一种改进的分枝定界半监督支持向量机学习算法

赵 莹[1],张健沛[1],杨 静[1],王冠军[2]
(1.哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001; 2.中国矿业大学计算机学院,江苏徐州 221116)

摘 要: 分枝定界半监督支持向量机,由于其实现的是全局最优化,因而可以作为其他半监督学习算法的一个基准.针对分枝定界半监督支持向量机中存在的缺陷,提出一种改进的分枝定界半监督支持向量机学习算法.该算法重新对下界的估计进行定义,从而降低了各结点计算下界的时间复杂度;同时利用支持向量机的几何特点确定分枝结点,以提高算法的运算速度.实验分析表明本文提出的算法具有精度高、鲁棒性强等优点.

关键词: 半监督学习; 支持向量机; 分枝定界; 统计学习理论
中图分类号: TP302.7 **文献标识码:** A **文章编号:** 0372-2112 (2010) 02-0449-06

An Improved Learning Algorithm for Branch and Bound for Semi-Supervised Support Vector Machines

ZHAO Ying[1], ZHANG Jian-pei[1], YANG Jing[1], WANG Guan-jun[2]
(1. *College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*;
2. *College of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*)

Abstract: Branch and bound semi-supervised support vector machines as an exact globally optimization is useful for benchmarking practical semi-supervised support vector machines implementations. An improved learning algorithm for branch and bound for semi-supervised support vector machines is presented, concerning the defects of the branch and bound for semi-supervised support vector machines. The estimations of the node lower bound are redefined, which can reduce time complexity of computing the lower bound on every node. Branching nodes are determined by using the geometric characteristic of the support vector machines, which can improve the operation speed simultaneously. Experimental results show that modified algorithm has high precision and strong robustness.

Key words: semi-supervised learning; support vector machines; branch and bound; statistic theory

1 引言

支持向量机(Support Vector Machines)是在统计学习结构风险最小化原理的基础上发展起来的一种模式识别方法,根据有限样本信息在模型的复杂性和泛化能力之间寻求最佳折衷.在实际应用中,收集大量无标签样本已相当容易,而获取大量有标签样本则需要耗费大量的人力和物力.如何利用大量无标签样本来改善学习性能成为当前机器学习研究的一个热点,半监督学习算法应运而生.半监督支持向量机最初应用于文本分类[1],随后研究人员提出一系列技术用于解决半监督支持向量机中非凸问题,如:梯度下降法[2]、凹凸法[3]、确定性熄火方法[4]、连续优化方法[5],半定规划[6]等.但是从以上这些文献的实验结果中可以发现一个问题:针对同一个数据集,上述算法得到的最优解存在一定的差异,造成该差异的主要原因是这些算法实现的是局部最优化.监督学习中,各算法的性能比较可以在同一个有标签的训练数据集或测试数据集上进行,但在半监督学习中,多数的样本为无标签样本,只在少数有标签样本集上进行的性能比较显然不够全面[7].

分枝定界(Branch-and-Bound)算法可以在小样本集上实现全局最优化,将它与半监督支持向量机相结合可以为其他半监督支持向量机学习性能的比较提供一个基准(benchmark).分枝定界最初由 Eennett 和 Demiriz 应用于解决半监督支持向量机中的整数规划算法中[8].最近 Chapelle 等人提出一种分枝定界半监督支持向量机(Branch and bound for semi-supervised support vector machines, BBS^{3}VM)[9].本文在此基础上提出一种改进的分枝定界半监督支持向量机学习算法(Improved learning algorithm for branch and bound for semi-supervised support vector machines, IBBS^{3}VM).算法重新对各结点下界的估计进行定义,降低计算各结点下界的时间复杂度,利用支持向量机的几何特点确定分枝结点,从而提高算法的速度,仿真实验表明 IBBS3VM 算法具有分类精度高,鲁

收稿日期:2009-03-20;修回日期:2009-07-01
基金项目:国家自然科学基金(No.60873037,No.60673131)

棒性强等特点.

2 半监督支持向量机

目前,半监督支持向量机根据优化参数的不同可分为两大类,基于组合的半监督支持向量机和基于连续优化算法的半监督支持向量机. BBS^{3}VM 是一种典型的基于组合的半监督学习方法,以下给出详细的介绍.

2.1 基于组合的半监督支持向量机

以两分类问题为例,训练样本为一组给定的独立同分布的有标签训练样本集: $L=\{(x_1,y_1),\cdots,(x_l,y_l)\}$, $x_i\in R^d$, $y_i\in\{+1,-1\}$和另一组来自同一分布的无标签样本集: $U=\{x_{l+1},x_{l+2},\cdots,x_n\}$, $x_i\in R^d$, $n=l+u$. 半监督支持向量机可以描述为以下的优化问题:

$$\min_{(w,b),y_u} I(w,b,y_U)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}V(y_i,o_i)+C^*\sum_{i=l+1}^{n}V(y_i,o_i) \tag{1}$$

其中, $\boldsymbol{y}_U=(y_{l+1},y_{l+2},\cdots,y_n)$为无标签样本集的标签向量. 参数 C 为有标签样本的惩罚系数,参数 C^* 为无标签样本在训练过程中的影响因子. $o_i=\boldsymbol{w}_i^{\mathrm{T}}x_i+b$, V 为损失函数,一般定义为 Hinge 函数形式:

$$V(y_i,o_i)=\max(0,1-y_io_i)^p$$

$p=1,2$. 为了使目标函数连续可导,本文取 $p=2$,此时的 Hinge 函数为二次软间隔损失函数.

基于组合的半监督支持向量机的思想:对无标签样本 $\boldsymbol{y}_u$ 的所有组合,通过监督 SVM 算法对各种组合的目标函数进行优化,从而得到半监督支持向量机的最优解. 根据此思想,定义 $F(\boldsymbol{w},b,\boldsymbol{y}_u)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}V(y_i,o_i)+C^*\sum_{i=l+1}^{n}V(y_i,o_i)$为任一给定 $\boldsymbol{y}_u$ 下监督学习的目标函数,此时,半监督支持向量机的目标函数定义为:

$$I(\boldsymbol{w},b,\boldsymbol{y}_U)=\min_{w,b}F(\boldsymbol{w},b,\boldsymbol{y}_u) \tag{2}$$

$F(\boldsymbol{w},b,\boldsymbol{y}_u)$的对偶问题可以表示为:

$$\begin{aligned}\max_{\alpha}W(\boldsymbol{\alpha})=&\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_jy_iy_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&-\frac{1}{2}\left(\sum_{i,j=1}^{l}\alpha_i\alpha_jy_iy_j\frac{\delta_{ij}}{2C}+\sum_{p,q=l+1}^{n}\alpha_p\alpha_qy_py_q\frac{\delta_{pq}}{2C^*}\right)\end{aligned} \tag{3}$$

$$\text{s.t.}\sum_{i=1}^{n}y_i\alpha_i=0$$

$$\alpha_i\geqslant 0,\ i=1,\cdots,n$$

其中, $\delta_{ij}=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}$

此时,半监督支持向量机转化为求解一个最小-最大化的问题,最大值通过求解有监督支持向量机得到,最小值则需要取遍无标签样本集所有标签组合得到. 对于非线性情况,可以采用核技巧,利用核函数将输入空间映射到高维的特征空间将非线性问题转化为该空间中的线性分类问题. 显然,这一最小-最大问题的精确解需要搜索无标签样本集上所有 2^u 种分类结果. 即使在无标签样本集规模有限的情况下,这种全局搜索也是不现实的,设计具有约束条件及搜索策略的启发算法成为提高搜索速度的一个可行方法.

2.2 均衡约束(Balancing constraint)

式(1)中,存在无标签样本数据倾斜划分问题,即有可能将所有的无标签样本分到一个类,此时,虽然可以最大化两类间隔,但泛化性较低. 为了避免倾斜划分问题,研究人员引入一个可以使得无标签样本被合理分到两个类的均衡约束. 文献[6]直接强制无标签样本集中两类样本的比例与有标签样本集中的两类样本的比例相同即:

$$\frac{1}{u}\sum_{i=l+1}^{n}\max(y_i,0)=r;$$

本文采用文献[4]所提出的比较宽松的约束:

$$\frac{1}{u}\sum_{i=l+1}^{n}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b=\frac{1}{l}\sum_{i=1}^{l}y_i \tag{4}$$

3 分枝定界算法

分枝定界算法由 Land Doig 等人提出,算法采用类似分而治之的策略,在分析组合最优化问题的一切可行解的过程中,采取了必要的限制条件,设法排除可行域中大量非最优解区域,从而可以实现全局最优化. 算法的描述如下:

设一个最优化问题的一个实例为(F,f),其中 F 为一有限集合或可行域, f 为代价函数或映射,所求解问题为可以表示为:

$$\min f(x)\qquad \text{s.t}\ x\in F \tag{5}$$

将原问题分解为一个或多个子问题,其中与子问题对应的可行域为 $F_1,F_2,\cdots,F_k$, $F_i\in F(i=1,\cdots,k)$. 在第 i 个子问题的可行域 $x\in F_i$ 内,求解 $\min f(x)$. 接着按同样的方式将可行域为 F_i 的子问题再分解为若干个子问题,直到某个子问题容易求解为止,如式(6)所示:

$$\min_{x\in F}f(x)=\min_{1\leqslant i\leqslant k}\{\min_{x\in F_i}f(x)\} \tag{6}$$

4 改进的分枝定界半监督支持向量机

4.1 分枝定界半监督支持向量机

文献[9]将分枝定界算法引入到半监督支持向量机中,提出一种基于分枝定界的半监督支持向量机学习算法(BBS^{3}VM). 算法利用分枝定界算法对 y_u 不同组合下的目标函数进行搜索,在优化局部最优解的过程

中实现全局搜索. BBS^{3}VM 算法有两个关键问题需要解决:(1)结点下界的定义;(2)分枝时无标签样本的选择. BBS^{3}VM 算法存在两个缺陷:①对结点下界进行估计时需要对无标签样本的组合进行大量的 0 ~ 1 二次规划;②在进行分枝时需要多次进行支持向量机训练,虽然文中采用了"add-one-in"方法用于实现对目标函数值的估计,可以在一定程度上降低计算复杂度,但在计算过程中涉及到大量的矩阵求逆的运算,另外当新加入样本为违反 KKT 条件的样本时,需要重新进行支持向量机训练.本文针对以上两个问题提出 IBBS^{3}VM.

4.2　分枝定界树的构造

分枝定界树的形式为二叉树,其上的每个结点表示为 $node_i\{L,U,lb\}$,其中 $L=\{(x_1,y_1),\cdots,(x_l,y_l)\}$ 为有标签样本集,$U=\{(x_{l+1},y_{l+1}),\cdots,(x_n,y_n)\}$ 为目前还未标记的无标签样本集,此时令 $y_i=0, i=l+1,\cdots,n$, lb 为该结点的下界. ub 为该分枝定界树的上界.两个孩子为对样本 $x_i, i\in U$ 标注不同标签后的结点.当 $U=\varnothing$ 时,该结点为叶子结点,其上所有的样本均有标签.

4.3　定界规则

基于分枝定界树的半监督支持向量机算法中上界和下界用于刻画分枝及剪枝的条件,对于提高全局搜索的速度起着至关重要的作用.

4.3.1　上界定义

IBBS^{3}VM 算法中上界为当前半监督支持向量机函数的最小值,即当前最优的可行解,用于定义子树被剪枝的条件.在分枝定界树全局搜索的过程中,叶子结点 $node_i\{L,U,lb\}$ 上所有无标签样本都有类别标注.

定义 1　设 $F_i(\boldsymbol{w},b,\boldsymbol{y}_u)$ 为 $\boldsymbol{y}_u$ 组合下叶子结点 $node_i\{L,U,lb\}$ 上的目标函数值,如果存在 $F^*(\boldsymbol{w},b,\boldsymbol{y}_u^*)$ 满足 $F^*(\boldsymbol{w},b,\boldsymbol{y}_u^*)=\min\{F_1(\boldsymbol{w},b,\boldsymbol{y}_u),F_2(\boldsymbol{w},b,\boldsymbol{y}_u),\cdots,F_k(\boldsymbol{w},b,\boldsymbol{y}_u)\}$,则称 $F^*(\boldsymbol{w},b,\boldsymbol{y}_u^*)$ 为该分枝定界树的上界,记为 ub.

k 为叶结点数,初始化时可将上界定义为无穷大.上界随着全局搜索的深入进行,不断的被更新,直至完成全部的搜索,得到全局最优解.

4.3.2　下界定义

对于分枝定界算法而言,很难直接求原始优化问题的下界,但可以通过不断修改分枝定界树上各结点下界估计的方法来实现.

定义 2　设 $F_i(\boldsymbol{w},b,\boldsymbol{y}_u)$ 为 y_u 组合下结点 $node_i\{L,U,lb\}$ 上的目标函数,如果该目标函数存在一个下界 $\inf F_i(\boldsymbol{w},b,\boldsymbol{y}_u)$,则称其为该结点的下界,记为 lb,即:

$$lb=\inf F_i(\boldsymbol{w},b,\boldsymbol{y}_u) \tag{7}$$

根据分枝定界树的原理,原始优化问题的求解可以转化为各个可行域内(即各分枝)的子优化问题 $F_i(\boldsymbol{w},b,\boldsymbol{y}_u)$,现根据其对偶形式构造函数:

$$D(\boldsymbol{\alpha},\boldsymbol{y}_u)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\frac{1}{2}\left(\sum_{i,j=1}^{l}\alpha_i\alpha_j\frac{\delta_{ij}}{2C}+\sum_{p,q=l+1}^{n}\alpha_p\alpha_q\frac{\delta_{pq}}{2C^*}\right) \tag{8}$$

为方便描述称其为伪对偶函数.

定理 1　分枝定界树各结点上伪对偶函数的函数值 $D(\boldsymbol{\alpha},\alpha(\boldsymbol{y}_u^*))$ 可以作为分枝定界树原始优化问题 $I(\boldsymbol{w},b,\boldsymbol{y}_u)$ 下界的一个估计,即 $D(\boldsymbol{\alpha},\alpha(\boldsymbol{y}_u^*))\leqslant I(\boldsymbol{w},b,\boldsymbol{y}_u)$.

证明　对于任意给定符合约束条件的 $\boldsymbol{y}_u$ 有

$$D(\boldsymbol{\alpha},\boldsymbol{y}_u)\leqslant\max_{\alpha}W(\alpha) \tag{9}$$

由对偶定理知,

$$\max_{\alpha}W(\boldsymbol{\alpha})=\min_{w,b}F(\boldsymbol{w},b) \tag{10}$$

根据分枝定界原理有

$$\min_{w,b}F(\boldsymbol{w},b,\boldsymbol{y}_u)=I(\boldsymbol{w},b,\boldsymbol{y}_U) \tag{11}$$

将式(9) ~ (11)联立有 $D(\boldsymbol{\alpha},\alpha(\boldsymbol{y}_u))\leqslant I(\boldsymbol{w},b,\boldsymbol{y}_u)$.即 $D(\boldsymbol{\alpha},\alpha(\boldsymbol{y}_u))$ 可以作为原始优化问题的下界.　　证毕.

显然,对于叶子结点所有的样本均有标签,不存在下界估计的问题,此时有 $lb=ub$.

结点 $node_i\{L,U,lb\}$ 下界的计算方法如下:

(1)给定一个非叶子结点 $node_i\{L,U,lb\}$,在 L 上进行支持向量机训练,求得 $\alpha_i(1\leqslant i\leqslant l)$ 及初始分类器 SVM_0;

(2)用初始分类器 SVM_0 对无标签样本进行学习,计算每一个无标签样本的判别函数输出,依据该输出对无标签样本进行标注,标注规则如下:

$$y_i=1,\ f(x_i)>0$$

$$y_i=-1,\ f(x_i)<0$$

(3)根据 $\boldsymbol{y}_u$ 构造一个向量 $\boldsymbol{\alpha}(\boldsymbol{y}_u)$ 规则如下:

$$\alpha_i=C^*,\ -1<f(x_i)<1$$

$$\alpha_i=0,\ f(x_i)<-1 \text{ 或 } f(x_i)>1$$

(4)计算 $D(\boldsymbol{\alpha},\alpha(\boldsymbol{y}_u))$ 的值作为该结点的下界.

在 IBBS^{3}VM 算法中采用一个指示向量 (e_1,e_2) 表示结点的当前状态,e_1,e_2 分别表示结点定界、访问状态.当一个非叶子结点实施定界操作后,该结点的定界状态指示位置为 1;当结点实施了分枝操作,或者由于应用了剪枝规则而无须再分解时,可将该结点的访问状态指示位置为 1,表示该结点已被访问.

4.4　分枝方法

根据分枝定界算法剪枝原理,经分枝操作产生的结点在满足剪枝准则时不需要进一步分解,定义 IBBS^{3}VM 剪枝准则如下:

剪枝准则 1:优化测试,当前结点的下界大于分枝定界树上界.

剪枝准则 2:约束测试,该结点上的样本集不符合

均衡约束.

剪枝准则 3: 可行性测试,对该子问题求解完毕,即达到叶子结点.

当一个已定界结点的下界不符合剪枝条件时,需要对该结点进行分枝操作.在进行分枝时有两个关键问题需要考虑:(1)应选取哪个无标签样本作为下一个标注的样本,也就是子结点的选取策略;(2)对这个无标签样本标注哪个标签,即子结点标注策略.本文采用类似中心距离比值法来确定无标签样本的标签,对当前可信度最高的样本进行标注,以减少搜索过程中修改分枝的次数.

定义 3(样本可信度) 在两分类问题中,无标签样本 x_i 到正类中心的距离与到负类中心的距离的比值为该无标签样本 x_i 关于正类的可信度 $\mu_+(x_i)$,即:

$$\mu_+(x_i)=\frac{D(x_i,a)}{D(x_i,b)} \tag{12}$$

其中 a,b 分别为正类和负类的中心,由于传统的欧氏距离不能完全反映数据的空间分布特性,本文采用文献[10]提出的密度可调节的线段长度用于表示无标签样本到类中心的距离,即 $D(x,y)=\rho^{\|x-y\|^2}-1,\rho>1$ 为伸缩因子.该距离定义可以依据密度扩大或缩小两点之间的欧氏距离.

对所有无标签样本 $x_i\in U,i=1,\cdots,n$ 进行关于正类的可信度计算,并对计算结果进行排序,根据排序结果选择无标签样本进行标注,选择及标注规则如下:

当 $\max\mu_+(x_i)>\dfrac{1}{\min\mu_+(x_j)}$ 则对 x_i 进行标注 $y_i=+1$;

否则,对 x_j 进行标注 $y_j=-1$.

IBBS^{3}VM 算法的问题描述如下:

算法:IBBS^{3}VM

输入:有标签样本集 $L=\{(x_1,y_1),\cdots,(x_l,y_l)\}$;

无标签样本集 $U=\{(x_{l+1},y_{l+1}),\cdots,(x_n,y_n)\}$;

输出:半监督支持向量机分类器 φ;

无标签样本集的标签向量 $\boldsymbol{Y}^*$;

方法:

Step1 在 $node_0\{L,U,lb\}$ 上的有标签样本集上训练,求得上界 ub_0 和下界 lb_0.

Step2 计算无标签样本的可信度,选择可信度最高的样本 x_i 进行分枝并标注其类别 y_i,生成结点 $node_j\{L\cup\{x_i,-y_i\},U/x_i,lb\}$、$node_{j+1}\{L\cup\{x_i,y_i\},U/x_i,lb\}$,分别对其执行 $push(node_j)$,$push(node_{j+1})$操作;

Step3 $pop(node_{j+1})$,计算 $node_{j+1}$该结点的下界 lb_{j+1},若 $lb_{j+1}<ub_0$,则返回 Step2,否则转 Step4;

Step4 依据剪枝准则进行判断,对满足准则 1 和 2 的结点进行剪枝;

Step5 当 $U=\varnothing$,到达叶子结点计算该结点的上界 ub_{j+1},$ub_{j+1}=ub_j$;

Step6 重复 Step2 至 Step5 直到栈为空,并输出 φ 和 $\boldsymbol{Y}^*$.

5 实验

为了验证本文提出的 IBBS^{3}VM 算法的有效性,分别在 3 组人工样本集 *g22c*、*g33c*、*2moons* 及 2 组真实样本集 Text、COIL3 上进行仿真实验.根据实验结果从分类精度、训练时间及参数敏感性三方面进行分析.

5.1 数据集

表 1 给出了实验数据集的数据特征. *g22c* 和 *g33c* 数据集的样本分别由 2 个、3 个正态分布函数生成,其中 *g22c* 有标签样本数为 2(每类各 1 个),如图 1 所示;而 *g33c* 数据集从每一类中选取 10% 作为有标签样本,另外 90% 作为无标签样本. *2moons* 数据集是半监督学习算法的一个基准数据集,该数据集的两个有标签样本是固定的,其他无标签样本在实验中随机生成,如图 3 所示. *Text* 样本集为 *Summer* 等人提供的 *newsgroups*20 样本集中的 *mac* 和 *windows* 子集,本文只选取其整理后的训练样本集,两个子集共有 780 个样本,每个样本有 7511 维,其中有标签的为 50 个. *COIL*3 为 3 个目标从不同角度得到的灰度图像数据,从每类中随机选择 2 个目标作为有标签样本.

表 1 实验数据集描述

Data set	Classes	Dims	Points	labeled
g22c	2	2	500	2
g33c	3	3	300	30
2moons	2	3	500	2
Text	2	7511	780	50
COIL3	3	1024	216	6

5.2 分类精度分析

实验选择高斯核函数 $k(x,y)=\exp(-\|x-y\|^2/2\sigma^2)$.为了方便比较,同一数据集上的所有算法都采用相同的参数,采用 5 重(5-fold)交叉验证测试分类性能.另外,由于本文只是在半监督学习算法内比较性能,均衡约束条件中的参数 r 根据无标签样本集实际的标签确定.表 2 给出了各算法精度比较结果.

表 2 各算法分类精度比较

	∇ S^{3}VM[5]	CCCP[3]	S^{3}VM^{light}[1]	BBS^{3}VM	IBBS^{3}VM
g22c	98.9	99.2	99.4	100	100
2moons	44.7	37.6	33.8	100	100
g33c	55.3	47.9	65.2	100	100
Text	94.8	91.5	91.4	100	100
COIL3	38.8	53.3	43.7	97.6	96.9

IBBS^{3}VM 在 *g*22*c* 和 2*moons* 数据集的结果如图 2、图 4 所示.从表 2 中可以看到本文提出的 IBBS^{3}VM 算法与 BBS^{3}VM 都达到了 100%的精度,这也表明这两种算法都可以作为不同半监督支持向量机学习算法性能比较的基准.COIL3 样本集对于半监督支持向量机学习算法来说是一个挑战,数据集本身并不完全符合聚类假设,从表 2 中也可以看到采用 ∇S^{3}VM、CCCP、S^{3}VM^{light}等算法的分类效果不理想.由于本文提出的 IBBS^{3}VM 算法及 BBS^{3}VM 实现的是全局搜索,其分类效果明显优于前三个算法.所有的算法在数据集 *g*22*c* 上的分类效果都很好,这主要是由于该人工数据集完全符合聚类假设.数据集 *g*33*c*、*COIL*3 为两个多分类问题,在其上的分类效果都不理想,这主要是由于本文对于多分类问题采用 1-*n* 策略,而该方法会导致样本数严重不均衡,另外,过多的类别也会影响聚类假设.但是,采用 1-1 的策略,对于样本数极少的半监督学习而言又十分困难.

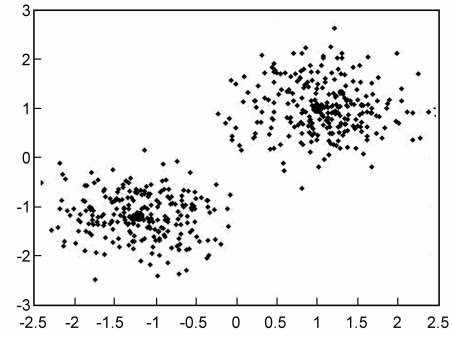

图1　*g*22*c*数据集分布图

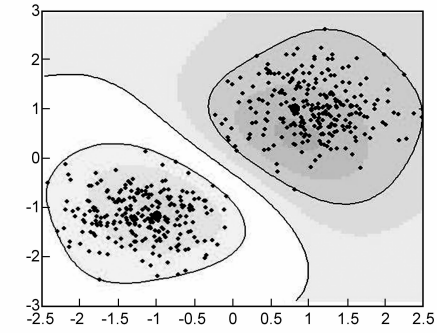

图2　*g*22*c*数据集半监督学习结果图

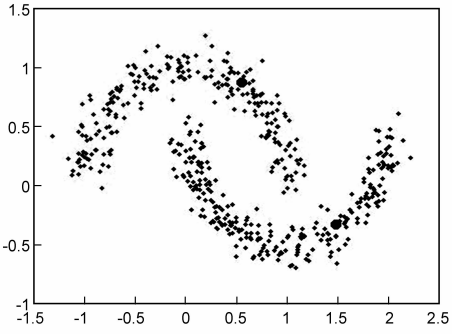

图3　2*moons*数据集分布图

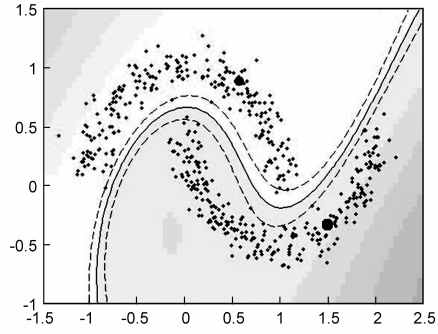

图4　2*moons*数据集半监督学习结果图

5.3　训练时间比较分析

由于∇S^{3}VM、CCCP、S^{3}VM^{light}在训练过程采用了熵渐近的策略,训练速度不仅与算法的设计有关更与迭代次数相关,这种情况下对各算法的训练时间进行比较意义不大,本文只对 BBS^{3}VM 和 IBBS^{3}VM 两种算法的训练时间进行比较,比较结果如表 3 所示.

表 3　两种算法训练时间比较

	BBS^{3}VM(s)	IBBS^{3}VM(s)
*g*22*c*	220	183
2*moons*	180	157
*g*33*c*	280	225
Text	898	569
*COIL*3	364	224

从表 3 中可以看出同 BBS^{3}VM 算法相比 IBBS^{3}VM 算法有更快的执行速度,而且随着样本数的增加这种优势更加明显.另外在实验的过程中,本文所提出的 IBBS^{3}VM 算法在 COIL3 数据集上的全局最优解明显小于其他的局部最优解,这也是其运算速度提高的原因之一.

5.4　参数敏感性分析

不同的算法在达到最优分类性能时的参数是不同的.有一些算法对于参数极其敏感,为了分析各算法对参数的敏感程度.表 2 中各算法所采用的参数如表 4 所示.现将表 2 中各算法的结果作为基准,比较各种算法所需参数$\{\sigma_c, C_c, C_c^*\}$在 $\sigma_c \in \{\frac{1}{2}\sigma, \sigma, 2\sigma\}$, $C_c \in \{\frac{1}{10}C, C, 10C\}$, $C_c^* \in \{\frac{1}{10}C^*, C^*, 10C^*\}$不同组合下的分类性能,每个参数组合独立进行 10 次实验取平均值,结果如图 5 所示. ∇S^{3}VM、CCCP、S^{3}VM^{light}算法的 C^* 均采用熵渐近方法,本文所给出的为 C^* 的初始值.

表 4　各算法参数取值表

	σ	C	C^*
g22c	1	10	100
g33c	0.5	10	100
2moons	0.5	10	100
Text	4	40	400
COIL3	3000	100	1000

从图 5 可以看出算法 ∇S³VM、CCCP、S³VM^light在不同参数组合下的实验结果差异幅度较大，特别是 ∇S³VM 平均差异度达到 27.6%．而 BBS³VM 和 IBBS³VM 两种算法实验结果差异幅度不大，这就说明这两种算法对于参数的要求不高，算法的性能对参数选择不敏感，但不同的参数对于训练时间有一定的影响．

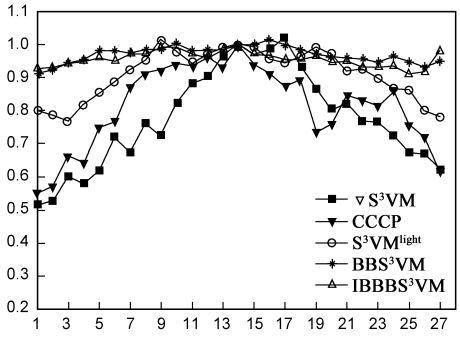

图5 各算法参数对精度影响

6 结论

实验表明本文提出的 IBBS³VM 具有分类精度高、参数不敏感等优点．需要指出的是虽然 IBBS³VM 同 BBS³VM 相比速度有一定的提高，但是由于其实现的是一个全局搜索，对于大规模样本集来说这两种方法仍然受到运算时间的限制，可以尝试从支持向量机算法的选取及并行运算两个角度解决问题，这也可以作为今后的一个研究方向．算法为其他半监督支持向量机甚至其他半监督学习[11]的比较提供了一个基准，也不失为一个新的研究方向．在半监督支持向量机研究领域里仍有许多亟待解决的问题．例如，半监督支持向量机学习算法中对无标签样本集各类的样本分布比例的估计、无标签样本预先选取、半监督支持向量机多分类问题的学习等都是一些具有挑战值得深入研究的问题．

参考文献：

[1] T Joachims. Transductive inference for text classification using support vector machines[A]. Proceedings of the 16th International Conference on Machine Learning[C]. Slovenia, 1999. 200 – 209.

[2] O Chapelle, A Zent. Semi-supervised classification by low density separation [A]. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics[C]. Barbados, 2005. 57 – 64.

[3] R Collobert, F Sinz, J Weston, L Bottou. Large scale transductive SVMs[J]. Journal of Machine Learning Research, 2006, 7: 1687 – 1712.

[4] V Sindhwani, S Keerthi, O Chapelle. Deterministic annealing for semi-supervised kernel machines[A]. Proceedings of the 23rd international Conference on Machine Learning[C]. Pittsburgh: ACM, 2006. 841 – 848.

[5] O Chapelle, M Chi, A Zien. A continuation method for semi-supervised SVMs[A]. Proceedings of The 23rd International Conference on Machine Learning[C]. Pittsburgh, 2006. 185 – 192.

[6] T De Bie, N Cristianini. Semi-supervised learning using semi-definite programming[A]. O Chapelle, B Schoëlkopf, A Zen. Semi-supervised Learning[M]. USA: MIT Press, 2006. 8134 – 152.

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

[8] K Bennett, A Demiriz. Semi-supervised support vector machines[A]. Advances in Neural Information Processing Systems[C]. Colorado, 1998. 368 – 374.

[9] O Chapelle, V Sindhwani, S Keerthi. Branch and bound for semi-supervised support vector machine[A]. Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems[C]. British Columbia, 2006. 217 – 224.

[10] 王玲，薄列峰，焦李成．密度敏感的半监督谱聚类[J]．软件学报，2007，18(10)：2412 – 2422.
Wang Ling, Bo Lie-Feng, Jiao Li-Cheng. Density-sensitive semi-supervised spectral clustering[J]. Journal of Software, 2007, 18(10): 2412 – 2422. (in Chinese)

[11] 戴新宇，田宝明，周俊生，陈家骏．一种基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT[J]．电子学报，2009，36(8)：1627 – 1630.
Dai Xin-yu, Tian Bao-ming, Zhou Jia-jun. LSASGT: An approach to text categorization based on latent semantic analysis and spectral graph transducer[J]. 2008, 36(8): 1626 – 1630. (in Chinese)

作者简介：

赵 莹 女，1981 年 6 月出生于黑龙江穆棱．现为哈尔滨工程大学博士研究生．主要研究方向为机器学习、数据挖掘等．
E-mail: yingzhao@hrbeu.edu.cn

张健沛 男，1956 年 11 月出生于黑龙江．哈尔滨工程大学教授、博士生导师．主要研究方向为数据库与知识库．
E-mail: zhangjianpei@hrbeu.edu.cn