能有所提高, 但在计算时需对阈值进行选择.

对于复杂分类问题, 由于有错分样本, 而错分样本一定是支持向量, 再加上邻近分类面的样本也都是支持向量, 所以普通支持向量机的支持向量较多, 从而影

收稿日期: 2006-05-11; 修回日期: 2006-07-28
基金项目: 国家自然科学基金重点项目(No. 60533080); 国家 973 预研项目(No. 2005CCA04400); 浙江省自然科学基金重点项目(No. Z603262)

响了识别速度. 为了解决这个问题, 我们提出了基于最小错分间隔的分类思想, 并对样本进行二次训练, 在二次训练时目标函数是最小错分间隔与机器泛化能力的折中而不是原来的最少错分样本与机器泛化能力的折中. 由这个二次训练得到一种新的简化支持向量机, 这种简化支持向量机与普通支持向量机相比有较少的支持向量、较高的识别速度, 而且实验结果表明, 它的识别精度完全可以与普通支持向量机的识别精度相媲美, 甚至更优.

2　简化支持向量机

支持向量机考虑下面优化问题:

$$\min_{w,b,N}\ \frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{L}N_i \qquad (1)$$
$$\text{s.t.}\quad y_i(w^{\mathrm{T}}\varphi(x_i)+b)\geq 1-N_i,\ N_i\geq 0,\ i=1,2,3,\cdots,L$$

它的目标函数折中考虑了最少错分样本和机器泛化能力(与 $w^{\mathrm{T}}w$ 有关), 其中 $C>0$ 是惩罚系数, 它控制对错分样本惩罚的程度, $\varphi(x_i)$ 是特征映射, x_i 是训练样本, y_i 是 x_i 的标记, 为 1 或 -1.

在实际求解中, 不是求解问题(1), 而是求解它的对偶问题:

$$\max_{A}\ \sum_{i=1}^{L}A_i-\frac{1}{2}\sum_{i,j=1}^{L}A_iA_jy_iy_jK(x_i,x_j) \qquad (2)$$
$$\text{s.t.}\quad \sum_{i=1}^{L}y_iA_i=0,\quad 0\leq A_i\leq C,\ i=1,2,3,\cdots,L$$

其中 $K(x_i,x_j)=\varphi(x_i)^{\mathrm{T}}\varphi(x_j)$ 是核函数. 这是一个二次凸规划问题, 其局部最优解就是全局最优解.

支持向量机的分类判别函数是 $f(x)=\mathrm{sgn}(w^{\mathrm{T}}\varphi(x)+b)=\mathrm{sgn}(\sum_{i\in SV}y_iA_iK(x_i,x)+b)$, 其中 SV 是支持向量(即不等于 0 的 A_i 对应的样本)下标集合. 显然, 支持向量个数越多, 支持向量机越复杂, 识别速度越慢. 下面讨论影响支持向量个数的因素.

根据优化理论的 KKT 条件, 对于问题(1)和(2), 我们有

$$A_i[y_i(w^{\mathrm{T}}\varphi(x)+b)-1+N_i]=0,\quad i=1,2,3,\cdots,L$$

所以, 当 A_i 不等于 0 时必有 $y_i(w^{\mathrm{T}}\varphi(x)+b)-1+N_i=0$. 由于每个等式中有一个变量 N_i, 所以 $y_i(w^{\mathrm{T}}\varphi(x)+b)-1+N_i=0$ 成立的可能性较大, 从而普通支持向量机的支持向量较多. 显然, 如果我们减少变量 N_i 的个数, 使得 $A_i[y_i(w^{\mathrm{T}}\varphi(x)+b)-1+N_i]=0$ 不能轻易成立, 有望降低支持向量个数.

为此, 我们采用下面的优化问题对样本进行二次训练, 由此定义基于最小错分间隔的广义最优分类超平面, 得出一个新的简化支持向量机, 用来减少支持向量数目:

$$\min_{w,b,N}\ \frac{1}{2}w^{\mathrm{T}}w+CN \qquad (3)$$
$$\text{s.t.}\quad y_i(w^{\mathrm{T}}\varphi(x_i)+b)\geq 1,\ i\in V-MV;$$
$$y_i(w^{\mathrm{T}}\varphi(x_i)+b)\geq 1-N,\ i\in MV,\ N\geq 0$$

它的目标函数折中考虑了最小错分间隔和机器泛化能力, 这里 MV 是一次训练形成的错分样本下标集合, V 是二次训练样本下标集合, 可以将 V 取为整个训练样本的下标集合, 也可以为了减少二次训练时间, 将 V 取为一次训练的支持向量的下标集合, 但不管怎样, 应有 $MV\subseteq V$.

问题(3)的对偶问题是如下二次凸规划问题:

$$\max_{A}L=\sum_{i\in V}A_i-\frac{1}{2}\sum_{i,j\in V}A_iA_jy_iy_jK(x_i,y_i) \qquad (4)$$
$$\text{s.t.}\quad \sum_{i\in V}y_iA_i=0,\quad \sum_{i\in MV}A_i\leq C,\ A_i\geq 0,\ i\in V$$

根据优化理论的 KKT 条件, 对于普通支持向量机, 若 $A_i=0$, 则 $y_i(w^{\mathrm{T}}\varphi(x_i)+b)\geq 1$; 若 $A_i>0$, 则 $y_i(w^{\mathrm{T}}\varphi(x_i)+b)\leq 1$. 所以, 普通支持向量机的支持向量是使 $y_i(w^{\mathrm{T}}x+b)\leq 1$ 成立的几乎所有样本, 包括所有错分样本、超平面 $w^{\mathrm{T}}\varphi(x)+b=-1$ 和 $w^{\mathrm{T}}\varphi(x)+b=1$ 之间的所有样本以及超平面 $w^{\mathrm{T}}\varphi(x)+b=-1$ 和 $w^{\mathrm{T}}\varphi(x)+b=1$ 上面的部分样本. 而简化支持向量机的支持向量是使 $y_i(w^{\mathrm{T}}x+b)=1$ 成立的部分样本, 和使 $y_i(w^{\mathrm{T}}x+b)=\min_j y(w^{\mathrm{T}}x_j+b)$ 成立的所有样本, 它仅仅包括错分样本中离分类超平面最远的样本, 以及超平面 $w^{\mathrm{T}}\varphi(x)+b=-1$ 和 $w^{\mathrm{T}}\varphi(x)+b=1$ 上面的部分样本.

显然, 简化支持向量机的支持向量要比普通支持向量机的支持向量少. 另外可明显看出, 问题(3)给出的解, 其对训练样本的识别率必高于等于问题(1)对训练样本的识别率. 在测试样本识别率上, 由于问题(3)是用错分间隔代替问题(1)的错分样本离分类超平面距离之和与机器泛化能力进行折中, 所以一般情况下, 问题(3)对测试样本的识别率也比问题(1)的要高.

由于简化支持向量机的支持向量仅在 4 个超平面上, 所以我们称这一简化支持向量机为超平面支持向量机, 简记为 HPSVM. 特别地, 在本文用 HPSVM1 表示二次训练采用整个训练样本的 HPSVM; 用 HPSVM2 表示二次训练采用一次训练的支持向量的 HPSVM.

3　简化支持向量机的求解

可以参照 SMO 方法及其修改方法[8,9] 来获得 HPSVM 的求解算法, 首先是要选取 2 个样本的工作集. 因为每次迭代以对 KKT 条件破坏最多的两个 Lagrange 乘子为工作集, 因此, 根据最优化理论的 KKT 条件, 对于 HPSVM 我们选取以下两个下标对应的样本组成工作集:

$$s \in \arg\max(\{y_i - w^T\phi(x_i) \mid \alpha_i \geq 0, y_i = 1, i \in V - MV\}, \{y_i - w^T\phi(x_i) \mid \alpha_i > 0, y_i = -1, i \in V - MV\}, \{y_i - w^T\phi(x_i) \mid \sum_{j \in MV} \alpha_j < C, y_i = 1, i \in MV\}, \{y_i - w^T\phi(x_i) \mid \alpha_i > 0, y_i = -1, i \in MV\})$$

$$t \in \arg\max(\{y_i - w^T\phi(x_i) \mid \alpha_i \geq 0, y_i = 1, i \in V - MV\}, \{y_i - w^T\phi(x_i) \mid \alpha_i > 0, y_i = -1, i \in V - MV\}, \{y_i - w^T\phi(x_i) \mid \alpha_i > 0, y_i = -1, i \in MV\}, \{y_i - w^T\phi(x_i) \mid \sum_{j \in MV} \alpha_j < C, y_i = 1, i \in MV\})$$

其次是给出解规模为2的优化问题解析解: 在对偶问题(4)中, 将 α_s, α_t 看成待求变量, 其他的看成已知参数, 得到求解 α_s, α_t 的优化问题如下:

$$\max_{\alpha_s, \alpha_t} L = \alpha_s + \alpha_t - \frac{1}{2}(k_{ss}\alpha_s^2 + k_{tt}\alpha_t^2 + 2sk_{st}\alpha_s\alpha_t) - \left(\sum_{i \in Xs, t} \alpha_i y_i \phi(x_i)\right)(y_s\phi(x_s)\alpha_s + y_t\phi(x_t)\alpha_t) + Const \tag{5}$$

$$\text{s.t.} \quad \alpha_s + s\alpha_t = C, \sum_{i \in MV} \alpha_i \leq C, \alpha_s, \alpha_t \geq 0$$

其中 $s = y_s y_t$, $r = \alpha_s^{old} + s\alpha_t^{old}$, $k_{i,j} = k(x_i, x_j)$.

不考虑不等式约束, 则上面的优化问题与SVM方法的一样, 其解析解如下所示:

$$\alpha_t = \alpha_t^{old} + \frac{y_t(g_s^{old} - g_t^{old})}{G}, \quad \alpha_s = C - s\alpha_t$$

这里 $G = 2k_{st} - k_{ss} - k_{tt}$, $g_s^{old} = y_s - \phi(x_s)^T w^{old}$, $g_t^{old} = y_t - \phi(x_t)^T w^{old}$, $w^{old} = \sum_{i \in SV} y_i \alpha_i^{old} \phi(x_i)$.

考虑不等式约束后需对上面求解的 α_s, α_t 进行裁减. 下面分两种情况进行讨论:

(1) $s = 1$ 的情况, 此时 $\alpha_s + \alpha_t = C$. 首先保证 α_s, $\alpha_t \geq 0$, 于是若 $\alpha_s < 0$, 则 $\alpha_s = 0$, $\alpha_t = C$; 若 $\alpha_t < 0$, 则 $\alpha_t = 0$, $\alpha_s = C$. 其次保证 $\sum_{i \in MV} \alpha_i \leq C$, 于是

当 $s \in MV$, $t \notin MV$ 时, 若 $\alpha_s > C - \sum_{i \in MV, i \neq s} \alpha_i^{old}$, 则 $\alpha_s = C - \sum_{i \in MV, i \neq s} \alpha_i^{old}$, $\alpha_t = C - \alpha_s$;

当 $s \notin MV$, $t \in MV$ 时, 若 $\alpha_t > C - \sum_{i \in MV, i \neq t} \alpha_i^{old}$, 则 $\alpha_t = C - \sum_{i \in MV, i \neq t} \alpha_i^{old}$, $\alpha_s = C - \alpha_t$;

当 $s \in MV$, $t \in MV$ 时, 则因为有 $\alpha_s + \alpha_t = \alpha_s^{old} + \alpha_t^{old}$, 所以 $\sum_{i \in MV} \alpha_i$ 的值不会改变, 因此这种情况无需对 α_s, α_t 进行剪裁;

当 $s \notin MV$, $t \notin MV$ 时, 显然 $\sum_{i \in MV} \alpha_i$ 的值不会改变, 因此这种情况也无需对 α_s, α_t 进行剪裁.

(2) $s = -1$ 的情况, 此时有 $\alpha_s - \alpha_t = C$. 首先保证 α_s, $\alpha_t \geq 0$, 于是若 $\alpha_s < 0$, 则 $\alpha_s = 0$, $\alpha_t = -C$; 若 $\alpha_t < 0$, 则 $\alpha_t = 0$, $\alpha_s = C$. 其次保证 $\sum_{i \in MV} \alpha_i \leq C$, 于是

当 $s \in MV$, $t \notin MV$ 时, 若 $\alpha_s > C - \sum_{i \in MV, i \neq s} \alpha_i^{old}$, 则 $\alpha_s = C - \sum_{i \in MV, i \neq s} \alpha_i^{old}$, $\alpha_t = \alpha_s - C$;

当 $s \notin MV$, $t \in MV$ 时, 若 $\alpha_t > C - \sum_{i \in MV, i \neq t} \alpha_i^{old}$, 则 $\alpha_t = C - \sum_{i \in MV, i \neq t} \alpha_i^{old}$, $\alpha_s = \alpha_t + C$;

当 $s \in MV$, $t \in MV$ 时, 若 $\alpha_s + \alpha_t > C - \sum_{i \in MV, i \neq s, t} \alpha_i^{old}$, 则 $\alpha_s = 0.5(C - \sum_{i \in MV, i \neq s, t} \alpha_i^{old} + C)$, $\alpha_t = 0.5(C - \sum_{i \in MV, i \neq s, t} \alpha_i^{old} - C)$;

当 $s \notin MV$, $t \notin MV$ 时, 显然 $\sum_{i \in MV} \alpha_i$ 的值不会改变, 因此这种情况也无需对 α_s、α_t 进行剪裁.

4 实验结果

我们在两个公共测试数据库UCI Adult 和Web 数据集上对SVM 和HPSVM 方法进行了对比计算, 总共计算了17个例子, 样本的具体情况见文献[8], 计算结果见表1 和表3. 计算采用径向基核函数 $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{R^2}\right)$, 核函数参数和支持向量机惩罚参数均

表1 SVM和HPSVM计算结果的比较(UCI Adult 数据集)

Dataset	Number of support vectors			Recognition time(ms)			Recognition rate(%)		
	SVM	HPSVM1	HPSVM2	SVM	HPSVM1	HPSVM2	SVM	HPSVM1	HPSVM2
A1a	785	63(8%)	63(8%)	13688	1844(13.5%)	1828(13.5%)	82.6689	82.6496	82.6496
A2a	1105	72(6.5%)	72(6.5%)	18937	1954(10.3%)	1922(10.1%)	83.4599	83.4037	83.4037
A3a	193	14(7.3%)	12(6.2%)	3734	1000(26.8%)	937(25.1%)	75.9395	75.9395	75.9375
A4a	455	29(6.4%)	29(6.4%)	7390	1141(15.4%)	1141(15.4%)	78.7077	79.8128	78.8128
A5a	1878	88(4.7%)	88(4.7%)	28234	1891(6.7%)	1906(6.8%)	83.6578	83.696	83.696
A6a	2044	103(5%)	103(5%)	24297	1719(7.1%)	1704(7.1%)	83.8011	83.8152	83.8152
A7a	4174	150(3.6%)	150(3.6%)	39703	1782(4.5%)	1735(4.4%)	84.2112	84.0593	84.0593
A8a	649	46(7.1%)	46(7.1%)	3657	500(13.7%)	500(13.7%)	81.997	82.4937	82.4937
A9a	1998	111(5.6%)	111(5.6%)	18141	1375(7.6%)	1359(7.5%)	84.2823	84.2516	84.2516

表 2　SVM 和 HPSVM 计算结果的比较(Web Data Sets)

Dataset	Number of support vectors			Recognition time(ms)			Recognition rate(%)		
	SVM	HPSVM1	HPSVM2	SVM	HPSVM1	HPSVM2	SVM	HPSVM1	HPSVM2
W1a	233	65(25.7%)	48(20.6%)	6610	2890(43.7%)	2407(36.4%)	97.3282	97.3811	97.3705
W2a	217	71(32.7%)	60(27.6%)	5891	2859(48.5%)	2578(43.8%)	97.6231	97.4805	97.3098
W3a	277	87(31.4%)	77(27.8%)	6812	3094(45.4%)	2781(40.8%)	97.5534	97.0248	95.8695
W4a	387	105(27.1%)	89(23%)	8422	3329(39.5%)	2907(34.5%)	97.3504	96.3146	95.5006
W5a	181	111(61.3%)	52(28.7%)	4469	3343(74.8%)	1968(44%)	65.2919	71.0494	67.4544
W6a	194	102(52.6%)	52(26.8%)	3641	2532(69.5%)	1640(45%)	79.3864	79.9177	80.9865
W7a	1484	193(13%)	173(11.7%)	17937	3125(17.4%)	2969(16.6%)	98.3597	98.3358	98.2879
W8a	2002	243(12%)	220(11%)	13703	2313(16.9%)	2172(15.9%)	97.358	97.1708	97.1574

采用文献[8] 中的数据, 即对于 UCI Adult, 核参数取 0.005, 惩罚参数取 1; 对于 Web 数据集, 核参数取 01005, 惩罚参数取 5, 但此时 W5a 和 W6a 两组数据的识别率只有 1719449% 和 5415069%, 明显不是合理的参数选择, 为此对这两组数据, 在这里核参数取 01005, 惩罚参数取 350.

从表 1、2 可以看出, 对于这 17 个算例, HPSVM1、HPSVM2 的支持向量明显比 SVM 要少得多, 最低只有其 316%, 从而使得 HPSVM1、HPSVM2 的识别时间比 SVM 的也要少得多, 最低只有其 414%. HPSVM1 的识别率最低的比 SVM 低 110358%, 最高的比 SVM 高 517575%, HPSVM2 的识别率最低的比 SVM 低 1.8499%, 最高的比 SVM 高 2.1625%, 总体上看, HPSVM 的识别率与 SVM 的相当.

5　结束语

针对普通支持向量机在处理复杂分类问题时出现的支持向量数较多, 识别速度较慢的问题, 我们提出了基于最小错分间隔的的分类思想, 并对样本进行二次训练, 根据这个二次训练得到一种新的简化支持向量机 HPSVM. 与普通支持向量机相比, HPSVM 有较少的支持向量、较高的识别速度, 而且实验结果表明, 它的识别精度完全可以与普通支持向量机的识别精度相媲美, 甚至更优, 因而是一个高性能的支持向量机.

但是 HPSVM 对支持向量的减少程度与一次训练的错分样本数有关, 错分样本越多, HPSVM 的效果越明显, 如果错分样本很少, 则 HPSVM 的效果不明显, 所以说 HPSVM 是一种针对复杂分类问题设计的支持向量机.

这一简化思想和方法对回归型支持向量机同样适用, 我们将另文叙述.

致谢　台湾大学计算机科学与信息工程系 Chang Chih2Chun 和 Lin Chih2Jen 的软件 LIBSVM 为我们的方法提供了一个平台, 在此向他们表示感谢.

参考文献:

[1] C J C Burges. Simplified support vector decision rule[A]. Proc 13th Int Conf Machine Learning[C]. San Mateo, CA, 1996. 71 - 77.

[2] Y J Lee, O L Mangasarian. RSVM: Reduced support vector machines[A]. Proc of the First SIAM International Conference on Data Mining[C]. Chicago, 2001.

[3] K M Lin, C J Lin. A study on reduced support vector machines [J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1449 - 1559.

[4] B Scholkopf, et al. New support vector algorithms[J]. Neural Computation; 2000, 12(5): 1207- 1245.

[5] P H Chen, C J Lin, B Sch lkopf. A tutorial on T2support vector machines[J]. Applied Stochastic Models in Business and Indus2try, 2005, 21(2): 111- 136.

[6] 刘向东, 陈兆乾. 一种快速支持向量机分类算法的研究[J]. 计算机研究与发展, 2004, 41(8): 1327- 1332.
Xiang2dong Liu, Zhao2qian Chen. A fast classification algorithm of support vector machines[J]. Journal of Computer Research and Development, 2004, 41(8): 132721332. (in Chinese)

[7] 李红莲, 等. 针对大规模训练集的支持向量机的学习策略[J]. 计算机学报, 2004, 27(5): 716- 719.
Hong2lian Li, et al. A learning strategy of SVM used to large training set[J]. Chinese Journal of Computers, 2004, 27(5): 716- 719. (in Chinese)

[8] J C Platt. Fast training of support vector machines using se2quential minimal optimization[A]. In Bernhard Sch lkopf, et al. editors, Advances in Kernel Methods2Support Vector Learn2ing, Cambridge[M]. MA: MIT Press, 1999. 185- 208.

[9] S Keerthi, et al. Improvements to Platt. s SMO algorithm for SVM classifier design[J]. Neural Computation, 2001, 13(3): 637- 649.

作者简介:

方景龙　男, 1964 年生于江西景德镇, 研究员, 主要研究领域为图像处理与模式识别、支持向量机. E2mail: fjl@hdu. edu. cn