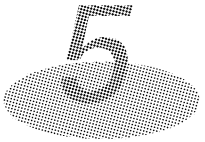

半监督支持向量机学习方法的研究*

李宇峰　周志华
南京大学 计算机软件新技术国家重点实验室，南京 210023

1　引言

监督学习是机器学习中研究最多、应用最广泛的一种学习框架[1]。监督学习通过对有标记示例进行学习，力图对未见示例进行准确的预测，即“强泛化性能”。众所周知，要获得泛化性能强的学习结果，需要使用大量的有标记示例。然而，在很多现实任务中，虽然可以收集得到大量未标记示例，但是有标记示例的数目往往非常有限，因为标记的获取通常需要耗费人力物力。例如在计算机辅助医疗诊断中，常常可以获得大量的未标记医学图像，但是我们无法要求医学专家给出所有图像的病灶(标记)，因为那样需要付出巨大的人力资源。在图像标注系统中，可以容易收集到大量图像，但是我们常常无法收集到大量的图像标注(标记)。在有标记示例稀少，得不到强泛化性能时，能否利用大量的未标记示例来提升学习性能？半监督学习[2-4]是这方面研究的一个主流方向。

近十年来，半监督学习取得了很多进展，涌现出大量学习方法[3-5]。根据工作方式的不同，它们可大致分为四类主流范型(paradigm)[5]。第一类为基于生成式模型的方法(generative model based method)[6-8]。它是生成式模型在半监督学习上的延伸。除了原来生成式模型的参数，它将未标记示例属于各个类别的概率也视为一组缺失参数，然后采用EM算法来进行标记估计和模型参数估计。第二类为基于图的方法(graph based method)[9-13]。它根据某种相似度度量把训练示例连成图，图的节点对应于训练示例，图的边对应于示例之间的相似度。基于光滑性假设(smoothness assumption)——相似的

* 本文得到国家自然科学基金重点项目(61333014)、国家自然科学基金青年科学基金项目(61403186)和江苏省基础研究计划(自然科学基金)——青年基金项目(BK20140613)资助。

示例具有相似的标记[4]，试图找到一种未标记示例的标记指派(label assignment)使得图上标记指派的总体误差最小。第三类为基于分歧的方法(disagreement based method)[5,14-16]。它使用两个或多个学习器(可以是监督学习器也可以是半监督学习器)，通过利用学习器之间的分歧这一重要信息[16]来改善学习性能。第四类为半监督支持向量机(semi-supervised SVM，简记为 S3VM)[2,17-18]。基于低密度假设(low density assumption)——决策边界经过数据低密度区域[4]，试图找到一个决策边界使得所有训练示例(包括有标记和未标记示例)到决策边界都具有大间隔(large margin)。值得一提的是，半监督学习的三大主流范型，即基于分歧的方法[14]、S3VM[18]、基于图的方法[11]的代表性工作先后于 2008、2009 和 2013 年获得了国际机器学习领域的十年最佳论文奖，这在一定程度上反映了半监督学习受到国际机器学习领域的高度重视。

在四个范型中，S3VM 在理论上根据 SVM 间隔理论[2]可以得到比较清楚的理论基础、在模型上由于它源于 SVM 具有比较完善的数学基础和很好的模型扩展能力，S3VM 提出后得到机器学习领域广泛关注，并已在众多领域得以应用。然而，S3VM 所涉及的一些重要问题，例如数据规模、学习效率、性能保障、代价抑制等，仍亟待研究。我们对以上问题进行研究，取得了一些进展[19-23]。具体来说：针对传统 S3VM 方法难以处理大规模数据的问题，提出了 WELLSVM 方法，理论上证明了其优化解的全局性与紧致性，实验验证了该方法能处理的数据规模达到传统方法的 10 倍以上；针对传统 S3VM 方法学习速度慢的问题，提出了 MeanS3VM 方法，理论上证明了该方法具有强逼近能力，实验验证了其学习效率比传统方法快 10 倍以上；针对传统 S3VM 方法利用未标记示例后常会出现性能下降的问题，提出了 S4VM 方法，理论上给出了性能保障条件，实验验证了该方法能将性能下降的比例从传统方法的 15%减少到 1%；针对传统 S3VM 方法难以处理非均衡代价的问题，提出了 CS4VM 方法，理论上证明了其具有处理非均衡代价的能力，实验验证了该方法通常比传统方法减少 20%以上的总体代价。

本文余下部分如下组织：第 2 节介绍 S3VM 的概况及其涉及的一些重要问题；第 3 节介绍我们在这些问题上得到的主要结果；最后第 4 节总结全文。

2 半监督支持向量机简介

S3VM 最早由美国国家工程院院士，统计学习理论的奠基人之一，同时也是 SVM 的创始人之一 V. Vapnik 在他的专著 *Statistical Learning Theory* 中提出构想[2]。V. Vapnik 在书中构想，当有标记示例不够多，得不到满意的性能时，SVM 可以利用未标记示例来改善性能。他在书中证明：当未标记示例到 SVM 超平面具有大间隔时，SVM 决策函数等价

类个数可以显著减少，从而有效地减少函数空间的 VC 维。换言之，如果有标记示例上的经验风险不变，未标记示例具有大间隔可以缩减决策函数的泛化风险上界，从而得到改善的泛化性能。当时 V. Vapnik 只给出了这个理论结果，并没有给出 S3VM 具体实现。后来，美国伦斯勒理工大学的 K. P. Bennett 和 A. Demiriz[17]，美国康奈尔大学的 T. Joachims[18] 分别从 SVM 扩展和文本分类的角度给出了 S3VM 的实现。其中 T. Joachims 实现效率更高，并发布了 S3VM 首个开源算法包 SVMlight。这个工作后来得到机器学习领域广泛关注，也因此于 2009 年获得了国际机器学习领域为数不多的十年最佳论文奖。

简单来说，S3VM 可视为 SVM 在半监督学习上的延伸。图 1 给出 S3VM 的一个示意图。给定少量有标记示例和大量未标记示例，S3VM 试图找到一个决策边界使得所有示例到该决策边界都具有大间隔。从形式化来讲，给定少量有标记示例 $\mathcal{L}=\{\boldsymbol{x}_i,y_i\}_{i=1}^{l}$ 和大量未标记示例 $\mathcal{U}=\{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$，$l\ll u$，$N=l+u$，$\boldsymbol{x}\in\mathcal{R}^d$ 为 d 维向量，$y\in\{+1,-1\}$ 表示示例的输出标记（为便于说明，本文只讨论二分类任务）。S3VM 希望找到一个决策函数 $f:\mathcal{R}^d\to\{+1,-1\}$ 使得如下目标最小化，

$$\begin{aligned}
\min_{\hat{y}_{l+1},\cdots,\hat{y}_N}\quad & \min_{\boldsymbol{w},\boldsymbol{\xi}}\quad \frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i=1}^{l}\xi_i+C_2\sum_{j=l+1}^{N}\xi_j\\
\text{s. t.}\quad & y_i\boldsymbol{w}'\phi(\boldsymbol{x}_i)\geqslant 1-\xi_i,\quad i=1,\cdots,l.\\
& \hat{y}_j\boldsymbol{w}'\phi(\boldsymbol{x}_j)\geqslant 1-\xi_j,\quad j=l+1,\cdots,N.\\
& \hat{y}_j\in\{+1,-1\},\quad \xi_j\geqslant 0,\xi_i\geqslant 0,i=1,\cdots,l,j=l+1,\cdots,N.\\
& -\beta\leqslant\frac{\sum_{j=l+1}^{N}\hat{y}_j}{N-l}-\frac{\sum_{i=1}^{l}y_i}{l}\leqslant\beta
\end{aligned}$$

$\{\hat{y}_{l+1},\cdots,\hat{y}_N\}$ 表示未标记示例的标记指派。$f(\boldsymbol{x})=\boldsymbol{w}'\phi(\boldsymbol{x})$ 为 S3VM 的决策函数（或超平面）其中 $\phi(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的高维特征映射（它由某个核函数 k 导出，满足 $k(\boldsymbol{x},\hat{\boldsymbol{x}})=\phi(\boldsymbol{x})'\phi(\hat{\boldsymbol{x}})$）。$\boldsymbol{\xi}=[\xi_1,\cdots,\xi_N]$ 为训练示例上的预测损失。C_1 和 C_2 是参数，用于权衡决策函数复杂性、

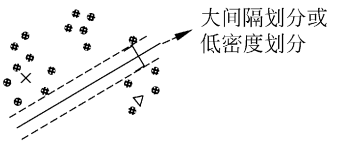

图 1　S3VM 示意图

注：叉和三角形表示有标记示例，灰点表示未标记示例。给定少量有标记示例和大量未标记示例，S3VM 试图找到一个大间隔划分使得所有示例到决策边界都具有大间隔。S3VM 的大间隔划分也常称为低密度划分[4,23]。

有标记示例及未标记示例上的损失。式中第一、二条约束分别要求有标记示例与未标记示例到决策边界都具有大间隔(如果间隔不足够大,所导致的损失会通过 $\boldsymbol{\xi}=[\xi_1,\cdots,\xi_N]$ 反映到目标函数)。第三条约束为优化变量的取值范围。最后一条为平衡约束(balance constraint),用于防止 S3VM 将所有未标记示例都指派为同一类别,β 是参数用于控制有标记示例与未标记示例上类别比例之差[18]。S3VM 的大间隔划分也常称为低密度划分(因为如果 S3VM 穿过高密度数据区域,S3VM 将导致很大的损失)[4,23]。

S3VM 提出后在众多领域取得应用。比如,T. Joachims 将 S3VM 用于文本分类,取得了比当时最好的技术更好的性能[18]。A. Kockelkorn 等人将 S3VM 用于邮件分类,发现 S3VM 在有标记示例稀少时,得到显著的性能提升[24]。L. Wang 等人将 S3VM 用于图像检索,发现 S3VM 结合主动学习技术显著地提升了性能[25]。N. Kasabov 和 S. Pang 将 S3VM 用于生物信息学,发现 S3VM 在基因启动序列(promoter sequence)识别上效果得到了明显提升[26]。C. Goutte 等人将 S3VM 用于医疗本文的实体识别,同样取得了显著的效果[27]。

然而,S3VM 所涉及的一些重要问题,比如数据规模、学习效率、性能保障、代价抑制等,仍急需要研究。

(1) 数据规模问题:与传统 SVM 不同,S3VM 涉及整数标记指派的优化,整体优化不再是凸问题(convex problem),而是个混合整数规划(mixed-integer program)问题。这通常得不到最优解[4,17-18]。现有技术要么难有求解效果保证,要么难以处理大规模训练示例。

(2) 学习效率问题:以往 S3VM 方法通过直接优化 S3VM 目标函数得到算法。然而 S3VM 目标函数求解的复杂度通常与未标记示例数量呈超线性增长,这导致现有 S3VM 方法的学习效率随着未标记示例数量的增加会显著下降。

(3) 性能保障问题:以往研究普遍认为 S3VM 利用了额外的未标记示例会改善学习性能。然而在不少情况下,S3VM 利用未标记示例不仅不能改善性能,有时候还会使性能下降。这个现象会严重影响 S3VM 的实际应用,特别是那些对未标记示例的利用要求高可靠性的应用场景。

(4) 代价抑制问题:以往 S3VM 假定训练示例的类别被错分所造成的代价是均衡的。这个假定在现实应用中通常并不成立,甚至错分类别所带来的代价会有很大的差别。S3VM 不能处理非均衡代价,将影响 S3VM 在更多应用中发挥作用。

3 半监督支持向量机学习方法

本节介绍我们在 S3VM 数据规模、学习效率、性能保障、代价抑制问题上得到的主要结果。

3.1 多：用于多训练示例的大规模半监督支持向量机

由于 S3VM 内在混合整数规划的高复杂性，学者们提出了很多优化技术对 S3VM 进行求解。第一类技术为全局优化技术，比如全局组合优化方法[28]、全局搜索方法[29]等。这类技术追求 S3VM 全局最优解，得到了不错的性能。然而这类技术的最坏情况时间复杂度与数据规模成指数增长，只能处理很少的训练示例。第二类技术为局部优化技术，比如梯度下降法[29]、局部标记交换法[18]、交替优化技术[30]、凸凹过程[31]等。这类技术牺牲了解的全局性，从而能够处理更大规模的数据。但这类技术容易陷入局部最小值，存在次优性能的问题。第三类技术为上述两种技术的折中。它将对 S3VM 松弛成一个凸问题，然后用凸问题的最优解来近似 S3VM 的最优解。直观上讲，当凸松弛紧致时，这类技术的解与 S3VM 的解会非常相近。半正定凸松弛[32-34]是这类技术的典型代表，得到了有前途的(promising)实验结果。然而这类技术的时间复杂度约为 $O(N^{4.5})$(其中 N 为训练示例规模)[30]，仍然难以处理大规模数据。

我们提出一种用于多训练示例的大规模半监督支持向量机学习方法 WELLSVM (weakly labeled SVM)[19]。WELLSVM 通过不断生成未标记示例的标记指派来最大化 S3VM 间隔。整个过程可证明是原 S3VM 目标函数的凸松弛形式，并可以证明 WELLSVM 凸松弛至少与现有半正定凸松弛一样紧。这个结果保证了 WELLSVM 求解效果的全局性和紧致性。另一方面，WELLSVM 的求解主要涉及一系列监督 SVM 求解，从而可利用当前可扩展性强的 SVM 算法如 LIBSVM[35]、SVM-perf[36]、LIBLINEAR[37]、CVM[38]等处理大规模训练示例。

具体来说，S3VM 可改写为下式

$$\begin{aligned}\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\min_{\boldsymbol{w},\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{j=l+1}^{N}\xi_i \\ \text{s.t.} \quad & \hat{y}_i\boldsymbol{w}'\phi(\boldsymbol{x}_i) \geqslant 1-\xi_i, \quad i=1,\cdots,N,\end{aligned} \tag{1}$$

其中

$$\mathcal{B}=\left\{\hat{\boldsymbol{y}}\,|\,\hat{\boldsymbol{y}}=[\hat{\boldsymbol{y}}_{\mathcal{L}};\hat{\boldsymbol{y}}_U],\hat{\boldsymbol{y}}_{\mathcal{L}}=\boldsymbol{y}_{\mathcal{L}},\hat{\boldsymbol{y}}_U\in\{\pm 1\}^{N-l};\ \frac{\mathbf{1}'\hat{\boldsymbol{y}}_U}{N-l}=\frac{\mathbf{1}'\boldsymbol{y}_{\mathcal{L}}}{l}\right\}$$

表示候选标记指派的集合。通过引入拉格朗日乘子 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]$，式(1)可等价为如下对偶最大化形式

$$\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\max_{\alpha\in\mathcal{A}} \quad G(\boldsymbol{\alpha},\hat{\boldsymbol{y}}):=\mathbf{1}'\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}'(\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}')\boldsymbol{\alpha}, \tag{2}$$

其中

$$\mathcal{A}=\{\boldsymbol{\alpha}|C_1\geqslant\alpha_i\geqslant 0,C_2\geqslant\alpha_j\geqslant 0,i\in\mathcal{L},j\in\mathcal{U}\}$$

是一个凸集。

WELLSVM 考虑优化如下目标

$$\max_{\alpha\in\mathcal{A}}\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\quad G(\boldsymbol{\alpha},\hat{\boldsymbol{y}}):=\mathbf{1}'\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}'(\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}')\boldsymbol{\alpha},\tag{3}$$

不难发现，WELLSVM 只作了个简单交换，将 S3VM 最小最大化问题换成最大最小化问题。根据 minimax 定理[39]可知，WELLSVM 的最优目标值为 S3VM 最优目标值的一个下界。式(3)有如下结果。

定理 1 WELLSVM 的目标函数即式(3)可以写成如下形式

$$\min_{\boldsymbol{\mu}\in\mathcal{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t G(\boldsymbol{\alpha},\hat{\boldsymbol{y}}_t),\tag{4}$$

其中 $\boldsymbol{\mu}$ 表示 μ_t 的向量形式，$\boldsymbol{M}=\{\boldsymbol{\mu}\mid\sum_t\mu_t=1,\mu_t\geqslant 0\}$ 为 $\boldsymbol{\mu}$ 的单纯形(simplex)。

由于 $G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})$ 是关于 $\boldsymbol{\alpha}$ 的凹函数，因此式(4)是个凸问题，换言之，WELLSVM 目标函数是 S3VM 的凸松弛形式。下面证明 WELLSVM 的凸松弛至少与现有半正定凸松弛一样紧。令 $\boldsymbol{M}_{\hat{\boldsymbol{y}}}=\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}'$，则 $G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})$ 可重写成 $\bar{G}(\boldsymbol{\alpha},\boldsymbol{M}):=\mathbf{1}'\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}'(\boldsymbol{K}\odot\boldsymbol{M}_{\hat{\boldsymbol{y}}})\boldsymbol{\alpha}$。定义

$$\mathcal{Y}_0=\{\boldsymbol{M}|\boldsymbol{M}=\boldsymbol{M}_{\hat{\boldsymbol{y}}},\hat{\boldsymbol{y}}\in\mathcal{B}\}.$$

则 S3VM 目标函数等价于

$$\min_{\boldsymbol{M}\in\mathcal{Y}_0}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\bar{G}(\boldsymbol{\alpha},\boldsymbol{M})$$

定义

$$\mathcal{Y}_1=\{\boldsymbol{M}|\boldsymbol{M}=\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\boldsymbol{M}_{\hat{\boldsymbol{y}}_t},\boldsymbol{\mu}\in\mathcal{M}\}.$$

则 WELLSVM 目标函数等价于

$$\begin{aligned}\min_{\mu\in\boldsymbol{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\bar{G}(\alpha,\mathcal{M}_{\hat{\boldsymbol{y}}_t})&=\min_{\mu\in\boldsymbol{M}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\bar{G}\Big(\boldsymbol{\alpha},\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\,\mathcal{M}_{\hat{\boldsymbol{y}}_t}\Big)\\&=\min_{\mathcal{M}\in\mathcal{Y}_1}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\bar{G}(\boldsymbol{\alpha},\mathcal{M})\end{aligned}$$

而现有半正定凸松弛方法[32-34]的目标函数等价于

$$\min_{\boldsymbol{M}\in\mathcal{Y}_2}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\bar{G}(\boldsymbol{\alpha},\boldsymbol{M})$$

其中 $\mathcal{Y}_2=\{\boldsymbol{M}|\boldsymbol{M}\geq 0,\boldsymbol{M}\in\mathcal{M}_{\mathcal{B}}\}$，$\mathcal{M}_{\mathcal{B}}$ 是一个跟 $\mathcal{B}$ 相关的凸集，即 $\mathcal{Y}_0\subseteq\mathcal{Y}_2$。根据以上改写，WELLSVM 跟现有半正定凸松弛的关系有如下结果。

定理 2 WELLSVM 至少与 S3VM 半正定凸松弛一样紧。

以上结果保证了 WELLSVM 求解效果的全局性和紧致性。下面介绍式(4)存在高

效算法。虽然式(4)涉及大量可能的标记指派,难以直接优化,然而对于式(4)的最优解而言,并非所有的指派都有助扩大间隔,有用的指派通常只是其中一个很小的子集。如果近似这个子集,则可以得到原问题一个不错的近似解。基于这个想法的一个直接实现是采用割平面算法[40]。

具体来说,WELLSVM 首先初始化一个标记指派 $\hat{\boldsymbol{y}}$ 并令工作集 $\boldsymbol{C}=\hat{\boldsymbol{y}}$,求解下式获得 $\boldsymbol{\alpha}$,

$$\min_{\mu\in\mathcal{M}}\max_{\alpha\in\mathcal{A}}\sum_{t:\hat{\boldsymbol{y}}_t\in\boldsymbol{C}}\mu_t G(\boldsymbol{\alpha},\hat{\boldsymbol{Y}}_t) \tag{5}$$

然后生成一个有利于扩大间隔的标记指派加入工作集 $\boldsymbol{C}$,重复"求解 $\boldsymbol{\alpha}$"与"生成标记指派"两个步骤直到 WELLSVM 的目标值不再下降。对于"求解 $\boldsymbol{\alpha}$",通过比对式(5)与多核学习目标函数[41],可以发现它等价于多核学习[41]问题,因此可以通过现有高效多核学习技术进行求解。而现有高效多核学习技术如文献[42]只涉及一系列监督 SVM 子问题,因此 WELLSVM 可利用当前高效 SVM 算法包如 LIBSVM[35]、SVM-perf[36]、LIBLINEAR[37]、CVM[38]等处理大规模训练示例。对于"生成标记指派",它可以通过一种简单快速的排序做法得到不错的近似解[19]。

下面给出 WELLSVM 的割平面算法可以在多项式迭代轮数内收敛。

假设 1 $G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})$满足如下条件:

1. $\mathcal{A}$是一个凸集;

2. 任意给定 $\hat{\boldsymbol{y}}$,$G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})$对于 $\boldsymbol{\alpha}$ 是一个凹函数;

3. $g_{\boldsymbol{y}}(\boldsymbol{\alpha})=-G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})|_{\hat{\boldsymbol{y}}=\boldsymbol{y}}$是一个 λ-强凸和 M-Lipschitz 函数。即,$\nabla^2 g_{\boldsymbol{y}}(\boldsymbol{\alpha})-\lambda\boldsymbol{I}\succeq\boldsymbol{0}$ 其中 $\boldsymbol{I}$ 为单位阵,$\|g_{\boldsymbol{y}}(\boldsymbol{\alpha})-g_{\boldsymbol{y}}(\bar{\boldsymbol{\alpha}})\|\leqslant M\|\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}}\|$,$\forall\,\boldsymbol{y}\in\mathcal{B},\boldsymbol{\alpha},\bar{\boldsymbol{\alpha}}\in\mathcal{A}$;

4. $\forall\,\hat{\boldsymbol{y}}\in\mathcal{B}$,$lb\leqslant\max_{\boldsymbol{\alpha}\in\mathcal{A}}G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})\leqslant ub$,其中 lb 和 ub 都是 N 的多项式函数;

5. $G(\boldsymbol{\alpha},\hat{\boldsymbol{y}})$可以改写成 $\bar{G}(\boldsymbol{\alpha},\boldsymbol{M})$,其中 $\boldsymbol{M}$ 是一个半正定阵,$\bar{G}$ 对于 $\boldsymbol{\alpha}$ 是个凹函数,对于 $\boldsymbol{M}$ 是一个线性函数。

定理 3 记 $p^{(t)}$ 为式(3.5)迭代到第 t 步的最优目标值,则有

$$p^{(t+1)}\leqslant p^{(t)}-\eta,$$

其中 $\eta=\left(\dfrac{-c+\sqrt{c^2+4\varepsilon}}{2}\right)^2$,$c=M\sqrt{2/\lambda}$。

推论 1 WELLSVM 的割平面算法将在$\dfrac{ub-lb}{\eta}$步收敛。

根据假设 1 可知,lb 和 ub 为 N 的多项式函数,因此 WELLSVM 迭代轮数最多是 N 的一个多项式函数。对于某些常用 SVM,例如 μ-SVM[43],lb 和 ub 都是与 N 无关的常数,WELLSVM 割平面算法迭代轮数最好可以达到常数级。

图 2 给出 WELLSVM 在 RCV1 上的实验结果。关于 WELLSVM 更多实验结果请参阅论文[19]。RCV1 数据包括 677 399 个示例和 47 236 个特征。WELLSVM 采用线性核，内部 SVM 实现采用 LIBLINEAR[37]。实验机器配置为：MATLAB 7.6、英特尔 Xeon (R)双核 3.20GHz、Windows 7 系统、8GB 内存。在这个数据集上，基于局部优化技术的传统 S3VM 方法如 TSVM[18]、LapSVM[44] 以及 UniverSVM[34] 和半正定凸松弛的方法[32-34]都无法在 24 小时内得到收敛。WELLSVM 与 SVMlin[45](专门用于加速线性核 S3VM 的方法)进行比较。有标记示例数量固定为 50 个，未标记示例数量分别采用 1%、2%、5%、15%、35%、55%和 75%的数据集大小，余下 25%数据作为测试。实验重复 10 次，报告平均性能。由图 2 可以看出，WELLSVM 取得了比 SVMlin 和监督 SVM 更好的性能。此外，WELLSVM 的可扩展性要优于 SVMlin 方法，并且随着未标记示例数量的增加，WELLSVM 的优势更加明显，WELLSVM 在相同时间内处理的数据规模可达 SVMlin 的 10 倍以上。

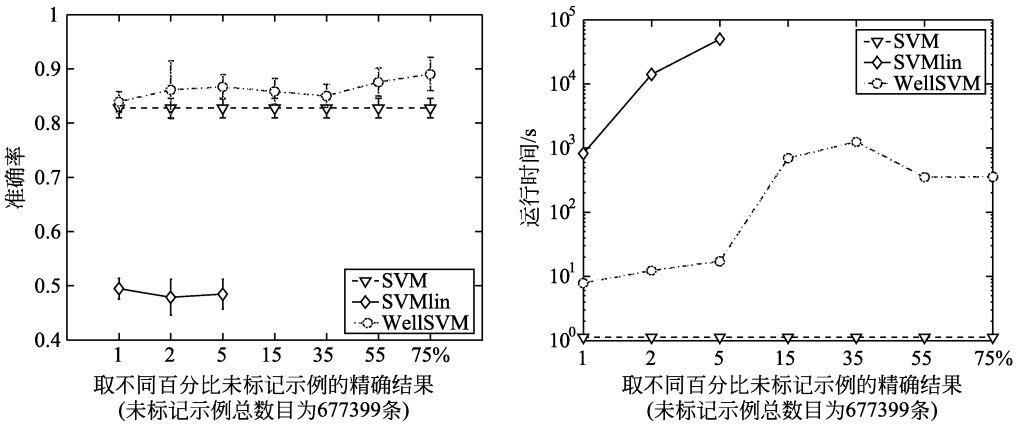

图 2 SVM、WELLSVM 和 SVMlin 在 RCV1 数据集上不同数量未标记示例下的比较结果

3.2 快：用于提升学习效率的快速半监督支持向量机

以往 S3VM 方法通过直接优化 S3VM 目标函数得到算法。然而 S3VM 目标函数求解的复杂度通常与未标记示例数量成超线性增长，这导致现有 S3VM 方法的学习效率随着未标记示例数量的增加会显著下降。比如 K. P. Bennett 和 A. Demiriz[17]、O. Chapelle 等人[29]的分支定界法，其时间复杂度随未标记示例数目呈指数级增长；TSVM[18]的局部标记交换策略虽然没有给出时间复杂度分析，但在实际应用中，TSVM 的迭代轮数通常

随未标记示例数目呈超线性增长[29]；Laplician SVM[44]涉及矩阵的求逆(矩阵规模与训练示例规模相当)，这导致立方级的时间复杂度。

我们提出一种用于提升学习效率的快速半监督支持向量机学习方法MeanS3VM[20]。有别于以往S3VM，MeanS3VM侧重于通过估计未标记示例的类中心来提升学习效率。可以证明，如果得到未标记示例的真实类中心，MeanS3VM将近似于得到未标记示例真实标记指派的监督SVM。具体来说，当训练示例可分时，MeanS3VM与监督SVM完全等价；当训练示例不可分时，MeanS3VM的损失函数将不超过监督SVM损失函数的两倍。因此，S3VM只需要估计未标记示例的类中心；而类中心估计可通过大间隔准则快速估计得到，因此MeanS3VM可以显著提升S3VM的学习效率。

具体来说，S3VM可形式化为下式：

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i\in\mathcal{I}_l}\xi_i+C_2\sum_{i\in\mathcal{I}_u}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i, \quad i\in\mathcal{I}_l, \\ & |\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b|\geqslant 1-\xi_i, \quad i\in\mathcal{I}_u;\ \xi_i\geqslant 0,\ i\in\mathcal{I}_l\cup\mathcal{I}_u, \\ & \sum_{i\in\mathcal{I}_u}\operatorname{sgn}(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)=r,\end{aligned} \tag{6}$$

其中最后一条约束对应于S3VM的平衡约束。

考虑如下形式

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{p}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i\in\mathcal{I}_l}\xi_i+C_2\sum_{i\in\mathcal{I}_u}(\xi_i+p_{i-1}-|f(\boldsymbol{x}_i)|) \\ \text{s.t.} \quad & y_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i, \quad i\in\mathcal{I}_l, \\ & \boldsymbol{w}'\phi(\boldsymbol{x}_i)+b\leqslant p_{i-l},\ -\boldsymbol{w}'\phi(\boldsymbol{x}_i)-b\leqslant p_{i-l};\ p_{i-l}\geqslant 1-\xi_i, \quad i\in\mathcal{I}_u; \\ & \xi_i\geqslant 0, \quad i\in\mathcal{I}_l\cup\mathcal{I}_u;\ \sum_{i\in\mathcal{I}_u}\operatorname{sgn}(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)=r\end{aligned} \tag{7}$$

则式(7)有如下结果。

定理4 式(7)与式(6)等价，即式(7)与S3VM等价。

由式(7)的平衡约束，未标记示例上正示例与负示例的数目为 $u_+=\frac{r+u}{2}$，$u_-=\frac{-r+u}{2}$。观察到，

$$\sum_{i\in\mathcal{I}_u}|f(\boldsymbol{x}_i)|-(u_--u_+)b=u_+\,\boldsymbol{w}'\hat{\boldsymbol{m}}_+-u_-\,\boldsymbol{w}'\hat{\boldsymbol{m}}_-,$$

其中

$$\hat{\boldsymbol{m}}_+=\frac{1}{u_+}\sum_{i\in\mathcal{I}_u,f(\boldsymbol{x}_i)\geqslant 0}\phi(\boldsymbol{x}_i),\quad \hat{\boldsymbol{m}}_-=\frac{1}{u_-}\sum_{i\in\mathcal{I}_u,f(\boldsymbol{x}_i)<0}\phi(\boldsymbol{x}_i)$$

为真实类中心的估计量。式(7)的目标函数可改写为

$$\frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i\in\mathcal{I}_l}\xi_i+C_2\sum_{i\in\mathcal{I}_u}(\xi_i+p_{i-l})-C_2(u_+\boldsymbol{w}'\hat{\boldsymbol{m}}_+-u_-\boldsymbol{w}'\hat{\boldsymbol{m}}_-+(u_+-u_-)b)\quad(8)$$

不难发现,式(8)只与类中心有关,与标记指派无关。换言之,S3VM 存在等价形式,它只与未标记示例的类中心估计有关。这启发了我们将未标记示例的真实类中心代入式(8),与得到未标记示例真实标记的监督 SVM 进行比较。记 y_i^* 表示未标记样本 $\boldsymbol{x}_i$ 的真实标记,$i\in\mathcal{I}_u$,$\boldsymbol{m}_+$ 和 $\boldsymbol{m}_-$ 为未标记示例上真实的类中心,

$$\boldsymbol{m}_+=\frac{1}{u_+}\sum_{y_i^*=1}\phi(\boldsymbol{x}_i),\quad \boldsymbol{m}_-=\frac{1}{u_-}\sum_{y_i^*=-1}\phi(\boldsymbol{x}_i)$$

将真实类中心 $\boldsymbol{m}_+$ 和 $\boldsymbol{m}_-$ 代入式(8),得到 MeanS3VM 的优化目标,

$$\begin{aligned}\min_{w,b,\xi,p}\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i\in\mathcal{I}_l}\xi_i+C_2\sum_{i\in\mathcal{I}_u}(\xi_i+p_{i-l})-C_2(u_+\boldsymbol{w}'\boldsymbol{m}_+-u_-\boldsymbol{w}'\boldsymbol{m}_-+(u_+-u_-)b)\\ \text{s.t.}\quad & y_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\quad i\in\mathcal{I}_l,\\ & \boldsymbol{w}'\phi(\boldsymbol{x}_i)+b\leqslant p_{i-l},\quad -\boldsymbol{w}'\phi(\boldsymbol{x}_i)-b\leqslant p_{i-l},\quad p_{i-l}\geqslant 1-\xi_i,\quad i\in\mathcal{I}_u\\ & \xi_i\geqslant 0,\ i\in\mathcal{I}_l\cup\mathcal{I}_u;\ \sum_{i\in\mathcal{I}_u}\operatorname{sgn}(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b)=r\end{aligned}$$

则 MeanS3VM 与监督 SVM 的比较有如下结果。

定理 5(可分情况) 当训练示例可分时,MeanS3VM 的损失函数等价于 SVM。

定理 6(不可分情况) 当训练示例不可分时,MeanS3VM 的损失函数不超过 SVM 损失的两倍。

图 3 给出了 MeanS3VM 与 SVM 损失函数的示意图,可见两者损失函数非常相似。以上分析表明,当未标记示例的真实类中心已知时,MeanS3VM 近似于得到真实标记指派的监督 SVM。换言之,监督 SVM 的性能可以近似地通过 MeanS3VM 加上类中心得到。这个分析启发了 S3VM 只需要估计类中心,而不需要估计所有未标记示例的标记指派。

类中心的估计存在高效算法。受大间隔准则的启发,MeanS3VM 采用一种大间隔的做法对类中心进行估计。该做法的直观意义是,若正负例的类中心间隔越大,则两类示例分得越开,越容易区分。形式化讲,MeanS3VM 考虑优化如下目标:

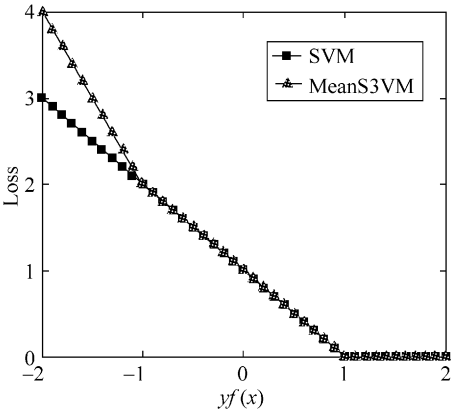

图 3　MeanS3VM 与 SVM 的损失函数

$$\begin{aligned}\min_{\boldsymbol{d}\in\Delta}\min_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_1\sum_{i=1}^{l}\xi_i - C_2\rho \\ \text{s.t.} \quad & y_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i)+b) \geqslant 1-\xi_i, \quad i=1,\cdots,l, \\ & \frac{1}{u_+}\Big(\boldsymbol{w}'\sum_{j=l+1}^{l+u} d_{j-l}\phi(\boldsymbol{x}_j)\Big)+b \geqslant \rho, \\ & \frac{1}{u_-}\Big(\boldsymbol{w}'\sum_{j=l+1}^{l+u}(1-d_{j-l})\phi(\boldsymbol{x}_j)\Big)+b \leqslant -\rho\end{aligned} \tag{9}$$

其中 $\Delta=\Big\{\boldsymbol{d}\,|\,d_i\in\{0,1\},\sum_{i=1}^{u}d_i=u_+\Big\}$，$\frac{1}{u_+}\Big(\sum_{j=l+1}^{l+u}d_{j-l}\phi(\boldsymbol{x}_j)\Big)$ 和 $\frac{1}{u_-}\Big(\sum_{j=l+1}^{l+u}(1-d_{j-l})\phi(\boldsymbol{x}_j)\Big)$ 分别表示正、负例的类中心。注意到，式(9)中的约束个数与未标记示例的数量无关，因此类中心估计可以高效求解。MeanS3VM 提出两种优化技术用于求解式(9)：基于多核学习的凸松弛技术(记为 MeanS3VM-mkl)以及基于交替优化的局部优化技术(记为 MeanS3VM-iter)。

表 1 给出了 MeanS3VM 与 TSVM、Laplician SVM 在 17 个数据集上的时间开销结果。关于 MeanS3VM 更多实验结果请参阅论文[20]。实验机器的配置为：双核 2GHz Intel Xeon (R)处理器、Windows XP 系统、4GB 内存。可以看出，MeanS3VM 方法要快于 TSVM 和 Laplician SVM，并且随着数据规模的增大，MeanS3VM 方法的优势更加明显。比如，在 text 等超过 1000 个训练示例的数据集上，MeanS3VM 方法可以比 TSVM 方法快 100 倍以上，比 Laplician SVM 方法快 10 倍以上。

表 1 时间开销/s

Data set(N,d)	TSVM	Laplician SVM	MeanS3VM-iter	MeanS3VM-mkl
BCI(400,117)	73.88	**0.19**	0.27	2.45
Text(1500,11960)	6181.12	17.27	**0.55**	14.12
g241d(1500,241)	596.23	5.88	**0.53**	0.94
g241c(1500,241)	552.19	7.08	**1.77**	2.17
Digitl(1500,241)	1222.90	6.54	**0.50**	0.83
USPS(1500,241)	560.05	7.48	**0.58**	1.25
house(232,16)	3.19	**0.09**	**0.09**	0.77
heart(270,9)	13.12	**0.06**	0.09	0.52
vehicle(435,36)	34.46	0.20	**0.11**	0.65
wdbc(569,14)	123.02	**0.29**	0.50	0.56
isolet(600,51)	62.10	0.55	**0.19**	0.97
austra(690,15)	44.37	0.40	**0.26**	0.80
optdigits(1143,42)	114.93	1.53	**0.39**	0.94
ethn(2630,30)	355.30	11.70	**1.09**	2.16
sat(3041,36)	494.38	18.78	**1.08**	1.86
(1,2)(2000,3736)	2176.65	13.46	**0.81**	3.33
(1,3)(2000,3757)	2151.67	13.48	**0.75**	3.09

注：每个数据集上最少的时间开销被加粗表示。每个数据集后面的数字分别表示示例总数量(N)及示例维度(d)。

3.3 好：用于提供性能保障的安全半监督支持向量机

以往半监督学习研究通常认为当有标记示例数目较少时，利用大量未标记示例改善学习性能。然而在不少情况下，半监督学习利用未标记示例不仅不能带来性能改善，有时还会导致性能下降[8,10,15,29,46-49]。这一现象，势必严重影响了半监督学习方法在现实任务中的应用，特别是对于那些对未标记示例使用要求高可靠性的现实任务——它们不希望看到“新技术”(半监督学习)不如“旧技术”(简单监督学习)。因此，有必要对安全半监督学习方法进行研究。这里“安全”指的是，半监督学习方法通常可以取得改善的性能，即使在最差情况下也不会显著差于只利用少量有标记示例得到的监督学习方法。

在半监督学习早期文献中，学者们对半监督学习导致性能下降的原因有所讨论。比如，F. G. Cozman 等人[47]对生成式模型方法导致性能下降的现象进行分析，指出当生成

式模型假设不正确时，即使有标记示例上的性能不断变好，生成式方法的泛化性能也可能会不断变差；然而在没有充分领域知识的情况下，如何得到正确的模型假设仍是个非常困难的问题。对于基于图的方法，研究者指出，图的构建是方法能否取得成功的关键[3]；然而如何在一般意义下构建一个好图仍是个公开问题。基于分歧的方法迭代地借助未标记示例的预估标记指派来完成学习过程。当预估标记指派含有噪声时，学习方法可能会导致性能下降。基于这一观察，M. Li 和 Z.-H. Zhou 对预估标记指派加以编辑，力图过滤掉不可靠的指派[50]。对于 S3VM，虽然有研究指出在小规模数据上，S3VM 得到最优解时可以得到不错的性能[29]，然而如何防止 S3VM 性能下降还没有相关解决方案。

如何构建安全的 S3VM？观察到在 S3VM 中，未标记示例在提供有用数据分布信息的同时，也可能提供了有误导甚至错误的信息，我们首先提出一种做法 S3VM-us[21]。它的基本想法是只利用那些有助于改善性能的未标记示例，而对于那些利用起来风险比较高的未标记示例应该不予理会。如何识别高风险未标记示例？我们首先提出两种底线方案（S3VM-c 和 S3VM-p）。S3VM-c 受聚类假设[4]的启发，它通过聚类（比如 k 均值法）将训练示例划分为多个簇。如果 S3VM 比 SVM（仅利用少量有标记示例）在某个簇的示例上取得了更一致的预测结果，那么该簇中示例被认为有助于提升性能，赋予它们 S3VM 的标记指派，否则被认为是高风险的，而不予理会。S3VM-p 受基于图的方法启发，它根据基于图的方法所提供的置信度估计来判断未标记示例的风险程度。S3VM-c 和 S3VM-p 均在一定程度上改进了 S3VM 的不安全程度，但它们均存在一些不足。比如，S3VM-c 是一个局部方法，没有考虑簇间联系，损失掉一些有用信息；S3VM-p 严重依赖置信度估计，而置信度估计对标记指派初始化非常敏感[51]。为克服以上困难，S3VM-us 方法采用层次式聚类来选择高风险的未标记示例，因为层次式聚类过程天然考虑了簇间联系，同时层次式聚类是个无监督方法，不受标记指派初始化的影响。实验结果证实了 S3VM-us 比 S3VM-c 和 S3VM-p 更有效地提高了 S3VM 的安全性。

然而 S3VM-us 在一些情况下性能提升十分有限，此外它是个启发式做法，缺乏基本的数学模型。为克服这些困难，我们提出一种安全半监督支持向量机学习方法 S4VM（safe S3VM）[21]。S4VM 重新审视 S3VM 的内在低密度假设，试图从低密度假设本身发现 S3VM 性能下降的原因。事实上，低密度假设确实存在不安全的地方。比如，当给定少量有标记示例和大量未标记示例时，整个数据分布可能会存在多个低密度划分（如图 4 所示）。在缺乏充分领域知识对这些低密度划分加以区分的情况下，错误选择其中一个划分可能会严重影响到最终的性能，甚至会导致 S3VM 不如简单监督 SVM（仅利用少量有标记示例）。根据这个观察，S4VM 考虑在最坏情况下最大化性能的提升。可以证明，当 S3VM 要求的低密度假设成立时，S4VM 的结果是安全的，而且取得了最大的性能提升。

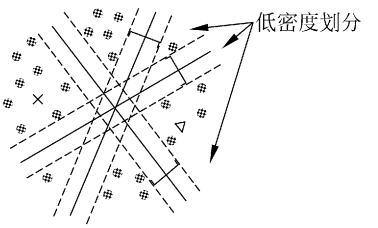

图 4 多个低密度划分的示意图

注：叉和三角形表示少量有标记示例，灰点表示大量未标记示例。

具体来说，假设$\{\hat{\boldsymbol{y}}_t\}_{t=1}^{T}$为多个候选低密度划分在未标记示例上的标记指派，令$\boldsymbol{y}^*$为未标记示例的真实标记指派，$\boldsymbol{y}^{svm}$为监督 SVM 在未标记示例上的标记指派。对于任意未标记示例的标记指派$\boldsymbol{y}$，根据真实标记指派，可知道在哪些未标记示例上$\boldsymbol{y}$优于$\boldsymbol{y}^{svm}$，记这部分的性能为$earn(\boldsymbol{y},\boldsymbol{y}^*,\boldsymbol{y}^{svm})$；类似地，可知道在哪些未标记示例上$\boldsymbol{y}$不如$\boldsymbol{y}^{svm}$，记这部分的性能为$lose(\boldsymbol{y},\boldsymbol{y}^*,\boldsymbol{y}^{svm})$。S4VM 通过下式最大化性能提升，

$$\max_{\boldsymbol{y}\in\{\pm1\}^u} earn(\boldsymbol{y},\boldsymbol{y}^*,\boldsymbol{y}^{svm})-\lambda lose(\boldsymbol{y},\boldsymbol{y}^*,\boldsymbol{y}^{svm}), \tag{10}$$

其中λ是参数用于权衡用户对风险的容忍程度。为了便于说明，记

$$J(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{y}^{svm}) = earn(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{y}^{svm})-\lambda lose(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{y}^{svm})$$

然而，式(10)中未标记示例的真实标记指派y^*未知，难以直接计算(事实上，当$\boldsymbol{y}^*$已知时，很明显$\boldsymbol{y}=\boldsymbol{y}^*$为最优结果)。因此，需要对真实标记指派$\boldsymbol{y}^*$做假设。根据 S3VM 内在低密度假设，真实标记指派$\boldsymbol{y}^*$应该由其中一个低密度划分得到，即$\boldsymbol{y}^*\in\mathcal{M}\triangleq\{\hat{\boldsymbol{y}}_t\}_{t=1}^{T}$。在无法进一步区分候选低密度划分的情况下，S4VM 考虑最坏情况，即最大化最坏情况下的性能提升，

$$\bar{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}\in\{\pm1\}^u} \min_{\hat{\boldsymbol{y}}\in\mathcal{M}} J(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{y}^{svm}) \tag{11}$$

记$\bar{\boldsymbol{y}}$为式(11)的最优解，则$\bar{\boldsymbol{y}}$有以下结果：

定理 7 当$\boldsymbol{y}^*\in\mathcal{M}\triangleq\{\hat{\boldsymbol{y}}_t\}_{t=1}^{T}$且$\lambda\geqslant1$时，$\bar{\boldsymbol{y}}$的精度不会差于$\boldsymbol{y}^{svm}$。

定理 7 揭示了当 S3VM 要求的低密度假设成立时，S4VM 是安全的。值得一提的是，定理 7 给出的是充分条件而非必要条件。换言之，即使定理 7 的条件不成立，S4VM 仍可能是安全的。比如，下面的推论说明，即使得不到式(11)的最优解，S4VM 也是安全的。

推论 2 当$\boldsymbol{y}^*\in\{\hat{\boldsymbol{y}}_t\}_{t=1}^{T}$且$\lambda\geqslant1$时，对于任意$\boldsymbol{y}$，只要满足$\min_{\hat{\boldsymbol{y}}\in\mathcal{M}} J(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{y}^{svm})\geqslant0$，则它的精度不会差于$\boldsymbol{y}^{svm}$。

S4VM 还需关心性能提升。下面的推论说明，S4VM 已取得最坏情况下最大的性能提升。

推论 3 当 $\boldsymbol{y}^* \in \{\hat{\boldsymbol{y}}_t\}_{t=1}^T$ 且 $\lambda=1$ 时，则 $\bar{\boldsymbol{y}}$ 得到最坏情况下最大的性能提升。

式(11)存在有效算法。进一步展开 $earn(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{y}^{svm})$ 和 $lose(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{y}^{svm})$ 可发现它们均可形式化为关于 $\boldsymbol{y}$ 的线性函数，因此，式(11)是个线性规划问题。但 $\boldsymbol{y}$ 的取值范围为整数，难以得到最优解。注意到推论 2 证明最优解不是必要的，因此，式(11)可如下求解：①将整数约束凸松弛成 $[-1,1]^u$，求解线性规划，得到最优解；②将该最优解投影到与它最近的整数解；③如果该整数解的目标值小于 0，则输出 $\boldsymbol{y}^{svm}$，否则输出该整数解。不难发现，这个解法满足推论 2 的要求。

如何得到候选的低密度划分？S4VM 期望候选的低密度划分具有大的间隔，同时彼此间具有较大的差异性。记 $h(f, \hat{\boldsymbol{y}})$ 为 S3VM 的目标函数，S4VM 优化如下目标：

$$\min_{\{f_t, \hat{\boldsymbol{y}}_t \in \mathcal{B}\}_{t=1}^T} \sum_{t=1}^{T} h(f_t, \hat{\boldsymbol{y}}_t) + M\Omega(\{\hat{\boldsymbol{y}}_t\}_{t=1}^T) \tag{12}$$

其中 T 为候选低密度划分数目，Ω 为低密度划分差异性的惩罚项，定义为

$$\Omega(\{\hat{\boldsymbol{y}}_t\}_{t=1}^T) = \sum_{1 \leqslant t \neq i \leqslant T} \delta\left(\frac{\hat{\boldsymbol{y}}_t' \hat{\boldsymbol{y}}_i}{u} \geqslant 1-\zeta\right)$$

其中 δ 为指示函数，$\zeta=0.5$ 是常数。Ω 可采用其他类似定义。$M=10^5$ 是一个大常数。不难发现，优化式(12)使得候选划分具有大间隔，同时具有大的差异性。式(12)是个非凸优化，S4VM 采用两种实现：①模拟退火的全局搜索，这种实现理论上可得到更好的解，但效率比较低下；②启发式的代表性采样，这个方法也取得了不错的效果，同时提高了效率。

表 2 给出 S4VM 与比较方法在 20 个数据集上高斯核的精度结果，这里 S4VM 指采用启发式的代表性采样方法。关于 S4VM 更多结果请参阅论文[21]。对于每个数据，随机抽取 10 个有标记示例，其类别比例与整个数据集的比例相当，余下数据作为未标记示例，实验重复 30 次，报告平均性能。S4VM 与 SVM(仅利用少量有标记示例)和 S3VM(用 TSVM[18]实现)进行比较。S4VM 还与 S4VM 的三个变种进行比较：S3VM^{best} 输出候选低密度划分中最好的性能(此方法在实践中不可行，此处列出仅作为讨论)；S3VM^{min} 输出候选低密度划分中目标值最小的一个；S3VM^{com} 集成候选低密度划分(在缺乏领域知识的情况下，候选划分的权重相同)。可以看出，S4VM 取得了与 S3VM 高度可比的性能。S4VM 与 S3VM 都在 11 个数据集上显著优于 SVM；更重要的是，S3VM 在 6 个数据集上显著地差于 SVM，而 S4VM 没有这个现象。S3VM^{min} 和 S3VM^{com} 仍会多次导致性能下降。Wilcoxon sign test (95%置信度) 的结果表明 S4VM 在高斯核显著优于 SVM，而 S3VM、S3VM^{min} 和 S3VM^{com} 没有这个显著性。S3VM-us 虽然较 S3VM-c、S3VM-p 提升了 S3VM 的安全程度，但它的性能提升比较小，经常不如 S4VM。

值得一提的是，低密度划分假设通常不成立（参见 S3VM^{best} 的性能）。即使在这种情况下，S4VM 仍然取得不错的性能。一个可能的原因是定理 7 只给出充分条件而非必要条件。另外一个原因是根据集成学习的理论[52]，低密度划分的差异性可能有助于提高性能。

表 2 S4VM 与比较方法在高斯核上的精度结果

RBF	SVM	S3VM	S3VM-c	S3VM-p	S3VM-us	S3VM^{best}	S3VM^{min}	S3VM^{com}	S4VM
austra	69.2±7.1	70.4±11.9	69±8.5	69.1±7.2	69.2±7.5	**76.3±10.1**	70.8±12	70.1±12.3	**70.6±8.8**
australian	71.4±6.8	**77.7±10.5**	**72.8±7.9**	**71.9±6.7**	71.2±7.2	**80.5±6.7**	71.1±14.4	71.3±10.6	71.2±7.1
breastw	95.2±2.4	93.2±0.4	94.9±2.1	95±2.4	95±2.4	**96.5±0.4**	**96.4±0.4**	**96.3±0.7**	**95.9±1.5**
cleanl	64.3±4.9	60.8±6.9	63.8±5.2	63.9±4.7	64.7±5	65.4±4.5	57.9±5.3	60.3±5.9	64.4±4.4
diabetes	66.1±4.4	65.1±7	66.3±4.2	**66.4±4.3**	**66.4±4.4**	66±5.7	65.2±5.5	64.8±5.4	65.5±5.5
haberman	65.8±5.4	61±3.7	65.8±5.2	65.9±5.3	65.7±5.4	65±3.1	62.5±3.3	65.4±3.6	66±4.2
heart	72.2±5.5	**73.9±5.1**	**72.9±5.5**	72.6±5.3	72.4±5.9	**75±5.1**	**73.4±5.8**	**73.4±6.1**	**73.5±5.6**
house-votes	87.9±2.4	**89.1±2**	**88.4±2.2**	**88.1±2.3**	**88.5±2.2**	**89.4±2.2**	88.5±2	88.5±2.4	88.6±2.2
house	89.3±2.3	**90.4±1.8**	**89.7±2.1**	**89.4±2.2**	**89.8±2.1**	**90.6±2.5**	89.2±2.4	89.5±2.7	**89.8±2.4**
ionosphere	79.7±5.6	**83.4±5.6**	**80.4±5.4**	79.9±5.6	**80±5.7**	**87.2±6.5**	**82.8±6.5**	**82±6.4**	**84.3±6.6**
isolet	91.9±3.1	**99.7±0.1**	**96.8±2.6**	**92.6±2.8**	**92.6±2.8**	**99.2±0.3**	**98.5±0.7**	**98.6±0.5**	**98.6±0.6**
liverDisorders	55.5±4.7	54.1±4.7	54.8±4.5	55.6±4.7	55.4±4.6	55.6±4.7	55.4±4.7	55.1±4.7	55.4±4.7
optdigits	94.6±3.2	**99.7±0.1**	**97.3±2.5**	**95.1±2.8**	**96.6±1.5**	**99.8±0.1**	**99.6±0.9**	**97.5±2.2**	**98±2**
vehicle	80.3±6.2	**84.8±11.5**	**83.2±8.1**	**81.1±6.2**	**82.7±7.2**	**91.1±5.7**	**87.5±8.4**	**84.6±8.7**	**85±7.5**
wdbc	85.3±5.1	**90.7±2.1**	**88.2±4.6**	**85.9±4.9**	85.6±4.9	**91.9±3.7**	**91.2±3.6**	**90.8±3.7**	**90.7±4.1**
digitl	75.4±8	**90.1±3.2**	**80.7±9.2**	**77.1±7.1**	**75.9±8**	**91.8±2**	**88.5±1.5**	**88.5±3.8**	**79.1±5.1**
usps	80±0	67.9±5.9	80±0	80±0	80±0	77.9±4.7	65.9±0.4	78.2±3.9	80±0
coil	62±6.4	61.6±6.1	62.5±6.8	61.2±6.4	**62.1±6.3**	**72.5±7.9**	64.4±9.8	59.9±8.2	61.9±6.4
bci	51.5±2.5	50±2	50.2±2.4	51.4±2.4	51.4±2.4	52.1±2.1	49.8±1.7	48.9±3	50.8±2.6
g241c	59.8±2.7	**60.8±2.8**	**60.5±2.9**	60±2.8	59.9±2.7	**63.7±2.6**	**62.2±3.5**	52.1±4.7	**60.2±2.8**
Win/Tie/Loss against SVM		11/3/6	10/8/2	9/9/2	10/9/1	**14/6/0**	9/6/5	8/8/4	**11/9/0**

注：对于每个数据集，加粗（下画线）表示该性能显著优于（差于）SVM。Win/Tie/Loss 的结果（双边成对 t-检验，95%置信度）在最后一行给出。最少次性能显著下降的方法被加粗表示。

3.4 省：用于代价抑制的代价敏感半监督支持向量机

以往 S3VM 假定训练示例错误分类的代价均衡。然而在许多现实任务中，不同错误

导致的代价往往不均衡，甚至有巨大差别。例如，在医疗诊断中，将一个病人错误诊断为健康人的代价，要远远高于将一个健康人诊断成一个病人的代价。在欺诈检测中，漏掉一个欺诈的代价将远远高于错判一个正常交易的代价。同时，这些现实任务同样面临有标记示例稀少的问题。比如，困难病状的诊断要耗费医生大量精力和相应的设备损耗；欺诈的检测需要耗费大量的人力资源。因此，在许多现实任务中，代价不均衡与有标记示例稀少通常同时存在。为使 S3VM 能够适用于更多实际应用，有必要对 S3VM 处理非均衡代价的能力进行研究，发展能够抑制总体代价的学习方法。

在以往研究中，代价敏感学习(cost sensitive learning)是监督学习环境下处理非均衡代价一个主流方向[53-57]，其目标是找到一个最佳决策使整体代价最小。值得一提的是，在学习过程伴随着多种不同类型的代价中[58]，其中研究最多的、最常见的是错分代价(misclassification cost)。进一步地，错分代价划分为类别相关代价 (class dependent cost)与样本相关代价(example dependent cost)两种。前者只与类别有关，与具体示例无关；后者与类别、具体示例都有关。

类别相关代价由于它实际应用中获得相对容易，得到了更多的关注。目前处理类别相关代价的方法大致可分为分类器驱动和 Rescaling 两类。前者对经典机器学习分类器加以改良适应代价敏感需求，如代价敏感决策树[55]、代价敏感神经网络[59]、代价敏感 AdaBoost[60]、代价敏感 SVM[61]等。后者通过对不同类别数据加以调整，使类别之间的错分代价均衡，从而可利用已有代价均衡方法进行求解，如样本加权法(example weighting)[55,57]、采样法(sampling)[54,57]、阈值移动法(threshold moving)[53,57]等。然而，这些方法均是监督学习方法，没有考虑未标记示例。最近有少量代价敏感方法[62-64]考虑到了未标记示例，然而它们都工作在主动学习[65-67]框架下。

为了赋予 S3VM 处理不均衡代价的能力，我们提出一种代价敏感半监督支持向量机学习方法 CS4VM(cost-sensitive S3VM)[22]。CS4VM 显式地将未标记示例的错分代价写进优化目标，通过求解 SVM 最优超平面和未标记示例上的标记指派，使得有标记与未标记示例上整体代价最小化。但 CS4VM 的损失函数不再是连续函数，以往 S3VM 基于连续函数的优化技术难以适用。为克服这个困难，我们通过分析 CS4VM 优化目标，证明了 CS4VM 可以通过类中心估计技术近似求解，从而得到 CS4VM 的有效算法。

具体来说，记 $c(+1)$和 $c(-1)$为示例被错分为正类与负类所导致的代价，$\hat{\boldsymbol{y}}=[\hat{y}_i;i\in\mathcal{I}_u]$ 为未标记示例的标记指派，CS4VM 优化如下目标，期望所有示例的总体代价最小化：

$$\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\min_{f}\frac{1}{2}\|f\|_{\mathcal{H}}^2+C_1\sum_{i\in\mathcal{I}_l}c(y_i)\ell(y_i,f(\boldsymbol{x}_i))+C_2\sum_{i\in\mathcal{I}_u}c(\hat{y}_i)\ell(\hat{y}_i,f(\boldsymbol{x}_i)),\qquad(13)$$

其中 $\ell(y,f(\boldsymbol{x}))=\max\{0,1-yf(\boldsymbol{x})\}$为 SVM 损失函数。$\mathcal{B}=\{\hat{\boldsymbol{y}}\,|\,\hat{y}_i\in\{\pm1\},\hat{\boldsymbol{y}}'\mathbf{1}=r\}$表示

候选的标记指派集合，$\hat{\boldsymbol{y}}'\mathbf{1}=r$ 对应于 S3VM 的平衡约束。图 5 给出 CS4VM 在未标记示例上损失函数的示意图。可见，当 $c(1)=c(-1)$，即代价均衡时，CS4VM 的损失函数退化成传统 S3VM 的损失函数[17,18]。当 $c(1)\neq c(-1)$，即代价不均衡时，CS4VM 的损失函数不再是一个连续函数。以往 S3VM 基于连续函数的学习方法[29]不再适用。

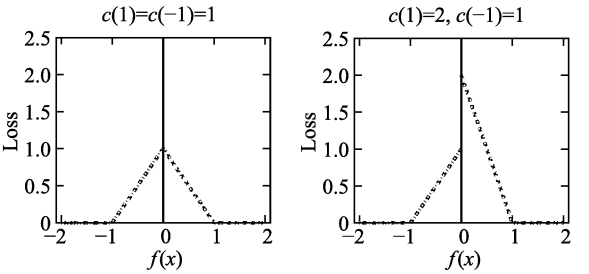

图 5 CS4VM 在未标记示例上损失函数的示意图

为克服这个困难，通过对 CS4VM 的优化目标进行分析，发现其可以改写为下式，

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{p}^{\pm}}\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i\in\mathcal{I}_l}\ell(y_i,\boldsymbol{w}'\boldsymbol{x}_i+b)+C_2(\mathbf{1}'\boldsymbol{p}^{+}+\mathbf{1}'\boldsymbol{p}^{-})\\ & -C_2(\boldsymbol{w}'(u_+\,c(+1)\hat{\boldsymbol{m}}_+-u_-\,c(-1)\hat{\boldsymbol{m}}_-)-n_1+n_2 b)\\ \text{s. t.}\quad & c(+1)f(\boldsymbol{x}_i)-c(+1)\leqslant p_i^{+},\\ & -c(-1)f(\boldsymbol{x}_i)-c(-1)\leqslant p_i^{-},\\ & p_i^{+},p_i^{-}\geqslant 0,\ \forall\, i\in\mathcal{I}_u;\ \sum_{i\in\mathcal{I}_u}\operatorname{sgn}(f(\boldsymbol{x}_i))=r\end{aligned}\tag{14}$$

其中

$$\hat{\boldsymbol{m}}_+=\frac{1}{u_+}\sum_{i\in\mathcal{I}_u,\,f(\boldsymbol{x}_i)\geqslant 0}\phi(\boldsymbol{x}_i),$$

$$\hat{\boldsymbol{m}}_-=\frac{1}{u_-}\sum_{i\in\mathcal{I}_u,\,f(\boldsymbol{x}_i)<0}\phi(\boldsymbol{x}_i)$$

为真实类中心的估计。这启发了我们对 MeanS3VM 的理论结果加以扩展从而对 CS4VM 近似求解，得到如下分析结果。

定理 8 记 f^* 为 CS4VM 的最优解。当未标记示例的损失不大时，即 $y_i^* f^*(\boldsymbol{x}_i)\geqslant -1,\forall\, i\in\mathcal{I}_u$，CS4VM 等价于真实标记指派的监督代价敏感 SVM(CS-SVM)。对于余下情况，CS4VM 的损失函数可被 CS-SVM 损失函数界定。具体来说，记 $\hat{\ell}(\boldsymbol{x}_i)$ 为 CS4VM 的损失函数，$\ell(y_i^*,f(\boldsymbol{x}_i))$ 为 CS-SVM 的损失函数，两者满足

$$\hat{\ell}(\boldsymbol{x}_i) \leqslant \frac{c(1)+c(-1)}{c(y_i^*)}\ell(y_i^*, f(\boldsymbol{x}_i))$$

以上结果说明，当未标记示例的真实类中心给定时，CS4VM 将近似于得到真实标记指派的监督代价敏感 SVM(CS-SVM)。从而，CS4VM 同样可以通过类中心估计技术得到有效算法。

表 3 给出 CS4VM 与比较方法在 20 个数据集上的总体代价结果。关于 CS4VM 更多实验结果可参阅论文[22]。每个数据集上 50%数据用于训练，余下数据用于测试。每个训练集随机选取 10 个有标记示例。$c(-1)$固定为 1，$c(+1)$设置为[0，1000]的一个随机数。实验重复 100 次，报告平均结果。这个设置考察了学习方法性能对代价的稳定性。CS4VM 与以下方法进行比较：代价敏感支持向量机 CS-SVM(仅考虑少量有标记示例)；TSVM 和 Laplacian SVM 两种传统 S3VM 方法。可以看出，CS4VM 有效降低了总体错误代价。双边成对 t-检验(95%置信度)结果表明：CS4VM 在 16 个数据集上显著优于 Laplacian SVM，在 17 个数据集上显著优于 TSVM。与 CS-SVM 相比，利用了未标记示例代价信息的 CS4VM 在 14 个数据集上取得了显著的性能提升，而 S3VM 方法则只在 5 个和 3 个数据集上取得显著的性能提高。Wilcoxon 符号检验（95%置信度）结果表明，CS4VM 显著优于 CS-SVM，Laplacian SVM 和 TSVM，而 Laplacian SVM 与 TSVM 均不显著优于 CS-SVM 方法。进一步可以发现，CS4VM 在 16 个数据集上将 Laplacian SVM 的总体错误代价减少了 1/5 以上，在 18 个数据集上将 TSVM 的总体错误代价减少了 1/5 以上。

表 3　CS4VM 与比较方法的总体代价和方差（$\times 10^3$）

Data set	Supervised CS-SVM	Laplacian SVM	TSVM	CS4VM
Heart-Statlog	9.745±6.906	**1.640±2.708**	10.28±6.985	6.261±4.920
lonosphere	17.02±12.84	27.19±17.03	11.98±7.749	**7.811±5.130**
Live Disorder	**0.178±0.388**	11.37±17.29	12.01±7.844	5.507±1.018
Echocardiogram	3.955±2.609	**1.314±2.305**	4.129±2.610	3.576±2.391
Spectf	6.022±6.451	**2.974±5.514**	12.52±8.384	**2.873±2.533**
Australian	23.63±19.06	25.01±27.15	24.80±19.00	**15.98±11.86**
Cleanl	17.96±13.44	20.63±14.88	21.97±14.26	**13.47±9.942**
Diabetes	**5.772±10.84**	**6.162±14.11**	32.08±19.30	10.01±8.946
German Credit	30.17±22.28	30.54±26.16	26.48±18.83	**18.63±13.30**
House Votes	8.594±7.187	9.693±8.515	12.50±8.551	**6.206±4.644**

续表

Data set	Supervised CS-SVM	Laplacian SVM	TSVM	CS4VM
Krvskp	144.9±87.03	131.5±81.30	158.0±90.43	**92.42±52.09**
Ethn	**9.919±16.25**	119.3±85.15	74.90±64.07	16.14±11.84
Heart	0.615±1.188	1.962±6.346	6.908±4.770	**0.127±0.205**
Texture	4.094±6.755	5.748±6.489	2.512±4.668	**0.045±0.205**
House	1.760±1.505	1.325±1.415	1.458±1.479	**0.935±1.061**
Isolet	4.976±4.218	7.207±6.382	0.943±1.394	**0.420±0.670**
Optdigits	6.642±6.881	4.025±4.177	**1.097±1.951**	**0.773±1.197**
Vehicle	1.978±3.812	18.70±26.50	7.191±7.800	**1.002±1.667**
Wdbc	**0.127±0.125**	32.92±38.52	11.33±8.367	0.264±0.415
Sat	**3.404±7.363**	6.968±10.01	**2.122±9.839**	**2.521±9.407**
CS4VM:W/T/L	14/2/4	16/1/3	17/3/0	—

注：每个数据集上最好的性能及与其没有显著性差别的性能被加粗表示。最后一行给出 CS4VM 与其他方法 win/tie/loss（双边成对 t-检验，95%置信度）的结果。

4 结语

传统机器学习技术通过对有标记示例的学习来构建模型，为了获得强泛化能力，通常需要有大量的有标记示例。在很多现实任务中，虽然很容易获得大量未标记示例，但是获取数据的标记却相对困难，因为标记过程需要花费人力物力资源；因此，如何有效地利用未标记数据来提高泛化性能，成为机器学习领域的一个关键问题。半监督学习是该方面的主流研究方向之一，而半监督支持向量机（semi-supervised SVM，S3VM）则是半监督学习中的一类主流范型。经过十年的研究，S3VM 已经取得了很多进展，并且在众多领域得以成功应用。然而，该范型所涉及的一些重要问题，例如数据规模、学习效率、性能保障、代价抑制等，仍有待解决。我们对这些问题进行研究，取得了一些进展：针对传统 S3VM 方法难以处理大规模数据的问题，提出了 WELLSVM 方法，理论上证明了其优化解的全局性与紧致性，实验验证了该方法能处理的数据规模达到传统方法的 10 倍以上；针对传统 S3VM 方法学习速度慢的问题，提出了 MeanS3VM 方法，理论上证明了该方法具有强逼近能力，实验验证了其计算效率可以比传统方法快 10 倍以上；针对传统 S3VM 方法利用未标记示例后常会出现性能下降的问题，提出了 S4VM 方法，给出了其性能保

障条件，实验验证了该方法能将性能下降概率从传统方法的15%减少到1%；针对传统S3VM方法难以处理非均衡代价的问题，提出了CS4VM方法，证明了其具有处理非均衡代价的能力，实验验证了该方法通常能比传统方法的代价减少20%以上。

半监督学习只是标记缺失的一种场景。现实应用存在多种不同类型标记缺失的场景，如弱标记学习[68]、众包学习[69]等。如何“多、快、好、省”地对那些场景的训练示例进行学习还有许多亟待研究的问题。

参考文献

[1] Mitchell T. *Machine Learning*. New York: McGraw-Hill, 1997.

[2] Vapnik V N. *Statistical Learning Theory*. New York: Wiley, 1998.

[3] Zhu X. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[4] Chapelle O, Schölkopf B, Zien A, eds. *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[5] Zhou Z-H, Li M. Semi-supervised learning by disagreement. *Knowledge and Information Systems*. 2010, 24(3): 415-439.

[6] Shahshahani B, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 1994, 32(5): 1087-1095.

[7] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Mozer M, Jordan M I, Petsche T, eds. *Advances in Neural Information Processing Systems 9 (NIPS'97)*, Cambridge, MA: MIT Press, 1997: 571-577.

[8] Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000, 39(2-3): 103-134.

[9] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency. In: Thrun S, Saul L, Schölkopf B, eds. *Advances in Neural Information Processing Systems 16 (NIPS'04)*, Cambridge, MA: MIT Press, 2004: 321-328.

[10] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, San Francisco, CA, 2001: 19-26.

[11] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, Washington, DC, 2003: 912-919.

[12] Belkin M, Niyogi P. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 2004, 56(1-3): 209-239.

[13] Belkin M, Niyogi P, Sindwani V. On manifold regularization. In: *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics* (*AISTATS'05*), Savannah Hotel, Barbados, 2005: 17-24.

[14] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: *Proceedings of the 11th Annual Conference on Computational Learning Theory* (*COLT'98*), Wisconsin, MI, 1998: 92-100.

[15] Zhou Z-H, Li M. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(11): 1529-1541.

[16] 周志华. 基于分歧的半监督学习. 自动化学报, 2013, 39(11): 1871-1878.

[17] Bennett K P, Demiriz A. Semi-supervised support vector machines. In: Kearns M J, Solla S A, Cohn D A, eds. *Advances in Neural Information Processing Systems 11* (*NIPS'99*). Cambridge, MA: MIT Press, 1999: 368-374.

[18] Joachims T. Transductive inference for text classification using support vector machines. In: *Proceedings of the 16th International Conference on Machine Learning* (*ICML'99*), Bled, Slovenia, 1999: 200-209.

[19] Li Y-F, Tsang I W, Kwok J T, Zhou Z-H. Convex and scalable weakly Labeled SVMs. *Journal of Machine Learning Research*, 2013, 14: 2151-2188.

[20] Li Y-F, Kwok J T, Zhou Z-H. Semi-supervised learning using label mean. In: *Proceedings of the 26th International Conference on Machine Learning* (*ICML'09*), Montreal, Canada, 2009: 633-640.

[21] Li Y-F, Zhou Z-H. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(1): 175-188.

[22] Li Y-F, Kwok J T, Zhou Z-H. Cost-sensitive semi-supervised support vector machine. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligences* (*AAAI'10*), Atlanta, GA, 2010: 500-505.

[23] Chapelle O, Zien A. Semi-supervised learning by low density separation. In: *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics* (*AISTATS'05*). Barbados, 2005: 57-64.

[24] Kockelkorn A, Lüneburg A, Scheffer T. Using transduction and multi-view learning to answer emails. In: *Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery* (*PKDD'03*). Cavtat-Dubrovnik, Croatia, 2003: 266-277.

[25] Wang L, Chan K L, Zhang Z. Bootstrapping SVM active learning by incorporating unlabeled images for image retrieval. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'03*). Madison, WI, 2003: 629-634.

[26] Kasabov N, Pang S. Transductive support vector machines and applications in bioinformatics for promoter recognition. *Neural Information Processing Letters and Reviews*, 2004, 3(2): 31-38.

[27] Goutte C, Déjean H, Gaussier E, et al. Combining labelled and unlabeled data: a case study on Fisher kernels and transductive inference for biological entity recognition. In: *Proceedings of the 6th Conference on Natural Language Learning* (*COLING'02*). Stroudsburg, PA, 2002: 1-7.

[28] Chapelle O, Sindhwani V. Keerthi S S. Branch and bound for semi-supervised support vector machines. In: Schölkopf B, Platt J C, Hoffman T, eds. *Advances in Neural Information Processing Systems 19* (*NIPS'06*). Cambridge, MA: MIT Press, 2006: 217-224.

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

[30] Zhang K, Tsang I W, Kwok J T. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks*, 2009, 20(4): 583-596.

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

[32] Xu L, Schuurmans D. Unsupervised and semi-supervised multi-class support vector machines. In: *Proceedings of the 20th National Conference on Artificial Intelligence* (*AAAI' 05*). Pittsburgh, PA, 2005: 904-910.

[33] De Bie T, Cristianini N. Convex methods for transduction. In: Thrun S, Saul L K, Schölkopf B, eds. *Advances in Neural Information Processing Systems 16* (*NIPS'04*). Cambridge, MA: MIT Press, 2004: 73-80.

[34] Valizadegan H, Jin R. Generalized maximum margin clustering and unsupervised kernel learning. In: Schölkopf B, Platt J, Hoffman T, eds. *Advances in Neural Information Processing Systems 19* (*NIPS'07*). Cambridge, MA: MIT Press, 2007: 1417-1424.

[35] Fan R-E, Chen P-H, Lin C-J. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 2005, 6: 1889-1918.

[36] Joachims T. Training linear SVMs in linear time. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'06*). Philadelphia, PA, 2006: 217-226.

[37] Fan R-E, Chang K-W, Hsieh C-J, et al. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 2008, 9: 1871-1874.

[38] Tsang I W, Kwok J T, Cheung P M. Core vector machines: fast SVM training on very large data sets. *Journal of Machine Learning Research*, 2006, 6: 363-392.

[39] Kim S-J, Boyd S. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 2008, 19(3): 1344-1367.

[40] Kelly J E, The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 1960, 8(4): 703-712.

[41] Lanckriet G R, Cristianini N, Bartlett P, et al. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 2004, 5: 27-72.

[42] Xu Z, Jin R, Yang H, et al. Simple and efficient multiple kernel learning by group lasso. In: *Proceedings of the 27th International Conference on Machine Learning* (*ICML'10*). Haifa, Israel, 2010: 1175-1182.

[43] Schölkopf B, Smola A J. *Learning with Kernels: Support vector machines, regularization, optimization and beyond*. Cambridge, MA: MIT Press. 2002.

[44] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, 7: 2399-2434.

[45] Sindhwani V, Keerthi S S. Large scale semi-supervised linear SVMs. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'06*). Seattle, WA, 2006: 477-484.

[46] Zhang T, Oles F J. A probability analysis on the value of unlabeled data for classification problems. In: *Proceedings of the 17th International Conference on Machine Learning* (*ICML'00*), San Francisco, CA, 2000: 1191-1198.

[47] Cozman F G, Cohen I. Unlabeled data can degrade classification performance of generative classifiers. In: *Proceedings of the 15th International Conference of the Florida Artificial Intelligence Research Society* (*FLAIRS'02*), Pensacola, FL, 2002: 327-331.

[48] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Saul L K, Weiss Y, Bottou L, eds. *Advances in Neural Information Processing Systems 17* (*NIPS'05*). Cambridge, MA: MIT Press, 2005: 529-536.

[49] Chawla N V, Karakoulas G. Learning from labeled and unlabeled data: an empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 2005, 23:331-366.

[50] Li M, Zhou Z-H. SETRED: self-training with editing. In: *Proceedings of 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD'05*), Hanoi, Vietnam, *LNAI* 3518, 2005: 611-621.

[51] Wang J, Jebara T, Chang S F. Graph transduction via alternating minimization. In: *Proceedings of the 25th International Conference on Machine Learning* (*ICML'08*). Helsinki, Finland, 2008: 1144-1151.

[52] Zhou Z-H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Chapman & Hall, 2012.

[53] Domingos P. MetaCost: a general method for making classifiers cost-sensitive. In: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'99*). San Diego, CA, 1999: 155-164.

[54] Elkan C. The foundations of cost-sensitive learning. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (*IJCAI'01*). Seattle, WA, 2001: 973-978.

[55] Ting K M. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 2002, 14(3): 659-665.

[56] Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. In: *Proceedings of the 3rd IEEE International Conference on Data Mining* (*ICDM' 03*). Melbourne, FL, 2003: 435-442.

[57] Zhou Z-H, Liu X-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(1): 63-77.

[58] Turney P. Types of cost in inductive concept learning. In: *ICML Workshop Cost-Sensitive Learning*, 2000: 15-21.

[59] Kukar M, Kononenko I. Cost-sensitive learning with neural networks. In: *Proceedings of the 13th European Conference on Artificial Intelligence* (*ECAI'98*). Brighton, UK, 1998, 445-449.

[60] Fan W, Stolfo S J, Zhang J, Chan P K. AdaCost: Misclassification cost-sensitive boosting. In: *Proceedings of the 16th International Conference on Machine Learning* (*ICML' 99*). Bled, Slovenia, 1999: 97-104.

[61] Morik K, Brockhausen P, Joachims T. Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring. In: *Proceedings of the 16th International Conference on Machine Learning* (*ICML'99*). Bled, Slovenia, 1999: 268-277.

[62] Greiner R, Grove A J, Roth D. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 2002, 139(2): 137-174.

[63] Margineantu D D. Active cost-sensitive learning. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (*IJCAI'05*). Edinburgh, Scotland, 2005: 1622-1623.

[64] Liu A, Jun G, Ghosh J. Spatially cost-sensitive active learning. In: *Proceedings of the 9th SIAM International Conference on Data Mining* (*SDM'09*), Sparks, NV, 2009: 814-825.

[65] Seung H, Opper M, Sompolinsky H. Query by committee. In: *Proceedings of the 5th ACM Workshop on Computational Learning Theory*, Pittsburgh, PA, 1992: 287-294.

[66] Lewis D, Gale W. A sequential algorithm for training text classifiers. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'94*), Dublin, Ireland, 1994: 3-12.

[67] Abe N, Mamitsuka H. Query learning strategies using boosting and bagging. In: *Proceedings of the 15th International Conference on Machine Learning* (*ICML'98*), Madison, WI, 1998: 1-9.

[68] Sun Y-Y, Zhang Y, Zhou Z-H. Multi-label learning with weak label. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence* (*AAAI'10*). Atlanta, GA, 2010: 593-598.

[69] Raykar V C, Yu S, Zhao L H, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: *Proceedings of the 26th International Conference on Machine Learning* (*ICML 2009*), Montreal, Canada, 2009: 889-896.