DOI：10.3724/SP.J.1219.2012.00007

基于两阶段学习的半监督支持向量机分类算法

陶新民，曹盼东，宋少宇，付丹丹

（哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001）

摘 要：提出了一种基于两阶段学习的半监督支持向量机（semi-supervised SVM）分类算法．首先使用基于图的标签传递算法给未标识样本赋予初始伪标识，并利用 k 近邻图将可能的噪声样本点识别出来并剔除；然后将去噪处理后的样本集视为已标识样本集输入到支持向量机（SVM）中，使得 SVM 在训练时能兼顾整个样本集的信息，从而提高 SVM 的分类准确率．实验结果证明，同其它半监督学习算法相比较，本文算法在标识的训练样本较少的情况下，分类性能有所提高且具有较高的可靠性．

关键词：SVM（support vector machine）；半监督；两阶段学习；伪标识

中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1002-0411(2012)-01-0007-07

The Semi-Supervised SVM Classification Algorithm Based on Two-Stage Learning

TAO Xinmin，CAO Pandong，SONG Shaoyu，FU Dandan

(*College of Information and Communication Engineering*, *Harbin Engineering University*, *Harbin* 150001, *China*)

Abstract: A semi-supervised support vector machine (semi-supervised SVM) classification algorithm is proposed based on two-stage learning. A graph-based label propagation algorithm is used to provide initial pseudo labels for the unlabeled samples. And k-nearest graph is applied to distinguishing and removing the possible noisy samples. Then the denoised samples are inputted into the support vector machine (SVM) as labeled samples, so that the global information of the whole samples can be utilized by SVM when it is used in the training to improve the classification accuracy. The experiment results show that compared with other semi-supervised learning algorithms, the proposed method improves classification performance and is of higher robustness in the case of fewer labeled training samples.

Keywords: SVM (support vector machine); semi-supervised; two-stage learning; pseudo label

1 引言（Introduction）

支持向量机（support vector machine）是由 Vapnik 首先提出来的，广泛应用于机器学习领域．其主要思想是建立一个最优决策超平面并使该平面与距离其最近的 2 类样本（即支持向量）之间的距离最大化 [1]，从而避免了以往神经网络学习过程中出现的过拟合、易陷入局部极值和维数灾难等诸多问题．传统的 SVM 方法作为一种有监督分类算法，需要一定数量的属于不同类别的标识样本进行训练才能获得较好的泛化能力．因此，要使用 SVM，首先要对样本集进行标识．然而在实际工作中，对于较易获得的海量样本，由于标识样本的代价较大，因此只有少数样本是被标识的，大多数是未标识的．

由于半监督学习方法能够将已标识和未标识样本提供的聚类信息有机结合起来，与传统分类算法相比更有助于解决实际问题，因此正逐渐成为当前机器学习领域的研究热点．目前，半监督支持向量机算法主要有：Joachims 提出的直推式支持向量机（TSVM）[2]，通过将未标识样本信息作为约束条件引入 SVM 中，引导分类超平面通过低密度区，减小了错分的概率，但同时带来了非凸优化的难题；Ting 等人提出的 BoostSVM（boosting support vector machine）[3] 算法将支持向量机和 AdaBoost 算法相结合，用于提高支持向量机的预测精度，但其受噪声影响很大；Belkin 等人提出的 LapSVM[4] 算法考虑了样本集的内部结构，但半正定优化问题导致其易陷入局部解．如何更好地利用未标识样本聚类信息且不受噪声影响，提高 SVM 分类性能，是值得进一步研究的问题．

基金项目：国家自然科学基金资助项目（61074076）；中国博士后科学基金资助项目（20090450119）；中国博士点新教师基金资助项目（20092304120017）．

通讯作者：陶新民，taoxinmin@hrbeu.edu.cn　收稿/录用/修回：2010-12-29/2011-05-12/2011-07-12

本文利用两阶段学习模型对整体样本集进行学习：第 1 阶段，在充分考虑全局结构信息的前提下，利用基于图的半监督模型给未标识样本赋予伪标签；第 2 阶段，将标识样本和伪标签样本作为整个训练样本集，运用 SVM 算法进行训练学习，使得 SVM 算法在训练时能充分利用未标识样本带来的结构信息，提高分类器的分类精度．考虑到伪标签生成后噪声样本的影响，在 SVM 训练前根据 k 近邻图，通过对比标签值识别并删除噪声样本，针对剩下的每个样本根据其所属类别的概率设置不同的惩罚因子，来增强 SVM 算法的鲁棒性和抗干扰能力．将本文提出的基于两阶段学习的半监督 SVM 算法同其它半监督 SVM 算法进行了对比，结果表明本文算法在只有少量标识样本的情况下分类精度有较大幅度的提高．

2 支持向量机分类算法（Support vector machine classification algorithm）

2.1 传统的支持向量机算法

支持向量机算法是建立在统计学习理论中的结构风险最小化原理基础上的，根据有限的样本信息，在模型复杂度和学习能力之间寻求最佳匹配，以获得最好的泛化能力．它通过核函数将原始特征空间中的非线性分类界面映射到更高维的特征空间中，使分类界面在高维特征空间中变得线性可分，分类效果更好．

例如，对 n 个样本进行 2 分类，需要建立最优分类面模型，此模型的构建可以表示成如下的约束优化问题：假设 $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$ 为给定的训练样本和其期望输出，寻找最优权值向量 $\boldsymbol{w}$ 和阈值 b，使下面的代价函数最小化 [5]：

$$\min \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{n}\varepsilon_i \tag{1}$$

约束条件：

$$\begin{aligned} &y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1-\varepsilon_i \\ &\varepsilon_i\geqslant 0,\ \ i=1,2,\cdots,n \end{aligned} \tag{2}$$

式中，$C>0$ 是惩罚因子，表示对错分样本的惩罚程度；ε_i 为松弛变量，表示对训练样本的错分程度．利用拉格朗日乘子法求解该问题的最优解，确定最优分类．

2.2 较少标识样本对 SVM 算法分类性能的影响

为了测试较少标识样本对 SVM 分类器性能的影响，选用人工合成的月牙形样本集．月牙形样本集由 2 类共 200 个样本组成：第 1 类 104 个样本，第 2 类 96 个样本．SVM 算法的参数设置为：高斯核函数，核宽度 0.35；惩罚因子 $C=1000$；已标识的训练样本数为 10 个，类别比例定为 6∶4，分别用实心棱形和实心圆表示；未标识样本为 190 个，用虚黑框表示，训练后得到的 SVM 最优分类界面如图 1 所示，横坐标和纵坐标表示样本点的坐标值．

从图 1 可以看出，由于训练样本中标识样本太少，无法准确代表整个样本集合的样本分布信息，只依靠这些标识样本训练得到的 SVM 分类器泛化性能并不理想，因此传统 SVM 分类算法在标识样本较少时得到的分类器泛化性能较差，但是这种缺乏样本标识的情况在现实中是普遍存在的．如何在标识样本较少情况下提高 SVM 分类器泛化性能是值得深入研究的．

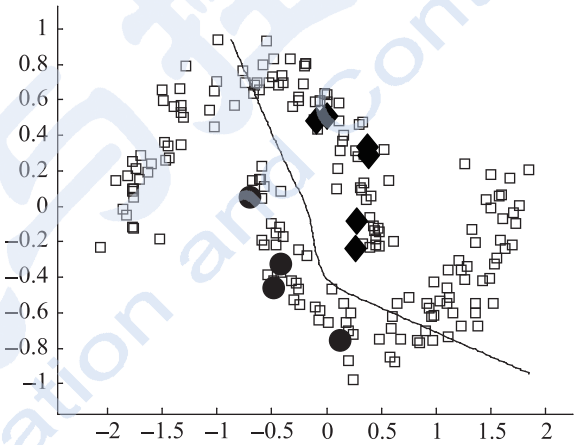

图 1 标识样本数量为 10（6:4）时的分类界面

Fig.1 The classification interface with 10 (6:4) labeled samples

3 基于两阶段学习的半监督 SVM 分类算法（Semi-supervised SVM classification algorithm based on two-stage learning）

SVM 是一个有监督的学习算法，没有考虑未标识样本的聚类信息，分类效果较差．以往的半监督 SVM 是通过改造 SVM 算法的优化函数来加入未标识样本信息，但是此过程会面临非凸优化或半正定优化难题．而基于图的半监督算法只能对未标识样本进行分类，对新来的样本（out-of-sample 样本）点无法进行处理，使训练出来的分类器缺乏实际应用能力．为了充分发挥两者优势，本文提出基于两阶段学习的半监督 SVM 分类算法，充分利用整个样本聚类结构的信息，得出准确的分类界面．

3.1 基于图的半监督伪标识生成算法

由于半监督学习算法只需少量标识样本就能得到较高的分类精度，因此在机器学习领域受到了无数科学工作者的青睐．根据算法的学习方式 [6]，目

前的半监督算法可分为：基于图的半监督算法 [7]；生成式模型算法，如用高斯模型模拟样本集的信息 [8]，其缺点是生成式模型的参数难以确定；基于主动学习和半监督结合的算法 [9]，利用一定的准则抽样出部分未标识样本进行标识，其缺点是性能往往取决于未标识样本的抽取顺序. 基于图的学习算法实现简单、易于理解，本文选择该算法作为半监督学习算法. 基于图的方法需在图中估计一个实值标识函数，并且规定此函数必须满足 2 个条件 [10]：在已标识样本中，用它估计出的标识必须非常接近真实值；它在整个样本集上必须光滑，即有连续的 1 阶和 2 阶偏导数 [11]. 伪标识生成算法主要分为 2 步：构建图模型，根据某种距离度量计算相似度矩阵，并将其转换成对称矩阵；依据概率转移矩阵生成伪标识值 [12].

(1) 确定相似度矩阵

定义由样本和标识组成的集合为 $S=\{X,Y|(\boldsymbol{x}_i,y_i)\}$，其中前 l 个已标识样本构成训练样本集合 L；后 u 个未标识样本构成测试样本集合 U，输入样本 $\boldsymbol{x}_i\in X$ 为 d 维向量，$y_i\in Y$ 为样本 $\boldsymbol{x}_i$ 的标签值. 利用基于欧氏距离的高斯函数构建相似度矩阵 $\boldsymbol{W}$，鉴于相似度矩阵由已标识和未标识部分构成，为方便矩阵运算，将矩阵 $\boldsymbol{W}$ 分割成 4 个子矩阵，见式 (3)，其中，$\boldsymbol{W}_{ll}$ 为已标识样本间的相似度，$\boldsymbol{W}_{lu}$ 为已标识样本和未标识样本间的相似度，$\boldsymbol{W}_{ul}$ 为未标识样本和已标识样本间的相似度，$\boldsymbol{W}_{uu}$ 为未标识样本间的相似度.

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{W}_{ll} & \boldsymbol{W}_{lu}\\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu}\end{pmatrix} \tag{3}$$

(2) 伪标签生成

定义标签转移概率 T_{ij}，表示标识从 j 变为 i 的概率，其具体表示形式见式 (4)：

$$T_{ij}=P(j\to i)=\frac{w_{ij}}{\sum_{k=1}^{l+u}w_{kj}} \tag{4}$$

其中 w_{ij} 表示相似度矩阵 $\boldsymbol{W}$ 中第 i 个样本和第 j 个样本间的相似度，$\sum_{k=1}^{l+u}w_{kj}$ 表示第 j 个样本和所有样本间的相似度的总和.

定义标签矩阵 $\boldsymbol{Y}\in\mathbb{R}^{(l+u)\cdot N}$，其中 N 代表总类别数，$\boldsymbol{Y}$ 的每行代表样本点 $\boldsymbol{x}_i$ 属于各类的概率，未标识样本属于每一类的初始概率定为 $1/N$，表示其所属类别待确定.

伪标签生成基本步骤为：传递标签 $\boldsymbol{Y}\leftarrow\boldsymbol{YT}$；行单位化 $\boldsymbol{Y}$；固定标签样本，重复执行传递，直到 $\boldsymbol{Y}$ 收敛. 将矩阵分割为

$$\boldsymbol{T}=\begin{pmatrix}\boldsymbol{T}_{ll} & \boldsymbol{T}_{lu}\\ \boldsymbol{T}_{ul} & \boldsymbol{T}_{uu}\end{pmatrix},\ \boldsymbol{Y}=\begin{pmatrix}\boldsymbol{y}_l\\ \boldsymbol{y}_u\end{pmatrix}$$

可导出未标识样本的伪标签矩阵为

$$\boldsymbol{y}_u=(\boldsymbol{I}-\boldsymbol{T}_{uu})^{-1}\boldsymbol{T}_{ul}\boldsymbol{y}_l \tag{5}$$

式中 $\boldsymbol{I}$ 为单位阵. 则伪标签样本点的预分类为

$$c_{iu}=\arg\max_j y_{u_{ij}} \tag{6}$$

式中 c_{iu} 为伪标签样本集中第 i 个样本所属的类别，$y_{u_{ij}}$ 为伪标签样本中第 i 个样本属于第 j 类的概率，其中 $i=l+1,l+2,\cdots,l+u$，$j=1,2,\cdots,N$.

利用本文的学习算法给月牙形样本集赋予伪标识后得出的分类结果如图 2 所示. 从图 2 可以看出，伪标识生成过程可以理解为，已标识样本的标识以概率的形式游走到周围的近邻点，将自身标识赋给其近邻点.

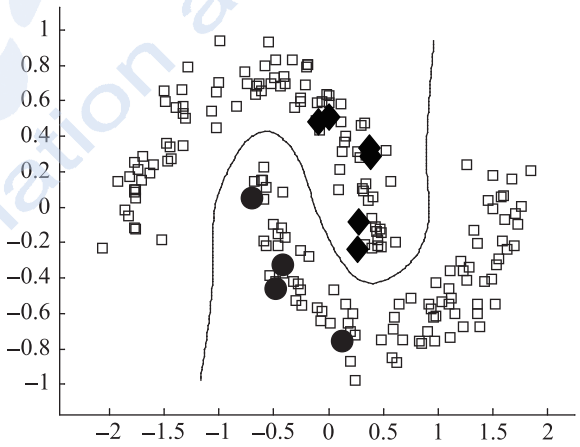

图 2　设置伪标识后的分类界面

Fig.2　The classification interface after pseudo labels are set

3.2　半监督 SVM 算法中核宽度的确定

在以往的伪标签生成过程中，相似度矩阵 $\boldsymbol{W}$ 中的核宽度 δ 都是通过反复实验得出的，增加了算法的计算时间和复杂度. 为此，本文采用自适应调节的方法，具体形式为

$$w_{ij}=\exp(-d^2(x_i,x_j)/\delta_i\delta_j) \tag{7}$$

其中，$d(x_i,x_j)=\|x_i-x_j\|$，核宽度 $\delta_i=d(x_i,x_{il})$，x_{il} 是 x_i 的第 l 个邻居. δ_i 随着近邻分布而自适应变化，保证了样本集内同类间的相似度不断增大，不同类别间相似度不断减小.

3.3 对伪标识样本进行去噪处理

生成的伪标签样本中可能存在边界噪声样本，导致 SVM 训练器得到的信息有偏差，使最终训练出的分类器出现错误，因此需要对伪标签样本进行去噪处理. 本文采取 2 种方法减少噪声的影响：(1) 通过对比标签值，识别并删除噪声样本；(2) 根据样本所属类别的信任度设置不同的惩罚因子.

将相似度矩阵 $\boldsymbol{W}$ 转换成 k 近邻图相似矩阵 $\boldsymbol{W}^*$，以表征样本间的聚类结构信息. 根据 k 近邻图稀疏矩阵，对比样本点和其近邻点的标签. 如果某样本点的标签值和其近邻点的伪标签值不相同，需判断该样本是否为噪声样本. 令

$$\boldsymbol{y}_u^{\text{new}} = \boldsymbol{W}^* \boldsymbol{y}_u \tag{8}$$

其中 $\boldsymbol{y}_u$ 是未标识样本点的伪标签矩阵，$\boldsymbol{W}^*$ 是 k 近邻图相似矩阵. 伪标签样本点的预分类变为

$$c'_{iu} = \arg\max_j y_{u_{ij}}^{\text{new}} \tag{9}$$

如果 $y_{u_{ij}}^{\text{new}}$ 满足式 (10) 或式 (11)，就将此样本作为噪声样本从训练样本集中删除.

$$c_{iu} \neq c'_{iu} \tag{10}$$

$$\frac{\max(y_{ui}^{\text{new}})}{\sum_i y_{ui}^{\text{new}}} < 0.8 \tag{11}$$

噪声处理后，$\boldsymbol{y}_u^{\text{new}}$ 变为 $\boldsymbol{y}_u^{\text{new}'}$，$\boldsymbol{y}^{\text{new}} = \begin{pmatrix} \boldsymbol{y}_l \\ \boldsymbol{y}_u^{\text{new}'} \end{pmatrix}$. 每个样本新的惩罚因子定义为

$$C_{i\text{new}} = C \cdot \max(y_{ic}^{\text{new}}) / \sum_i y_{ui}^{\text{new}} \tag{12}$$

式中，y_{ic}^{new} 代表点 x_i 属于每一类的概率，$i = 1,2,\cdots,l+u$，$c = 1,2,\cdots,N$，$C = 1000$.

3.4 基于两阶段学习的半监督 SVM 算法流程

(1) 计算样本集的相似度矩阵 $\boldsymbol{W}$ 并变换成 k 近邻图矩阵 $\boldsymbol{W}^*$.

(2) 确定标签转移概率矩阵 $\boldsymbol{T}$，计算未标识样本的伪标签矩阵 $\boldsymbol{y}_u$.

(3) 去除噪声样本，得到新的样本标签矩阵 $\boldsymbol{y}_u^{\text{new}}$.

(4) 根据式 (12) 计算每个样本新的惩罚因子 $C_{i\text{new}}$. 将处理后的样本集 $(\boldsymbol{x}_i, y_{ui}^{\text{new}})$ 及参数 $C_{i\text{new}}$ 代入到 SVM 的优化函数 (1) 中，$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 选用高斯核函数，核参数的选择采用平均 7 近邻距离法.

(5) 通过优化函数得出分类超平面.

4 实验（Experiment）

4.1 人工和 UCI 数据集实验

为了验证本文算法（TSLSVM）的优越性，选用 8 组不同的样本集进行实验，并与 SVM、BoostSVM、TSVM、LapSVM 四种算法的分类正确率进行比较. 8 组样本中前 5 组取自人工样本集，后 3 组取自国际机器学习标准样本库 UCI 中的 Pendigits、WDBC 和 IRIS. 为了测试两分类问题，在 IRIS 样本集中，选择 Iris Setosa 和 Iris Versicolour 为一类，Iris Virginica 为一类，Pendigits 中选用数字 0 和数字 5 的样本作为待分样本集. 样本特征信息见表 1.

表 1 实验样本集描述

Tab.1 The description of the test data sets

样本集	属性	样本总数	类别数
two-spirals	2	299	2
two-moons	2	200	2
two-circles	2	500	2
two-smiles	2	266	2
two-Gauss	2	400	3
Pendigits	16	7494	10
WDBC	30	569	2
IRIS	4	150	3

训练样本中标识样本数 L 分别为 5、10、15，实验迭代次数为 20 次，取 20 次的平均值作为最终的结果. SVM 中核函数为高斯函数，采用 10 次交叉验证法：惩罚因子 C 的搜索范围为 $\{2^1,2^2,2^3,2^4,2^5,2^6,2^7,2^8,2^9,2^{10}\}$，核宽度 δ 的搜索范围为 $\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0\}$. 将 C 和 δ 随机分成 10 组分别测试，统计出参数的范围. 最终得出最优参数为：$C = 1000$，$\delta = 0.35$. LPSVM 中 k 近邻图的近邻数选为 7；BoostSVM 中弱样本学习迭代次数为 20 次；LapSVM 中再生核希尔伯特空间（RKHS）规划因子 $\gamma_A = 10^{-5}$，流行规划因子 $\gamma_I = 1$，近邻数为 7；TSVM 中拉格朗日因子 $\lambda = 10^{-7}$，惩罚函数 $C = 1/2\lambda$，核函数为径向基函数，核宽度为 0.35. 实验结果如表 2 ～ 4 所示.

从实验结果可以看出：对于文中选用的样本集，在少量标识样本情况下，SVM 算法直接利用标识样本进行边界最大化优化，没有充分考虑样本整体的聚类信息. TSVM 算法利用了未标识样本的信息，但是存在非凸优化难题. BoostSVM 算法对噪声样本敏感，因此训练出来的最终分类器性能不理想. 本文 TSLSVM 算法的分类性能相对较好，这是由于 TSLSVM 算法充分考虑了样本集合的整体聚类结构

表 2　$L=5$ 时的分类准确率比较

Tab.2　Accuracy comparison of classification when $L=5$

Dataset	TSLSVM	SVM	BoostSVM	TSVM	LapSVM
two-spirals	**1.000±0.0**	0.565 3±0.034 7	0.583 7±0.027 8	0.515 8±0.066 5	0.659 2±0.094 8
two-moons	**1.000±0.0**	0.635 7±0.082 8	0.685 7±0.072 5	0.750 7±0.115 4	**1.000±0.0**
two-circles	**1.000±0.0**	0.601 9±0.089 7	0.782 0± 0.065 7	0.720 3±0.088 0	0.927 1±0.089 3
two-smiles	**1.000±0.0**	0.736 3±0.166 4	0.857 6±0.193 0	0.844 6±0.108 8	0.975 6±0.048 9
two-Gauss	0.995 5±0.024 2	0.863 1±0.005 1	0.881 2±0.006 5	0.734 6±0.048 6	**0.996 6±0.018 1**
Pendigits	0.937 1±0.090 9	0.650 4±0.092 4	0.780 5±0.041 2	0.834 7±0.094 5	**0.947 3±0.079 4**
WDBC	**0.812 5±0.178 2**	0.692 4±0.147 2	0.751 8±0.113 4	0.692 8±0.162 1	0.789 1±0.120 4
IRIS	**0.793 6±0.115 7**	0.744 0±0.109 3	0.778 0±0.121 3	0.779 6±0.127 8	0.784 2±0.108 9

表 3　$L=10$ 时的分类准确率比较

Tab.3　Accuracy comparison of classification when $L=10$

Dataset	TSLSVM	SVM	BoostSVM	TSVM	LapSVM
two-spirals	**1.000±0.0**	0.609 3±0.047 7	0.657 8±0.032 1	0.602 5±0.063 7	0.748 4±0.082 0
two-moons	**1.000±0.0**	0.768 7±0.108 2	0.835 6±0.122 7	0.842 0±0.101 5	**1.000±0.0**
two-circles	**1.000±0.0**	0.638 9±0.093 3	0.891 0±0.045 3	0.814 5±0.095 9	0.985 4±0.042 3
two-smiles	**1.000±0.0**	0.925 7±0.082 7	0.955 7±0.076 1	0.898 3±0.088 5	0.999 3±0.002 9
two-Gauss	0.995 3±0.025 3	0.865 4±0.005 8	0.865 4±0.005 8	0.851 0±0.118 8	**0.996 6±0.018 3**
Pendigits	0.971 8±0.073 3	0.731 3±0.108 3	0.832 7±0.121 1	0.850 7±0.085 4	**0.982 0±0.052 0**
WDBC	**0.885 1±0.107 3**	0.869 4±0.046 9	0.872 8±0.031 3	0.881 5±0.056 5	0.871 6±0.096 2
IRIS	**0.906 2±0.073 0**	0.849 8±0.058 9	0.881 1±0.031 3	0.872 2±0.056 9	0.894 9±0.082 5

表 4　$L=15$ 时的分类准确率比较

Tab.4　Accuracy comparison of classification when $L=15$

Dataset	TSLSVM	SVM	BoostSVM	TSVM	LapSVM
two-spirals	**1.000±0.0**	0.681 4±0.042 5	0.715 1±0.032 1	0.666 7±0.053 4	0.818 2±0.062 5
two-moons	**1.000±0.0**	0.860 0±0.082 9	0.880 0±0.057 8	0.878 7±0.082 7	**1.000±0.0**
two-circles	**1.000±0.0**	0.738 4±0.087 2	0.789 6±0.072 3	0.877 4±0.067 4	0.994 8±0.026 2
two-smiles	**1.000±0.0**	0.956 5±0.054 0	0.966 5±0.031 2	0.877 7±0.079 8	0.997 8±0.012 1
two-Gauss	**0.995 5±0.024 2**	0.868 0±0.008 0	0.916 0±0.012 0	0.894 8±0.094 3	**0.995 5±0.023 9**
Pendigits	**0.999 3±0.000 0**	0.854 0±0.094 5	0.903 1±0.016 5	0.896 1±0.069 5	**0.999 3±0.000 0**
WDBC	**0.897 5±0.061 5**	0.888 2±0.039 3	0.8895±0.0176	0.8963±0.0374	0.8948±0.0750
IRIS	**0.9360±0.0509**	0.8978±0.0357	0.909 1±0.015 5	0.884 4±0.061 2	0.931 6±0.042 1

信息，使 SVM 训练出的分类器的分类性能大大提高．随着标识样本数目的增加，其它算法的分类精度也有明显的提高．因此，可以认为本文算法在少量标识样本的情况下分类性能较好，而这项指标对半监督分类算法是非常重要的．LapSVM 算法在 $L=5$ 和 $L=10$ 时对 two-Gauss 和 Pendigits 样本集的分类性能与本文算法旗鼓相当，甚至略优于本文算法，这是由于本文算法可能在去噪阶段剔除了过多的样本点，使少许样本信息丢失，导致最终提供给 SVM 的信息不够完整，分类精度受到一些影响．总的来说，与其它算法相比，在少量标识样本数的情况下，本文算法能利用自身优势，充分发掘整体样本集信息，最终算法的分类准确率较高，且抗干扰能力较强，验证了本文算法的可行性．

4.2　故障检测实验

为了验证本文算法的可用性，选用工程中的轴承工作样本，对本文算法和上述其它算法进行分类准确率对比．实验选取 20000 个正常样本，20000 个

外圈故障样本，20000 个内圈故障样本和 20000 个滚动体故障样本. 训练样本中包含 2000 个正常样本、500 个外圈故障样本、500 个内圈故障样本和 500 个滚动体故障样本，测试样本中包含 1500 个正常样本、500 个外圈故障样本、500 个内圈故障样本和 500 个滚动体故障样本. 各算法的实验参数设置同 4.1 节，实验结果如图 3 ～ 5 所示.

从图 3 ～ 5 可以看出，本文算法在标识样本不同的情况下，分类性能均优于其它算法. 当标识样本数为 250 和 300 时，本文算法的平均分类准确率达到了 80%以上，体现了本文算法的优势. 随着标识样本数继续增加，各算法的性能都有不同程度的提高.

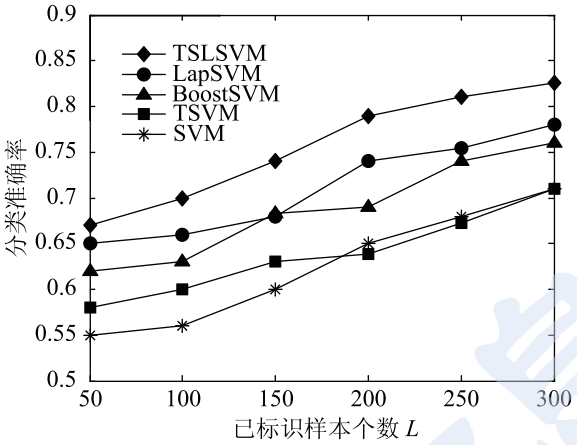

图 3 外圈故障样本分类结果

Fig.3 The classification results of outer ring fault samples

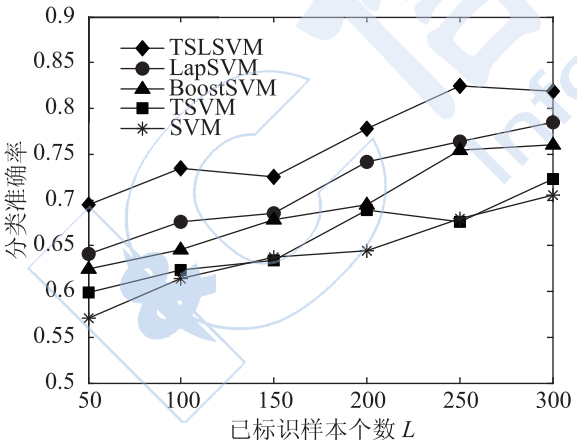

图 4 内圈故障样本分类结果

Fig.4 The classification results of inner ring fault samples

4.3 k 值对本文算法的影响

为了测试不同近邻数 k 值（1:15）对本文算法性能的影响，选择样本集 IRIS 和 Pendigits，并选择 10 个已标识样本，其它参数同上，实验结果如图 6 所示. 由实验结果可以看出，k 设置在区间 [6,11] 处分类精度较好且较稳定. k 选取太小则成为最近邻法，标识样本的信息向未标识样本区域辐射的范围将很小，本属于同类的样本不能被涵盖到一个类别中，致使最终分类精度较差；k 值过大则有可能将类别外的样本纳入其中，带来了类间的相互干扰，导致最终的分类精度呈下降趋势.

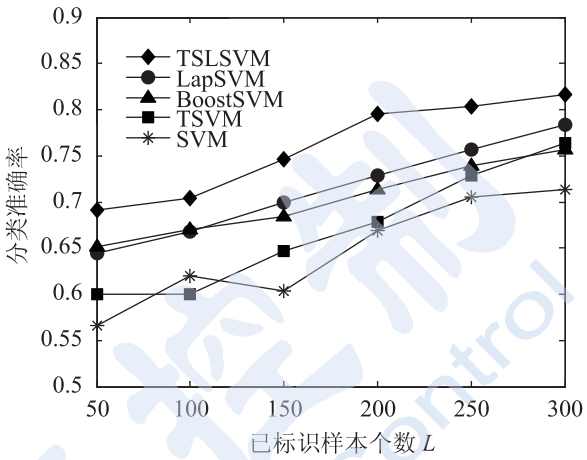

图 5 滚动体故障样本分类结果

Fig.5 The classification results of rollers fault samples

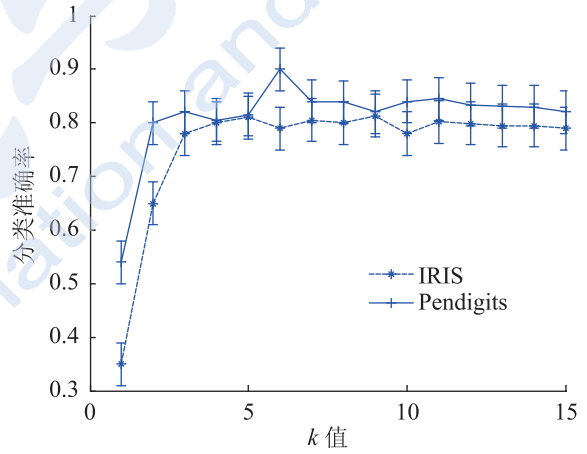

图 6 不同 k 值对本文算法的影响

Fig.6 The effects of different k values on our algorithm

5 结论（Conclusion）

针对传统 SVM 算法在标识样本数较少的情况下分类性能较差的问题，提出一种基于两阶段学习的半监督支持向量机算法. 该算法首先利用基于图的半监督算法给未标识样本赋予伪标签，接着利用 k 近邻图识别并删除其中的噪声样本，然后将处理后的样本集作为已标识样本集交由 SVM 处理，得出准确的分类界面. 实验部分使用了人工合成样本集、UCI 样本集和轴承故障样本集. 本文算法和其它半监督算法的分类准确率比较结果表明，本文算法在大部分数据集上的分类性能都优于其它算法. 需要说明的是，two-Gauss 和 Pendigits 数据集的实验结果表明本文算法与 LapSVM 算法的分类性能

接近，这是由于本文算法在处理噪声时，有可能错将有价值的边界样本当成噪声点删除而导致分类错误．如何解决二者间的矛盾，是下一阶段要深入研究的课题．

参考文献（References）

[1] 韩立群．人工神经网络教程 [M]．第 1 版．北京：北京邮电大学出版，2006：185-194．

Han L Q. The course of artificial neural networks[M]. 1st ed. Beijing：Beijing University of Posts and Telecommunications Press, 2006: 185-194.

[2] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th International Conference on Machine Learning. New York, USA: ACM, 1999: 200-209.

[3] Ting K M, Zhu L. Boosting support vector machines successfully[M]//Lecture Notes in Computer Science: vol.5519. Berlin, Germany: Springer-Verlag, 2009: 509-518.

[4] Belkin M, Niyogi P. Manifold regularization[J]. Machine Learning Research, 2006, 7(8): 31-42.

[5] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 7(3): 273-297.

[6] Zhu X J. Semi-supervised learning literatures survey[D]. Wisconsin, USA: University of Wiscinsin-Madison, 2005.

[7] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions[C]//Proceedings of the Twentieth International Conference on Machine Learning. Piscataway, NJ, USA: IEEE, 2003: 912-919.

[8] Monroy I, Benitez R, Escudero G, et al. A semi-supervised approach to fault diagnosis for chemical processes[J]. Computers and Chemical Engineering, 2010, 34(5): 631-642.

[9] Li C L, Wu C G. A new semi-supervised support vector machine learning algorithm based on active learning[C]//Proceedings of the 2010 2nd International Conference on Future Computer and Communication．Piscataway, NJ, USA: IEEE, 2010: 3638-3641.

[10] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[D]. Pennsylvania, USA: Carnegie Mellon University, 2002.

[11] Johnson R, Zhang T. On the effectiveness of Laplacian normalization for graph semi-supervised learning[J]. Machine Learning Research, 2007, 8(1): 1489-1517.

[12] Gong Y C, Chen C L. Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions[C]//Proceedings of the 19th International Conference on Pattern Recognition. Piscataway, NJ, USA: IEEE, 2008: 2217-2220.

作者简介：

陶新民（1973–），男，博士，副教授．研究领域为智能信号处理，软计算．

曹盼东（1986–），男，硕士生．研究领域为信号与智能信息处理．