云南大学学报（自然科学版），2006，28（2）：93~97　　CN 53-1045/N　ISSN 0258-7971
Journal of Yunnan University

基于正交表的支持向量机并行学习算法*

邱宇青[1]，胡光华[1]，潘文林[2]
（1. 云南大学 数学系，云南 昆明　650091；2. 哈尔滨工程大学 机电学院，黑龙江 哈尔滨　150001）

摘要：对大规模训练样本的支持向量机训练问题进行探索，提出了一种基于正交表的并行学习算法. 这种方法通过求解一些相互独立的小的训练问题来求解大的训练问题，采用多处理机可求解大规模的训练问题.

关键词：正交表；并行计算；SVM；HBSVM
中图分类号：TP 183　　文献标识码：A　　文章编号：0258-7971(2006)02-0093-05

支持向量机（Support Vector Machine，简称 SVM）是在统计学习理论[1]基础上发展起来的一种新型的机器学习方法. 由于它是基于小样本的统计学习理论，在许多方面表现出了优越的性能. 根据统计学习理论结构风险最小化原则，为了最小化期望风险的上界，SVM 在固定机器学习经验风险的条件下最小化 VC 维置信度. SVM 的求解采用求解拉格朗日乘子的优化方法，最后归结为求解二次函数的极值问题. 若采用经典的二次规划方法，难以求解大规模的学习问题. 针对大规模的 SVM 学习问题，本文提出一种并行的学习算法，把大规模的学习问题转化为一些可同时并行计算的小规模学习问题，用多个处理机同时计算以达到提高速度的目的. 通过实验，可以看到此方法的可行性.

1　支持向量机

对模式分类问题，已知观测数据样本：$P = \{(x_1, y_1), \cdots, (x_l, y_l)\}$，$x \in R^n$，$y \in \{1, -1\}$，根据统计学习理论，得到 SVM 的求解为下面的优化问题

$$\begin{aligned} &\min \quad \frac{1}{2}\alpha Q^{\mathrm{T}}\alpha - e^{\mathrm{T}}\alpha, \\ &\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant C,\ i = 1, \cdots, l,\ y^{\mathrm{T}}\alpha = 0, \end{aligned} \tag{1}$$

其中 $Q_{ij} = y_i y_j K(x_i, x_j)$，$K(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j)$ 为核函数. 分类判别函数为

$$f(x) = \operatorname{sign}\left(\sum_{i \in S}\alpha_i y_i K(x, x_i) + b\right). \tag{2}$$

记 $S = \{i \mid \alpha_i > 0\}$，则称 S 为支持向量的指标集，$SV = \{(x_i, y_i), i \in S\}$ 为支持向量集，$SV \subseteq P$.

在样本数量不多时，求解(1)可用传统的二次规划方法. 但样本数量比较多时，求解比较困难，消耗的时间和空间较大. 许多学者对寻找更有效的方法进行了研究，提出了一些算法，如分解算法[2]、SVMlight 算法[3]、SMO 算法[4]、近邻 SVM 算法[5]等.

由于支持向量机得到的判别函数只与支持向量有关，如果只取支持向量作为训练样本集，得到的判别函数与所有样本作为训练样本集得到的判别函数是一致的. 本文提出的算法就是基于这一思想.

命题 1　设集合 D，满足 $SV \subseteq D \subseteq P$，若以 P 为训练样本集的 SVM 问题的解是唯一的，以 D 为训练样本集的 SVM 问题的解也是唯一的，则以 P 为训练样本集求解 SVM 得到的判别函数(2)与以 D 为训练样本集得到的判别函数相同.

* 收稿日期：2005-04-25
基金项目：国家自然科学基金资助项目(10271103)；云南大学理(工)科校级科研资助项目(2002Q019SL).
作者简介：邱宇青(1981-　)，男，云南人，硕士生，主要从事智能学习算法方面的研究.
通讯作者：胡光华(1962-　)，男，博士，教授，主要从事智能学习算法方面的研究.

证明 以 $D = \{(x_i^*, y_i^*)\}$, $i = 1, \cdots, r$, $r \leqslant l$ 为训练样本集, 求解 SVM, 得到如下规划

$$\min \quad \frac{1}{2}\alpha^* (Q^*)^{\mathrm{T}}\alpha^* - (t^*)^{\mathrm{T}}\alpha^*,$$
$$0 \leqslant \alpha_i^* \leqslant C,\ i = 1, \cdots, l,\ (y^*)^{\mathrm{T}}\alpha^* = 0, \tag{3}$$

其中 $Q_{ij} = y_i^* y_j^* K(x_i^*, x_j^*)$.

则判别函数为

$$g(x) = \operatorname{sign}\left(\sum_{i \in S^*} \alpha_i^* y_i^* K(x, x_i^*) + b^*\right), \tag{4}$$

其中 S^* 为支持向量的指标集, 即 $S^* = \{i \mid a_i^* > 0\}$.

设 α 为(1)的解, 则 α 使得 $h(\alpha) = \frac{1}{2}\alpha Q^{\mathrm{T}}\alpha - e^{\mathrm{T}}\alpha$ 最小, 且满足(1)的约束条件.

令 $\alpha^* = (\alpha_1^*, \cdots, \alpha_r^*)^{\mathrm{T}}$, $\alpha_j^* = \alpha_i$, 若 $x_i, x_j \in D$, $x_i = x_j$, 则 $h^*(\alpha^*) = \frac{1}{2}\alpha^* (Q^*)^{\mathrm{T}}\alpha^* - (e^*)^{\mathrm{T}}\alpha^* = h(\alpha)$, 并且满足(3)的约束条件. 下面证明 α^* 使得(3)式最小.

假设存在 $\tilde{\alpha}^* \neq \alpha^*$ 使得(3)式达到最小值, 则有 $\bar{h}^*(\alpha^*) > h(\tilde{\alpha}^*)$. 令 $\tilde{\alpha} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_l)^{\mathrm{T}}$,

$$\begin{cases} \tilde{\alpha}_j^*, & \text{if } x_i, x_j \in D,\ x_i = x_j, \\ 0, & \text{if } x_i \notin D. \end{cases}$$

所以 $h\bar{\alpha} = h^*(\tilde{\alpha}^*) < h^*(\alpha^*) = h(\alpha)$, 这与 α 使得 $h(\alpha)$ 最小矛盾. 因此, α^* 使得(3)式最小.

反之若 α^* 使得(3)式最小, 令 $\alpha = (\alpha_1, \cdots, \alpha_l)^{\mathrm{T}}$, $\alpha_i = \begin{cases} \alpha_i^*, & \text{if } x_i, x_j \in D,\ x_i = x_j, \\ 0, & \text{if } x_i \notin D, \end{cases}$ 同理可证 α 使得(1)式最小.

综上可知(2)和(4)相同. 证毕.

推论 **1** 以 P 为训练样本集求解 SVM 得到的判别函数(2)与以 SV 为训练样本集得到的判别函数相同.

推论 1 是命题 1 的极端情况, 取 $D = SV$, 则推论 1 得证.

为了简化约束条件, Friess 提出了 BSVM(Bounded SVM)算法[6]

$$\min \quad \frac{1}{2}\alpha(Q + yy^{\mathrm{T}})^{\mathrm{T}}\alpha - e^{\mathrm{T}}\alpha,$$
$$\text{s. t. } 0 \leqslant \alpha_i \leqslant C,\ i = 1, \cdots, l. \tag{5}$$

这种方法提高了收敛速度, 而且推广误差也可以接受. 此外, 命题 1 及推论 1 对 BSVM 的情况也适用.

2 基于正交表的并行学习算法

文献[7]中介绍了按正交表做试验比较极差的方法, 可筛选出对模型影响大的重要因素. 受正交试验设计思想的启发, 支持向量机的训练学习就是要找出支持向量, 我们可以用正交设计的方法找出那些对支持向量机的影响最为明显的样本即支持向量机.

对模式分类问题, 对于 SVM, 设 $t = (t_1, \cdots, t_l)^{\mathrm{T}}$, 其中 $t_i \in \{0, 1\}$, $i = 1, \cdots, l$; $r = \sum_{i=1}^{l} t_i$, $G = \{i \mid t_i = 1\}$, 令

$$f(t) = \min_{\alpha}\left[\frac{1}{2}\alpha(t)^{\mathrm{T}} Q(t)\,\alpha(t) - e^{\mathrm{T}}\alpha(t)\right], \tag{6}$$

满足条件: $0 \leqslant \alpha(t)_i \leqslant C$, $y^{\mathrm{T}}(t)\alpha(t) = 0$, 其中 $\alpha(t) = (\alpha_1, \cdots, \alpha_r)$, $y(t) = (y_1^*, \cdots y_r^*)$, $y_j^* = y_i$, $i \in G$;

$$Q_{ij}(t) = y_i y_j K(x_i, x_j),\ i, j \in G.$$

由命题 1, 若 $G \supseteq S$, 则(6)与(1)同解.

命题 **2** 令 $e = (1, \cdots, 1)^{\mathrm{T}} \in R^l$; $u(\theta) = (u_1, \cdots, u_l)^{\mathrm{T}}$, 其中 $\theta = \{1, 2, \cdots, l\}$, $u_i = \begin{cases} 1, & i \neq \theta, \\ 0, & i = \theta, \end{cases}$

则 $R^* = |f(e) - f(u(\theta))| \geqslant 0$, 当且仅当 $\theta \notin S$ 取等号.

证明　由命题 1, 若 $\theta \notin S, f(e) = f(u(\theta))$ 所以, $R^* = 0$. 若 $\theta \in S$, 则 $f(e) \neq f(u(\theta))$, 所以 $R^* > 0$. 证毕.

同样对于 BSVM, 令

$$f_B(t) = \min\left[\frac{1}{2}\alpha^{\mathrm{T}}(t)(Q(t) + y(t)y^{\mathrm{T}}(t))\alpha(t) - e^{\mathrm{T}}\alpha(t)\right], \quad (7)$$

若 $D \supseteq S$, 则(7) 与(5) 同解. 命题 2 也对 BSVM 适用.

我们可把每个样本看作一个因素, 每个因素有 2 个水平, 即是或不是支持向量. 这样就是一个 l 个因素 2 水平的试验. 因此, 可以用正交表计算极差的方法找出对(6) 式影响最大的因素, 支持向量就包含在这些因素中. 在找到这些因素后, 再进行一次小规模的训练就可得到支持向量和判别函数. 为了简化约束条件, 可用(7) 代替(6), 从而提高速度. 具体的算法如下, 我们称其为 HBSVM.

算法 **1**:

(1) 产生正交表;

(2) 对正交表每一行(即每一次试验), 把低水平因素(把表的高水平对应样本的低水平, 而表的低水平对应样本的高水平) 对应的样本组成一个样本集, 用这个样本集对支持向量机进行训练, 即求解(7) 式, 记下最优的目标函数值 f_i;

(3) 计算每个因素的 $R_i = \bar{f}_{i,1} - \bar{f}_{i,0}$, $\bar{f}_{i,1}$ 为因素 i 为高水平时所有 f_i 的平均值, $\bar{f}_{i,0}$ 为因素 i 为低水平时所有 f_i 的平均值;

(4) 选取极差最大的 k 个样本组成一个样本集, 对支持向量机进行训练, 得到所有样本的最优的支持向量.

这个算法相当于每次把一定数量的不同样本集进行训练, 相当于把其它的样本的拉格朗日乘子置为零. 如果样本集包含最优的支持向量, 那么一定可以得到最优的目标函数值和支持向量. 由于每次只有一部分样本进行训练, 比所有样本训练需要的时间要少得多. 再加上 2 次试验之间是独立的, 我们可以把每一次的试验作为 1 个进程, 用并行的很多个处理机同时进行计算, 从而提高速度. 图 1 给出了并行计算的流程图, 进程 $P_1, \cdots, P_n$ 对应算法 1 的第 2 步的 n 次试验, 可由多个处理机来并行计算, 处理机越多, 速度越快. 假设有 m 个处理机, 运算所需的时间近似等于 1 个处理机所需时间的 $1/m$.

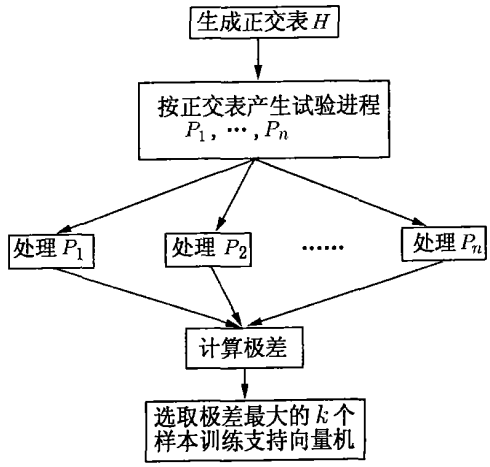

图 **1**　流程图

Fig. 1 Flow chart

3 实　验

实验的环境是: 1.7 G CPU, 256 M 内存的 PC 计算机, 软件为 MATLAB6.1. 由于没有并行计算机实现算法, 我们用串行的方式计算, 从每次试验(即每个进程) 平均所用的时间可以看到, 如果进行并行计算, 学习速度将会有很大的提高.

例 **1**　实验选用的数据为: 线性不可分的 data1(如图 2(a)), 随机生成的在 $0 \leqslant x \leqslant 2$, $0 \leqslant y \leqslant 2$ 的范围内均匀分布的 2 000 个点, 抛物线 $y = (x-1)^2 + 1$ 把它们分成 2 类; 核函数选用的是 RBF 径向基核函数 $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2)$, 取 $\sigma = 0.1$. 实验结果见表 1, HBSVM 的分类超平面见图 3(a), SVM 的分类分类超平面见图 3(b).

例 **2**　实验选用的数据为 data2(如图 2(b)), 是 NEC Research Institute 的 MNIST(http://yann.lecun.com/exdb/mnist/) 手写数字的数据库, 选用了其中的 2 000 个样本做二分类识别实验, num0 表示把手写数字 0 的样本分为一类, 其它的分为一类的分类问题; num1, num2 分别是手写数字 1 和 2 的二分类问题. 核函数选用的是 RBF 径向基核函数 $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2)$, 取 $\sigma = 1000$. 实验结果见表 2.

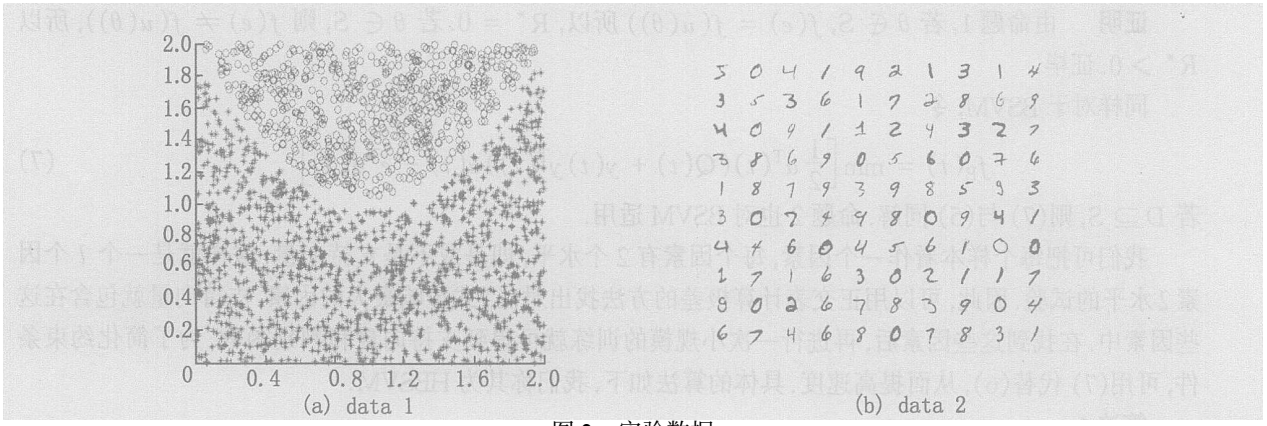

图 **2** 实验数据

Fig. 2 Data for experiment

表 **1** **Data1** 的实验结果

Tab. 1 Performance on data1

算法	训练样本数	测试样本数	正确率/%	训练时间/min	试验次数	平均每次试验时间/min
SVM	1 023	977	100	133		
HBSVM	1 023	977	100	351	1 024	0.34

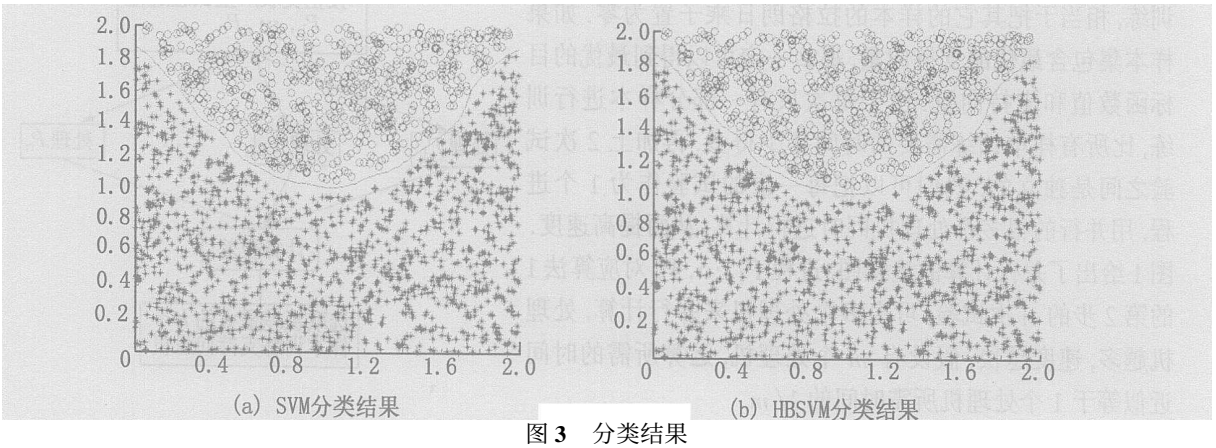

图 **3** 分类结果

Fig. 3 Performance for classification

表 **2** **Data2** 的实验结果

Tab. 2 Performance on data2

分类法	训练样本数	测试样本数	SVM		HBSVM			
			正确率/%	训练时间/min	正确率/%	训练时间/min	试验次数	平均每次试验时间/min
num0	1 000	1 000	98	45	98	107	1 024	0.11
num1	1 000	1 000	98.9	36	98.9	159	1 024	0.16
num2	1 000	1 000	96.3	33	96.5	126	1 024	0.12

4 结　论

HBSVM 是一种基于统计学习理论的改进的支持向量机学习算法, 实验证明这种算法是一种有效的方法. 由于它可以把一个大的支持向量求解问题转化为, 一些相互独立的小的支持向量求解进程, 虽然在总时间上不如 SVM, 但用多个处理机同时并行计算的话, 速度会有很大的提高, 而 SVM 不可以并行计算. 在空间复杂度方面, 每一个进程需要的内存远远小于 SVM, 这就意味着 SVM 由于内存的限制而不能求解的问题 HBSVM 可以求解. 支持向量机的求解是一个 NP 问题, 在一定程度上可以说 HBSVM 能在可以预料的时间内求解这个问题的.

本文提出的 HBSVM 算法是基于把大的规划问题化为一些相互独立的小的规划问题的思想和正交设计的筛选变量的思想, 但能否找到更好的试验表格和更快速的处理每个进程的方法是以后研究的方向.

参考文献:

[1] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer, 1998.

[2] OSUNA E, FREUND R, GIROSI F. An improved training algorithm for support vector machines[C]// Proc of NNSP' 97, 1997.

[3] JOACHIMS T. Making Large-scale SVM learning practical[C]// B Scholkopf, C J C Burges, A J Smola. Advances in Kernel Methods-support vector learning. London: Cambridge, MA, MIT Press, 1999.

[4] PLATT J. Sequential minimal optimization: A fast algorithm for training support vector machine[R]. Technical Report MSR – TR– 98– 14, Microsoft Research, 1998.

[5] 杞娴, 胡光华, 彭新俊. 基于最佳距离度量近邻法的邻域风险最小化方法[J]. 云南大学学报: 自然科学版, 2004, 26(5): 373-377.

[6] FRIESS T T, CRISTIANIMI N, CAMPBELL C. The kermel adatron algorithm: a fast and simple learning procedure for support vector machine[C]// In Proceeding of 15th Intl. Conf. Machine Learning. Burlington: Morgan Kaufman Publishers, 1998.

[7] 方开泰, 马长兴. 正交与均匀试验设计[M]. 北京: 科学出版社, 2001.

Parallel algorithm of support vector machine based on orthogonal array

QIU Yu-qing[1], HU Guang-hua[2], PAN Wen-lin

(1. Department of Mathematics, Yunnan University, Kunming 650091, China;
2. School of Mechanical & Electrical Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: Explores the training problems of support vector machine with large training pattern set, and a new parallel algorithm based on orthogonal array is presented. This method breaks a large training problem into some small independent problems. Then the large problems can be solved via solving those small problems individually. Thus we will be able to use this algorithm and the computer with many CPU to calculate large problems.

Key words: orthogonal array; parallel compute; support vector machine(SVM); HBSVM