新型光滑正则半监督SVM方法及其在信用评级中的应用

薛 飞[1] 鲁利民[2] 王 磊[1]

(西南财经大学经济信息工程学院 成都 610072)[1] (西南财经大学西部商学院 成都 610072)[2]

摘 要 提出了一种基于光滑正则的半监督支持向量机方法,并将其用于建立中小信用评级模型。它从少量标签样本和大量无标签样本中构造反映数据流形结构的光滑正则项,并结合到支持向量机的最大间隔分类器的学习过程。然后,提出一种渐进式方法来迭代获得"半标签"样本,稳健地提升支持向量机的泛化性能。在真实数据集上的实验结果表明,新方法获得的测试精度显著优于多种现有方法,非常适用于中小企业的信用评级任务。

关键词 支持向量机,信用评级,半监督学习,光滑正则

中图法分类号 TP391.4 **文献标识码** A

Novel Smooth Regularization Based Semi-supervised SVM Approach and its Application in Credit Evaluation

XUE Fei[1] LU Li-min[2] WANG Lei[1]

(School of Economics Information Engineering, Southwest University of Finance & Economics, Chengdu 610072, China)[1]

(School of Western Business, Southwest University of Finance & Economics, Chengdu 610072, China)[2]

Abstract This paper proposed a novel smooth regularization based semi-supervised support vector machine approach, and applied it to set up credit evaluation model of small-and-medium enterprises. It computes a manifold-related smooth regularization term on both few labeled samples and plenty of unlabeled samples, which is combined into the learning process of maximal margin classifiers. Then, it adopts a progressive method to acquire semi-labeled samples iteratively so that the generalization performance of support vector machine can be improved gradually. Experiments on reality dataset show that the testing accuracy of proposed approach outperforms several popular ones, and is very suitable for evaluating credit grades of small-and-medium enterprises.

Keywords Support vector machines, Credit evaluation, Semi-supervised learning, Smooth regularization

1 引言

从20世纪90年代起,一些学者开始将机器学习方法应用在信用评级问题中[1]。Desai等较早地将神经网络引入消费者贷款的信用风险分析[2]。李萌等将上市公司不良贷款率作为衡量信用风险高低的标准,基于BP神经网络和主成分分析法构造商业银行信用等级识别模型,取得了优良的识别精度[3]。

支持向量机(support vector machines, SVM)具有收敛到全局最优、非线性学习、维数不敏感等优点[4]。它通过在再生核Hilbert空间(RKHS)中计算的最大间隔超平面实现任务分类,是目前公认的最有效的机器学习方法之一。庞素琳等对我国2000年106家上市公司进行了信用评级,发现从分类准确度来看,SVM方法明显优于BP神经网络和LDA方法[5]。Gestel等利用最小均方SVM对银行的证券投资组合的风险进行评级[6]。Martensa等人利用SVM建立企业信用评级模型,并设计两种方法导出显式的评级规则,提高了评级模型的可解释性和适应性[7]。考虑到SVM对于噪声样本比较敏感,张永等提出了一种模糊补偿多类支持向量机算法,用于建设银行个人房贷的信用评估系统中,取得了较好的效果[8]。宋晓东等针对"样本重叠"问题,提出了一种双隶属模糊支持向量机算法对中小企业信用风险进行评级[9]。吴冲等提出一种新颖的基于模糊积分的支持向量机集成方法,该方法充分考虑了不同基分类器的重要性,并用于电子商务客户的信用评级,具有较好的鲁棒性和精度[10]。最近,Wang等人提出一种混合支持向量机集成方法,它的关键在于同时利用boostrap机制和随机子空间方法增强基分类器的多样性,提升集成效果[11]。

可见,目前绝大多数方法都属于监督式方法,它们假设事先已经存在相当数量的标签样本(已经获得确定信用等级的企业)用于建立信用评级模型。实际的情况是,中小企业由于数量庞大、财务和经营数据不透明、生命周期短等原因,很难获得大量准确的标签样本。此时,仅依靠少量的标签数据难以准确度量实际的数据分布规律,因而上述方法很难建立有效的信用评级模型。另一方面,获取大量的无标签样本却十分容易,但它们不能体现数据的分布规律而且隐藏着数据的判别信息。因此,半监督机器学习方法近几年受到广泛的关注[12]。Belkin等人指出[14],合理利用无标签样本的局部流形

到稿日期:2012-12-24 返修日期:2013-03-22 本文受教育部人文社会科学研究一般项目(10YJCZH153)资助。

薛 飞(1975-),女,讲师,主要研究方向为商务智能、信息技术及其应用等,E-mail:xuef_t@swufe.edu.cn;鲁利民(1968-),男,博士,副教授,主要研究方向为企业管理、市场营销等;王 磊(1978-),男,博士,副教授,主要研究方向为机器学习、模式识别等。

结构有助于显著提高机器学习的性能，从而提出一种基于流形正则的半监督拉普拉斯支持向量机(LapSVM)方法。近期，Shalit 等人利用矩阵低秩分解技术提出一种在线版本的流形正则的支持向量机方法，该方法能够大幅降低 LapSVM 学习的时空复杂度[15]。Zhao 等人利用流形假设和聚类假设给出 RKHS 空间的一种半监督核函数，有效结合无标签样本的分布提出一种渐进式半监督支持向量机方法[16]。此外，Li 等人通过层次聚类技术对无标签样本进行采样，明显提高了半监督机器学习的效率[17]。

在上述研究工作的基础上，本文提出一种新颖的渐进式基于光滑正则的半监督支持向量机学习方法，并将其应用在标签数据不足情况下的中小企业信用评级问题中。

2 基于光滑正则的半监督支持向量机

令 $S_l^0=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n \mid \forall \boldsymbol{x}_i\in\boldsymbol{R}^d, i=1,2,\cdots,n\}$ 是少量已标签的样本，它们所属的类别标记(信用等级)分别为 $y_i\in\{c_1,c_2,\cdots,c_k\}$。同时，存在大量的未标签样本 $S_u^0=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m \mid \forall \boldsymbol{x}_i\in\boldsymbol{R}^d, i=1,2,\cdots,m, n<<m\}$，其类别标签未知。目标是在半监督数据集 $S_l^0\cup S_u^0$ 上，建立有效的 SVM 分类模型用于中小企业的信用评级。简单起见，我们假设 $k=2$ 且 $y_i\in\{+1,-1\}$；对于 $k>2$ 的多分类问题，可采用 OAA，OAO，ECOC 等方法扩展解决[13]。

2.1 基于光滑正则的半监督加权支持向量机模型

考虑到当标签样本较少时，它们不足以反映数据的实际分布，并且由它们训练的 SVM 的决策函数在再生核 Hilbert 空间(RKHS)中容易产生过拟合。为了克服该问题，我们借鉴文献[14]的方法，假设决策函数在 RKHS 空间中具有“光滑”的特点，利用大量的无标签样本构建一种“光滑”正则项，如下：

$$\Omega(f)=\sum_{i,j=1}^{n+m}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2\cdot W_{i,j} \tag{1}$$

式中，$f(\boldsymbol{x})$ 是核函数 $k(\boldsymbol{x}',\boldsymbol{x})$ 诱导的 RKHS 空间 $\boldsymbol{H}_k$ 中的某个决策函数；$\boldsymbol{x}_i,\boldsymbol{x}_j\in S_l^0\cup S_u^0$ 是全部训练样本；$W_{i,j}$ 是在 $S_l^0\cup S_u^0$ 构造的邻接图上任意两点的连接权值，由以下热核函数计算：

$$W_{i,j}=\begin{cases}\exp\left(-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma}\right), & \text{if } \boldsymbol{x}_j\in KNN(\boldsymbol{x}_i)\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

设 $\boldsymbol{W}\in\boldsymbol{R}^{(n+m)\times(n+m)}$ 是 $W_{i,j}$ 的矩阵形式，$\boldsymbol{D}\in\boldsymbol{R}^{(n+m)\times(n+m)}$ 是元素为 $D_{ii}=\sum_{j=1}^{n+u}W_{i,j}$ 的对角矩阵，则邻接图的 Laplacian 矩阵为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$。这样，令 $\boldsymbol{f}=(f(\boldsymbol{x}_1),f(\boldsymbol{x}_2),\cdots,f(\boldsymbol{x}_{n+m}))^{\mathrm{T}}$，我们可以将式(1)简化为 $\Omega(f)=\boldsymbol{f}^{\mathrm{T}}\cdot\boldsymbol{L}\cdot\boldsymbol{f}$。

分析可知，Laplacian 矩阵从样本之间的局部近邻关系的角度描述了数据分布的流形结构。而正则项 $\Omega(f)$ 的目的是充分利用大量的无标签样本描述数据分布的流形结构，然后使得支持向量机在 RKHS 空间中的决策函数尽量保持光滑(也即，将对它在近邻样本上的较大决策差异给予“惩罚”)，通过这种方式克服它“过拟合”少量标签样本的问题。

因此，我们给出如下的基于光滑正则的半监督加权支持向量机模型。

$$\begin{aligned}&\min_{f\in\boldsymbol{H}_k,\xi_i\in\boldsymbol{R}}\frac{1}{2}\|f\|_k^2+\gamma_1\cdot\sum_{i=1}^{n}\delta_i\cdot\xi_i+\gamma_2\cdot\boldsymbol{f}^{\mathrm{T}}\cdot\boldsymbol{L}\cdot\boldsymbol{f}\\&\text{s.t. } y_i\cdot f(\boldsymbol{x}_i)\geqslant 1-\xi_i, i=1,2,\cdots,n\\&\xi_i\geqslant 0, i=1,2,\cdots,n\end{aligned} \tag{3}$$

这里，目标函数的前两项分别是 RKHS 空间中的分类模型复杂度、标签样本的经验损失。我们认为标签样本在模型训练过程中的贡献是不一样的(见 2.3 节的分析)，为每个标签样本赋予权值 δ_i。此外，γ_1,γ_2 是权衡系数。

借助于著名的 Representer 定理[4]，我们得到 RKHS 中决策函数的一般形式，即 $f(\boldsymbol{x})=\sum_{j=1}^{n+m}\beta_j\cdot K(\boldsymbol{x},\boldsymbol{x}_j)+b$，其中，$\beta_j$ 是系数，b 是偏置项。因此，可以对优化(3)的目标函数进行整理：

$$\begin{aligned}&\min_{\beta\in\boldsymbol{R}^{n+m},\xi_i\in\boldsymbol{R}}\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\cdot\boldsymbol{K}\cdot\boldsymbol{\beta}+\gamma_1\cdot\sum_{i=1}^{n}\delta_i\cdot\xi_i+\gamma_2\cdot\boldsymbol{\beta}^{\mathrm{T}}\cdot\boldsymbol{K}\cdot\boldsymbol{L}\\&\cdot\boldsymbol{K}\cdot\boldsymbol{\beta}\end{aligned} \tag{4}$$

设 $\alpha_i\geqslant 0,\mu_i\geqslant 0$ 为拉格朗日乘子，可以得到优化式(3)的拉格朗日对偶函数：

$$\begin{aligned}L(\boldsymbol{\beta},\xi_i,b,\alpha_i,\mu_i)=&\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\cdot\boldsymbol{K}\cdot\boldsymbol{\beta}+\gamma_1\cdot\sum_{i=1}^{n}\delta_i\cdot\xi_i+\gamma_2\cdot\boldsymbol{\beta}^{\mathrm{T}}\\&\cdot\boldsymbol{K}\cdot\boldsymbol{L}\cdot\boldsymbol{K}\cdot\boldsymbol{\beta}-\sum_{i=1}^{n}\alpha_i(y_i(\sum_{j=1}^{n+m}\beta_j\cdot K\\&(\boldsymbol{x}_i,\boldsymbol{x}_j)+b)-1+\xi_i)-\sum_{i=1}^{n}\mu_i\xi_i\end{aligned} \tag{5}$$

将式(5)分别对 $b,\xi,\boldsymbol{\beta}$ 求偏导数可得：

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

$$\frac{\partial L}{\partial\xi_i}=0\Rightarrow\gamma_1\delta_i-\alpha_i-\mu_i=0\Rightarrow 0\leqslant\alpha_i\leqslant\gamma_1\delta_i \tag{7}$$

$$\begin{aligned}&\frac{\partial L}{\partial\boldsymbol{\beta}}=0\Rightarrow(\boldsymbol{K}+2\gamma_2\boldsymbol{KLK})\boldsymbol{\beta}-\boldsymbol{KM}\boldsymbol{\alpha}=0\\&\Rightarrow\boldsymbol{\beta}=(\boldsymbol{I}+2\gamma_2\boldsymbol{LK})^{-1}\boldsymbol{M}\boldsymbol{\alpha}\end{aligned} \tag{8}$$

式中，$\boldsymbol{M}=\boldsymbol{G}\cdot\boldsymbol{Y}$，$\boldsymbol{G}=(\boldsymbol{I}\quad 0)$，$\boldsymbol{Y}=diag(y_1,y_2,\cdots,y_n)$，其中 $\boldsymbol{M}$ 和 $\boldsymbol{G}$ 是 $(n+m)\times n$ 维矩阵。令 $\boldsymbol{e}$ 是 n 维全 1 向量，将式(6)—式(8)合并到式(5)中，可以整理得到优化式(3)的 Wolf 对偶问题：

$$\begin{aligned}&\max_{\alpha\in\boldsymbol{R}^n}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{e}-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{I}+2\gamma_2\boldsymbol{LK})^{-1}\boldsymbol{M}\boldsymbol{\alpha}\\&\text{s.t. }\sum_{i=1}^{n}\alpha_i y_i=0, i=1,2,\cdots,n\\&0\leqslant\alpha_i\leqslant\gamma_1\delta_i, i=1,2,\cdots,n\end{aligned} \tag{9}$$

可以证明，优化问题式(9)仍然是线性约束的凸二次规划问题。它与标准 SVM 的区别是：1)目标函数中增加了由无标签数据产生的反映数据分布流形的光滑正则项(即 L)，因此是一种半监督学习模型；2)标签样本的权值视为不同(即 δ_i)。

求解出最优值 $\boldsymbol{\alpha}^*$ 后，利用式(8)容易推导 SVM 的决策函数：

$$f(\boldsymbol{x})=\boldsymbol{\alpha}^{*\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{I}+2\gamma_2\boldsymbol{LK})^{-1}\cdot\boldsymbol{k}(\cdot,\boldsymbol{x})+b \tag{10}$$

式中，$\boldsymbol{k}(\cdot,\boldsymbol{x})=(k(\boldsymbol{x}_1,\boldsymbol{x}),k(\boldsymbol{x}_2,\boldsymbol{x}),\cdots k(\boldsymbol{x}_{n+m},\boldsymbol{x}))^{\mathrm{T}}$ 是由核函数 $k(\boldsymbol{x}',\boldsymbol{x})$ 构造的 $n+m$ 维向量。偏置 b 可以由任意支持向量确定[4]。

2.2 “半标签”样本的获取过程

优化模型式(9)利用无标签样本的流形结构克服了支持向量机的决策超平面过度拟合标签样本问题，但并没有充分利用无标签样本中隐含的分类信息。注意到，在 RKHS 空间中距离 SVM 的决策超平面较远的点具有非常高的概率被正确分类，因此，我们可以迭代地对这部分无标签样本进行标注，从而逐渐扩大标签样本集的规模。

假设第 $t-1$ 个迭代步利用优化模型式(9)求解的 SVM

决策函数为 $f^{t-1}(\boldsymbol{x})$，它在任意样本 $\boldsymbol{x}_i \in S_u^{t-1}$ 上的决策值为 $f^{t-1}(\boldsymbol{x}_i)$，$\boldsymbol{x}_i$ 的标注类别 $\hat{y}_i = sgn(f^{t-1}(\boldsymbol{x}_i))$。则在第 t 个迭代步骤，可以根据 $f^{t-1}(\boldsymbol{x}_i)$ 的值将当前的无标签样本集合 S_u^{t-1} 划分成两个子集：

$$\begin{aligned} \boldsymbol{A}_+^t &= \{(\boldsymbol{x}_i, \hat{y}_i) \mid \boldsymbol{x}_i \in S_u^t, f^{t-1}(\boldsymbol{x}_i) \geqslant 0\} \\ \boldsymbol{A}_-^t &= \{(\boldsymbol{x}_i, \hat{y}_i) \mid \boldsymbol{x}_i \in S_u^t, f^{t-1}(\boldsymbol{x}_i) \leqslant 0\} \end{aligned} \tag{11}$$

我们分别从两个集合中选取 $|f^{t-1}(\boldsymbol{x}_i)|$ 最大的前 p 个样本作为第 t 迭代步的“半标签”集合，即 S_{b+}^t 和 S_{b-}^t。

此外，为了给“半标签”样本重新修正标签的机会，我们采用一种校正策略：采用 $f^{t-1}(\boldsymbol{x})$ 对此前所有的“半标签”样本进行重新决策，如果它们标注的标签与以前不一致，则应将它们重新置入无标签集合。我们记 $t-1$ 迭代步的“半标签”集合为 S_B^{t-1}，并记 t 迭代步在 S_B^{t-1} 中识别出的不一致样本集合为 S_c^t，满足

$$S_c^t = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in S_B^{t-1}, sgn(f^{t-1}(\boldsymbol{x}_i)) \neq sgn(f^{t-2}(\boldsymbol{x}_i))\} \tag{12}$$

因此，第 t 迭代步时“半标签”集合和无标签集合分别调整为：

$$\begin{aligned} S_B^t &= S_B^{t-1} \cup (S_{b+}^t \cup S_{b-}^t) - S_c^t \\ S_u^t &= (S_u^{t-1} \cup S_c^t) - (S_{b+}^t \cup S_{b-}^t) \end{aligned} \tag{13}$$

2.3 权值调整策略

由于“半标签”样本只是从无标签数据集中选择出具有较大概率被正确标注的部分样本，因此不能保证它们获得的“标签”绝对正确。但是可以假设：对于任意“半标签”样本，如果在 t 迭代步时对它重新决策，并且它的标签不发生改变，则认为它被正确标注的概率进一步增大，因而在 t 迭代步对决策超平面有更大的影响。

基于该假设，我们采用样本权值的变化来反映“半标签”样本被正确标注的概率，即

$$\delta_i^t = \begin{cases} 1, & \boldsymbol{x}_i \in S_l^0 \\ \log_Z(p_i^t + 1), & \boldsymbol{x}_i \in S_B \end{cases} \tag{14}$$

式中，p_i^t 表示在第 t 迭代步时，样本 $\boldsymbol{x}_i$ 从进入“半标签”集合开始所经历的迭代步数，显然，对于 $\forall \boldsymbol{x}_i \in S_{b+}^t \cup S_{b-}^t$ 有 $p_i^t = 1$。而 Z 是表示最大迭代步数的常量。

分析可知，刚进入“半标签”集合的样本的权值较少，它们对于决策超平面的影响程度有限，但随着迭代进行，它们的权值逐渐增大直至 1。这种渐进式调整权值的方式和前文的假设一致，因而也是合理的，保证了 SVM 的决策超平面随着“半标签”样本的增加而稳步调整。

2.4 渐进式半监督学习算法

基于上述分析，我们设计了一种渐进式的基于光滑正则的半监督 SVM 学习算法(SRS-SVM)用于中小企业的信用评级。它们的主要思想是：迭代地从无标签样本中选择距离 RKHS 空间的决策超平面最远的部分样本进行标注，然后在产生的“半标签”集合和原始标签集合上重新训练支持向量机模型(“半标签”样本的权值迭代地调整)，从而以渐进的方式利用无标签样本蕴含的分类信息对 SVM 的决策超平面进行优化。

算法的主要步骤如下所示。

算法 1 SRS-SVM

输入：标签集合 S_l^0，无标签集合 S_u^0，模型参数 γ_1，γ_2，最大迭代次数 Z

输出：SVM 的最终决策函数 $f^*(\boldsymbol{x})$

Step1：利用梯度下降方法在 S_l^0 和 S_u^0 上求解凸二次优化式(9)，利用式(10)得到初始决策函数 $f^0(\boldsymbol{x})$，并令初始“半标签”集合 $S_B^0 = \varnothing$；

Step2：for t=1,2,…,Z

a)利用 $f^{t-1}(\boldsymbol{x})$ 对 S_u^{t-1} 中的样本进行标注，并从每一类别种选取 $|f^{t-1}(\boldsymbol{x}_i)|$ 最大的前 p 个样本作为当前的半标注集合，即 S_{b+}^t 和 S_{b-}^t。

b)对 S_B^{t-1} 中的“半标签”样本进行重新标注，按式(12)识别不一致集合 S_c^t。

c)按式(13)计算 S_u^t 和 S_B^t。

d)按式(14)确定集合 S_l^0 和 S_B^t 中样本的权值。

e)以 $S_l^0 \cup S_B^t$ 为标签样本集合，联合 S_u^t 求解优化式(9)，并由式(10)得到 SVM 的决策函数 $f^t(\boldsymbol{x})$。

end for

3 实验仿真及结果分析

本节将提出的基于光滑正则的半监督支持向量机方法(SRS-SVM)应用到中小企业的信用评级问题中，以检验它在标签样本不足条件下(半监督)的泛化能力，并与几种经典方法进行比较，包括传统支持向量机(SVM)[2]、双隶属模糊支持向量机(DFSVM)[9]、半监督拉普拉斯支持向量机(LapSVM)[14]。

3.1 评级指标选取

将支持向量机算法应用在企业信用分类问题时，关键步骤之一在于评级指标的选择。总结国外已有研究成果，本文选取能综合反映中小企业的盈利能力、偿债能力、经营能力、发展能力和现金流量能力 5 个方面的 20 个指标作为评价中小企业信用状况的指标，如表 1 所列。

表 1 评价中小企业信用状况的指标体系

类型	指标名称	
盈利能力	①X_1：总资产净利润率	②X_2：成本费用利润率
	③X_3：销售净利率	④X_4：每股收益
偿债能力	①X_5：流动比率	②X_6：速动比率
	③X_7：现金比率	④X_8：资产负债率
	⑤X_9：所有者权益比率	
营运能力	①X_{10}：存货周转率	②X_{11}：应收账款周转率
	③X_{12}：总资产周转率	④X_{13}：营业成本率
发展能力	①X_{14}：净利润增长率	②X_{15}：营业收入增长率
	③X_{16}：总资产增长率	
现金流量能力	①X_{17}：营业收入现金比率	②X_{18}：现金流量比率
	③X_{19}：每股经营活动现金净流量	
	④X_{20}：全部资产现金回收率	

3.2 样本的选取和预处理

实验数据来自于四川省某大型商业银行最近 5 年的中小企业贷款客户，总计 1750 个样本。经过相关专家经验分析，将这些中小企业评定为“信用良好”和“信用不良”两大类(分别包含 1109 和 641 个样本)。我们采用随机划分的方法将样本分成训练集合和测试集合两部分(分别包含 1000 个和 750 个样本)，重复划分 10 次，产生 10 组训练和测试集合。对于每组数据的训练集，随机选取 n 个(n=50,100,200,300,500)作为标签训练样本集 S_l^0，剩余部分则去掉标签成为无标签训练样本集 S_u^0。

此外，我们对样本的所有指标进行归一化处理，取值区间为[0,1]。

3.3 不同算法的性能比较实验

分别采用 SRS-SVM，SVM，DFSVM 和 LapSVM 算法在

训练集合上学习支持向量机分类模型，核函数均采用高斯核函数，即 $k(x',x)=\exp(-\|x-x'\|^2/2\sigma^2)$。上述算法涉及的模型参数包括：经验损失项的权衡因子 γ_1，光滑正则项的权衡因子 γ_2，核函数的带宽参数 σ，本文采用粒子群优化方法(PSO)在独立的验证集数据上进行选取，结果如表 2 所列。此外，SRS-SVM 算法的最大迭代次数设置为 100 次，“半标签”集合的规模取为 $p=20$。

表 2　算法的模型参数的设置

算法	模型参数
SRS-SVM	$\gamma_1=10.28,\gamma_2=0.77,\sigma=0.08$
SVM	$\gamma_1=17.02,\sigma=0.11$
DFSVM	$\gamma_1=32.41,\sigma=0.18$
LapSVM	$\gamma_1=17.70,\gamma_2=0.24,\sigma=0.10$

为了便于比较，我们将 SRS-SVM 和 LapSVM 算法在标签数据集和无标签数据集上进行半监督学习，SVM 和 DFSVM 仅在标签数据集上监督学习，然后计算 $n=50,100,200,300,500$ 时不同算法在测试集上的测试误差。10 次重复实验的平均结果如表 3 所列。

表 3　不同算法的平均测试误差

误差率(×100%)	SRS-SVM	SVM	DFSVM	LapSVM
n=50	**9.54**	19.47	15.68	12.31
n=100	**7.61**	10.36	9.73	9.05
n=200	**6.98**	8.02	7.86	7.70
n=300	**6.80**	7.26	7.19	7.02
n=500	**6.65**	6.94	6.71	6.72

从表 3 的结果可知，本文提出的基于光滑正则的半监督 SVM 算法用于企业信用评级时取得了最优异的性能。与监督式的传统 SVM 算法相比，测试误差大幅降低(尤其是当 $n=50$ 时，误差率降低了 9.93 个百分点)。该现象表明，本文算法采用的光滑正则项度量了无标签数据的流形分布结构，并利用它们隐含的判别信息大幅提高了 SVM 模型的泛化性能；与另外一种半监督算法 LapSVM 相比，本文算法也取得了明显的优势(当 $n=50$ 时，误差率降低了 2.77 个百分点)。该现象表明，利用渐进式方法对无标签样本进行“半标注”，能有效扩充标签数据集的规模，从而更有效地学习 SVM 的判别模型。此外，DFSVM 算法利用模糊技术一定程度地提高了传统 SVM 算法的性能，但提升幅度有限。

此外，图 1 描述了不同情况下随着迭代次数的增加，SRS-SVM 算法的测试误差的变化曲线。显然，测试误差能随着迭代次数逐渐降低并趋于稳定。这表明“无标签”样本采取的渐进式半标注的方法能稳健地提升半监督支持向量机的泛化性能。

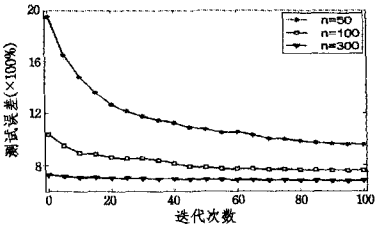

图 1　测试误差随着迭代次数的变化曲线

综合上述实验分析可知，在标签样本不足的情况下，本文提出的基于光滑正则的半监督支持向量机方法能够稳健、有效地建立模型实现中小企业信用评级，其泛化性能与多种现有算法相比具有明显的优势。

结束语　本文针对中小企业信用评级时标签数据不足的问题，提出一种基于光滑正则的半监督支持向量机方法。它利用无标签样本的局部流形结构定义了一种光滑正则项，并将其用于支持向量机的最大间隔分类器的学习。同时设计一种渐进式方法迭代地获得“半标签”样本，稳健地提高了支持向量机的泛化性能。初步实验结果验证了本文提出的算法的优势，其非常适用于中小企业的信用评级应用。在今后的工作中，我们将进一步研究“半标签”样本的加权机制，以便有效地发掘它们隐含的判别信息，用于 SVM 分类器的学习。

参 考 文 献

[1] 徐志春. 我国商业银行中小企业信用风险管理研究[D]. 武汉：华中科技大学，2012

[2] Desai V S, Crook J N, Jr G A. A comparison of neural networks and linear scoring models in the credit union environment[J]. European Journal of Operational Research, 1996, 95(1): 24-39

[3] 李萌，陈柳钦. 基于 BP 神经网络的商业银行信用风险识别实证分析[J]. 南京社会科学，2007(1)：18-29

[4] 张学工. 统计学习理论的本质[M]. 北京：清华大学出版社，2000

[5] 庞素琳. 信用评价与股市预测模型研究及应用[M]. 北京：科学出版社，2005

[6] Gestel V, Baesens B, Garcia J, et al. A support vector machine approach to credit scoring[J]. Bank en Financiewezen, 2003(2): 73-82

[7] Martensa D, Baesensb B, Gestelc V, et al. Comprehensible credit scoring models using rule extraction from support vector machines[J]. European Journal of Operational Research, 2007, 183(3): 1466-1476

[8] 张永，迟忠先，闫德勤. 一种新的模糊补偿多类支持向量机[J]. 计算机科学，2006，33(12)：152-155

[9] 宋晓东，韩立岩. 基于双隶属模糊支持向量机的中小企业信用评价[J]. 工业工程，2012，15(1)：93-99

[10] 吴冲，夏晗. 基于支持向量机集成的电子商务环境下客户信用评估模型研究[J]. 中国管理科学，2008，16(10)：362-367

[11] Wang G, Ma J. A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine[J]. Expert Systems with Applications, 2012, 39(5): 5325-5331

[12] Zhu X J. Semi-supervised learning literature survey [R]. TR-1530. Madison: University of Wisconsin-Madison, 2008

[13] Hs C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transaction on Neural Networks, 2002, 13(2): 415-425

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

[15] Shalit U, Weinshall D, Chechik G. Online learning in the embedded manifold of low-rank matrices[J]. Journal of Machine Learning Research, 2012, 13: 429-458

[16] Zhao Z K, Qian J S, Cheng J. Semi-supervised kernel based progressive SVM for coal mine gas safety data classification[J]. Journal of Information&Computational Science, 2012, 9(7): 1771-1780

[17] Li Y F, Zhou Z H. Improving semi-supervised support vector machines through unlabeled instances selection[C]// Proceedings of AAAI'11. 2011: 386-391

[18] 郭春花. 基于邻域粗糙集和距离判别的信用风险评级[J]. 重庆理工大学学报：自然科学版，2013，27(2)：130-134