迁移组概率学习机

倪彤光[1,2],王士同[1],应文豪[1],邓赵红[1]
(1.江南大学数字媒体学院,江苏无锡 214122;2.常州大学信息科学与工程学院,江苏常州 213164)

摘 要: 基于组概率的学习方法因其能够很好地保护数据的隐私性而成为近年来机器学习领域的研究热点.已有的组概率学习方法虽然取得了一定的效果,但是在模型训练时仅考虑单一的场景信息,如果在当前领域所采集的数据信息有限,则在当前领域下建立的分类模型泛化能力较差.针对此问题,提出了一种基于组概率和结构风险最小化模型的迁移组概率学习机(TGPLM).该方法通过构造领域相似距离项来引入历史领域的先验知识,提出了针对类标签保护数据的增强型分类器优化目标学习准则,以期在有效利用当前领域数据类标签组概率信息的同时借鉴历史领域相关知识来指导当前领域下的学习任务.基于模拟、UCI 及 PIE 人脸等数据集上的实验结果表明,本文所提之方法是有效的.

关键词: 迁移学习;分类;支持向量机;组概率
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2013) 11-2207-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.11.015

Transfer Group Probabilities Based Learning Machine

NI Tong-guang[1,2], WANG Shi-tong[1], YING Wen-hao[1], DENG Zhao-hong[1]
(1. *School of Digital Media, Jiangnan University, Jiangsu, Wuxi, Jiangsu* 214122, *China*;
2. *School of Information Science & Engineering, Changzhou University, Changzhou, Jiangsu* 213164, *China*)

Abstract: Learning from group probabilities helps to protect the privacy of users and has become a hot topic in the community of machine learning. The traditional group probabilities based learning methods have gained certain success, however, they still fall short when the prior information are not fully provided. In order to solve this problem, a novel transfer learning method called transfer group probabilities based learning machine (TGPLM in abbreviation) is proposed by integrating group probabilities into the principle of structure risk minimization. In TGPLM, a novel learning criteria is proposed based on reusing the related domain knowledge by minimizing domain similarity distance, which makes the proposed TGPLM not only make full use of the group probabilities in the current scene, but also learn the existing useful knowledge in the history scene effectively. Experimental results on the artificial, UCI and PIE face datasets show the effectiveness of the proposed method.

Key words: transfer learning; classification; support vector machine; group probabilities

1 引言

目前,在政治选举、欺诈检测和垃圾邮件过滤等领域,利用数据的组概率知识来解决模式分类问题受到越来越多的关注[1~4].如文献[2]中的投票选举事件,对于整个参加选举的地区(或区域)每张选票的结果是不公开的,但对该地区(区域等)划分的各个子区域而言,关于每个候选人的得票情况是清楚的,这些数据就构成了该区域内各个候选人的类标签组概率,若能对这些组概率数据进行有效地的分析,那么其分析的结果可以为未来所进行的选举提供一种有价值的参考.此类情况在欺诈监测[2]和垃圾邮件鉴别[4]等领域同样大量存在,其特点是仅已知数据分组的类标签组概率的条件下的类别判定任务,其是一种介于有监督学习和无监督学习之间的一种特殊的半监督学习方法.组概率信息的优势是对原始数据提供的隐私保护性,而隐私保护恰是当前社会关注的重点及研究的热点之一[5~7],这就使得如何利用组概率信息构建一种精度高,泛化能力强的数据分类模型变得十分重要.

目前为了解决上述场景下的数据分类任务已存在一些研究成果及相应的处理策略[2,3],其中有代表性的工作是 2010 年 Stefan Rueping 提出的反向标定支持向量

收稿日期:2013-03-29;修回日期:2013-05-08;责任编辑:马兰英
基金项目:国家自然科学基金(No.61272210,No.61170122);江苏省自然科学基金(No.BK2012552)

机(Inverse Calibration Support vector machine, IC-SVM),但该方法却忽略了对历史相似场景数据的利用.迁移学习[8~10]作为一种有效利用历史数据的新型学习策略,放松了对训练数据和测试数据同分布的要求,从而使得当前的学习过程更为快速有效[8,9].众多学者已就迁移学习相关问题展开了深入研究,有代表性的有洪佳明等人[11]提出的基于领域相似性的迁移学习方法(TrSVM),Gao 等人[12]提出的局部加权嵌入学习算法(LWE)和 Brian 等人[13]提出一种基于特征空间的大间隔直推式迁移学习方法(LMPROJ)等.这些研究充分说明了迁移学习作为一种机器学习新方法的有效性和实用性.

本文从迁移学习的角度重新审视了上述仅含组概率信息的数据分类问题,构造了一种新颖的迁移组概率学习机(Transfer Group Probabilities based Learning Machine, TGPLM).该方法的主要思想在于将当前领域(目标领域)的数据组概率知识和历史领域(源领域)的已标注数据融入结构风险最小化学习框架中,通过构造领域相似距离项来实现不同领域知识的迁移,从而构造了一种基于迁移学习机制的优化目标学习准则,并通过相关的理论得证新分类器的求解过程依然是一个二次规划(Quadratic Programming, QP)问题.相较于现有的相关方法,本文工作具有如下优势:(1),新方法同时借鉴了历史数据和新领域数据组概率,最大程度体现了当前和历史领域迁移学习过程中的相似知识提取和领域间相互学习的能力;(2),由于继承了基于经验风险最小化框架的支持向量机的优点,使得所提方法的寻优能力在理论上得到了保证;(3)将组概率信息作为一种知识的具体表现形式用以进行地知识迁移,提高了算法的隐私保护性,这也是以往的类似算法所不具备的.

2 反向标定技术(IC)

文献[2]提出了基于 Platt 模型[14]的反向标定技术(Inverse Calibration, IC),从而将组概率信息应用在数据分类问题上.为了在后文中引出本研究的 TGPLM 方法本节将对反向标定技术 IC 做简要描述.

Platt 标定技术最初用来标定支持向量机[15](Support Vector Machine, SVM)的输出,文献[16]提出了利用 Sigmoid 函数估计 SVM 后验概率的输出方法,即

$$p(y=1|\boldsymbol{x})=1/(1+\exp(-Af(\boldsymbol{x})+B)), \quad (1)$$

式(1)中 $\boldsymbol{x}$ 为样本特征向量,y 为样本标签,且 $y\in\{-1,1\}$,$p(y=1|\boldsymbol{x})$为标签为正的概率,参数 A 和 B 通过最小交叉熵获得.文献[2]所提反向标定技术(Inverse Calibration)的主要思想是:取 $A=1$ 和 $B=0$,则式(1)可化成如下形式:

$$p=\sigma(y)=\frac{1}{1+\exp(-y)} \quad (2)$$

变形得: $$y=\sigma^{-1}(p)=-\log(\frac{1}{p}-1) \quad (3)$$

式(3)中 p 为标签为正的样本概率.实际应用时为了避免出现无效的 y 值,限定 $p\in[\varepsilon,1-\varepsilon]$,$\varepsilon$ 为分类估计器精度.

由于实际情况下是很难获取每个样本数据所对应的类标签概率,所以更合理的方式是用组 S_i 中类标签估计的平均值来逼近分组类标签的预测值,即式(4):

$$\forall\, i:\frac{1}{|S_i|}\sum(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j+b)\approx y_i \quad (4)$$

其中$(\boldsymbol{w},b)$为样本分类超平面 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ 的参数对,该式可构建适用于类标签隐藏仅知组概率的数据的支持向量机方法,详细过程参见文献[2].

3 迁移组概率学习机

本文以结构风险最小化模型及相关的组概率知识构造了迁移组概率学习方法模型,其原理如图 1 所示.

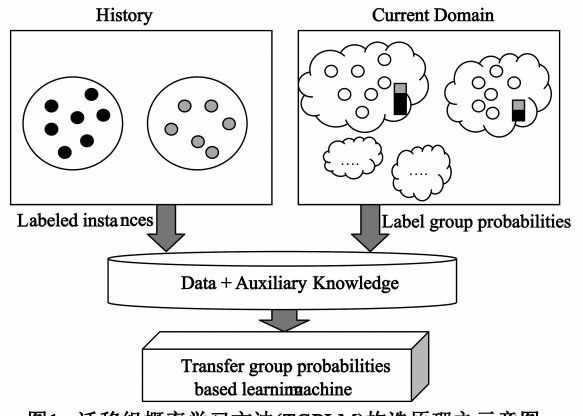


图1 迁移组概率学习方法(TGPLM)构造原理之示意图

由图 1 可知所提方法从已标记的历史数据和具有类标签组概率信息的当前数据这两者中全面地获取了有用知识,这样既保护了当前所研究数据类标签的隐蔽性,又同时借助历史数据和当前领域数据类标签概率来改善算法因类标签保护所带来的性能下降之缺陷.

3.1 融合数据和组概率的目标函数构造

本文将研究重点置于最基本的二元分类问题上,对 TGPLM 算法的具体形式构造如下:

$$\min_{f_h,f_c\in H_K} J_{\text{history}}(f_h)+J_{\text{current}}(f_c)+\lambda d(f_h,f_c) \quad (5)$$

其中,$J_{\text{history}}(f_h)=C_hV_h(D^h,f_h)+\frac{1}{2}\|f_h\|_K^2$,

$$J_{\text{current}}(f_c)=C_cV_c(D^c,f_c)+\frac{1}{2}\|f_c\|_K^2$$

针对式(5)给出如下说明:

(1)f_h 为历史领域 D^h 和当前领域 D^c 学习得到的决

策函数;H_K 为特征空间下的函数集合.

(2)$J_{\text{history}}(f_h)$为历史领域的风险函数,包含结构风险项$\|f_h\|_K^2$和经验风险项 $V_h(D^h,f_h)$,其中$\|f_h\|_K^2$为f_h在特征映射核空间的 L_2 范式,C_h 为历史领域正则化参数.

(3)$J_{\text{current}}(f_c)$为针对当前领域的风险函数,包含结构风险项$\|f_c\|_K^2$和经验风险项 $V_c(D^c,f_c)$,其中$\|f_c\|_K^2$为f_c在特征映射核空间的 L_2 范式,C_c 为当前领域正则化参数.

(4)$d(f_h,f_c)$为当前领域和历史领域间差异项,λ为 $d(f_h,f_c)$的惩罚程度控制参数.

对于 $d(f_h,f_c)$,本文引入了如下具体实现形式:

$$d(f_h,f_c)=\frac{1}{2}(\|\boldsymbol{w}_c-\boldsymbol{w}_h\|^2+(b_c-b_h)^2) \tag{6}$$

式(6)中$(\boldsymbol{w}_c,b_c)$和$(\boldsymbol{w}_h,b_h)$分别定义了当期领域和历史领域数据的分类超平面.

进一步地,以 SVM 学习框架为基础,结合公式(4),(5)和(6),本文给出 TGPLM 的原始优化问题:

$$\begin{aligned}\min_{w_c,w_h,b_c,b_h}\ &\frac{1}{2}\|\boldsymbol{w}_c\|^2+\frac{1}{2}\|\boldsymbol{w}_h\|^2+C_h\sum_{i=1}^{n}\xi_i^h+C_c\sum_{i=n+1}^{n+d}(\xi_i+\xi_i^*)\\&+\frac{\lambda}{2}(\|\boldsymbol{w}_c-\boldsymbol{w}_h\|^2+(b_c-b_h)^2)\end{aligned} \tag{7}$$

$$\text{s.t. } y_i(\boldsymbol{w}_h^{\mathrm{T}}\boldsymbol{x}_i+b_h)\geqslant 1-\xi_i^h,\ i=1,\cdots,n,$$

$$\forall_{i=1}^{d}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^{\mathrm{T}}\boldsymbol{x}_j+b_c)\geqslant y_i-\varepsilon_i-\xi_i,$$

$$i=n+1,\cdots,n+d,$$

$$\forall_{i=1}^{d}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^{\mathrm{T}}\boldsymbol{x}_j+b_c)\leqslant y_i+\varepsilon_i+\xi_i^*,$$

$$i=n+1,\cdots,n+d.$$

其中 $\xi_i^h,\xi_i,\xi_i^*\geqslant 0$,$\xi=[\xi_1^h,\cdots,\xi_n^h,\xi_1,\cdots,\xi_d,\xi_1^*,\cdots,\xi_d^*]^{\mathrm{T}}$为松弛向量;$y_i,i=n+1,\cdots,n+d$ 为反向标定输出值;n 为历史领域样本数,d 为当前领域分组数,本文参考文献[2]的相关分组策略,所分每组样本个数相同;C_h 和 C_c 为为历史领域和当前领域领域正则化参数(惩罚误差程度).

对式(7)所示优化目标函数,本文给出如下说明:

(1)$\frac{1}{2}\|\boldsymbol{w}_c\|^2+\frac{1}{2}\|\boldsymbol{w}_h\|^2+C_h\sum_{i=1}^{n}\xi_i^h+C_c\sum_{i=n+1}^{n+d}(\xi_i+\xi_i^*)$分别表示历史领域数据和当前领域数据的结构风险项和经验风险项.

(2)$\frac{\lambda}{2}(\|\boldsymbol{w}_c-\boldsymbol{w}_h\|^2+(b_c-b_h)^2)$反映了当前领域和历史领域分类器的差异程度.

(3)约束条件 $y_i(\boldsymbol{w}_h^{\mathrm{T}}\boldsymbol{x}_i+b_h)\geqslant 1-\xi_i^h,i=1,\cdots,n$,是为了保证历史领域中分类器尽可能正确分类.而约束条件$\forall_{i=1}^{d}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^{\mathrm{T}}\boldsymbol{x}_j+b)\geqslant y_i-\varepsilon_i-\xi_i$ 和 $\forall_{i=1}^{d}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^{\mathrm{T}}\boldsymbol{x}_j+b)\leqslant y_i+\varepsilon_i+\xi_i^*,i=n+1,\cdots,n+d$,则表示在当前领域中关于数据子集的 S_i 的决策值与 p_i 的反向标定值尽可能接近.

(4)ε_i 为当前领域数据子集 S_i 中反向标定而得 y_i 的逼近精度,本文采用文献[2]相同的方法,取 $\varepsilon_i=\frac{\varepsilon'}{p_i(1-p_i)}$,其中 p_i 为 S_i 中标签为正的数据的组概率,ε'为一个较小的正常数.

这里值得指出的是,针对数据分布相近时的学习问题已有不少的研究工作,如在多视角学习[17]和模糊系统建模[8]方面的研究,与之相比本文工作的特点在于采用了不同的知识迁移策略,即通过历史领域的分类超平面参数来指导当前领域的分类超平面参数的学习.另外,值得指出的是本文工作的迁移学习涉及到组概率信息,这也是相关文献未曾涉及到的.

3.2　相关定理推导和证明

式(7)所示原始问题可转化为如下的对偶问题进行求解:

定理 1　TGPLM 原始优化问题的对偶问题为:

$$\min_{\beta}\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\beta}+\tilde{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{\beta},\qquad \text{s.t. } \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\beta}=0. \tag{8}$$

其中 $\boldsymbol{\beta}=[\boldsymbol{\alpha}^h,\boldsymbol{\alpha},\boldsymbol{\alpha}^*]^{\mathrm{T}}$,

$$0\leqslant\boldsymbol{\beta}\leqslant[\underbrace{C_h,\cdots,C_h}_{n}\ \underbrace{C_c,\cdots,C_c}_{d}\ \underbrace{C_c,\cdots,C_c}_{d}],$$

$$\boldsymbol{f}^{\mathrm{T}}=[y_1,\cdots,y_n,\underbrace{1,\cdots,1}_{d},\underbrace{-1,\cdots,-1}_{d}],$$

$$\tilde{\boldsymbol{e}}=[\underbrace{0,\cdots,0}_{n},\varepsilon-y,\varepsilon+y],$$

$$\tilde{K}=\begin{bmatrix}\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{h,h}+\frac{1}{\lambda} & \frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c} & -\frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}\\ \frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}^{\mathrm{T}} & \frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c} & -\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c}\\ -\frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}^{\mathrm{T}} & -\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c} & \frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c}\end{bmatrix}_{(n+2d)\times(n+2d)},$$

$$\boldsymbol{K}_{h,h}=(y_iy_jk(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,\cdots,n},$$

$$\boldsymbol{K}_{h,c}=(\frac{y_i}{|S_k|}\sum_{j\in S_k}k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i=1,\cdots,n,k=1,\cdots,d},$$

$$\boldsymbol{K}_{c,c}=(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}k(\boldsymbol{x}_{i'},\boldsymbol{x}_{j'}))_{i,j=1,\cdots,d}.$$

证明　最小值问题式(7)的拉格朗日函数为:

$$\begin{aligned}&L(\boldsymbol{w}_c,\boldsymbol{w}_h,b_c,b_h,\boldsymbol{\xi},\boldsymbol{\xi}^*,\boldsymbol{\xi}^h,\boldsymbol{\alpha},\boldsymbol{\alpha}^*,\boldsymbol{\alpha}^h)\\&=\frac{1}{2}\|\boldsymbol{w}_c\|^2+\frac{1}{2}|\boldsymbol{w}_h|^2+C_h\sum_{i=1}^{n}\xi_i^h+C_c\sum_{i=n+1}^{n+d}(\xi_i+\xi_i^*)\\&+\frac{\lambda}{2}(|\boldsymbol{w}_c-\boldsymbol{w}_h|^2+(b_c-b_h)^2)-\sum_{i=1}^{n}r_i^h\xi_i^h-\sum_{i=n+1}^{n+d}r_i\\&\xi_i-\sum_{i=n+1}^{n+d}r_i^*\xi_i^*-\sum_{i=1}^{n}\alpha_i^h(y_i(\boldsymbol{w}_h^{\mathrm{T}}\boldsymbol{x}_j+b_h)-1+\xi_i^h)\end{aligned}$$

$$-\sum_{i=n+1}^{n+d}\alpha_i\Big(\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^T\boldsymbol{x}_j+b_c)-y_i+\varepsilon_i+\xi_i\Big) \tag{9}$$

$$-\sum_{i=n+1}^{n+d}\alpha_i^*\Big(y_i+\varepsilon_i+\xi_i^*-\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_c^T\boldsymbol{x}_j+b_c)\Big)$$

其中，$\boldsymbol{\alpha}^h=(\alpha_1^h,\cdots,\alpha_n^h)$，$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_d)$，$\boldsymbol{\alpha}^*=(\alpha_1^*,\cdots,\alpha_d^*)$，$\boldsymbol{r}^h=(r_1,\cdots,r_n)$，$\boldsymbol{r}=(r_1,\cdots,r_d)$，$\boldsymbol{r}^*=(r_1^*,\cdots,r_d^*)$是拉格朗日系数. 根据 Karush-Kuhn-Tucker(KKT)[18]条件：

$$\frac{\partial L}{\partial\xi_i^h}=0\Rightarrow C_h=\alpha_i^h+r_i^h \tag{10}$$

$$\frac{\partial L}{\partial\xi_i^{(*)}}=0\Rightarrow C_c=\alpha_i^{(*)}+r_i^{(*)} \tag{11}$$

$$\frac{\partial L}{\partial\boldsymbol{w}_c}=0\Rightarrow\boldsymbol{w}_c+\lambda(\boldsymbol{w}_c-\boldsymbol{w}_h)$$

$$-\sum_{i=n+1}^{n+d}\frac{\alpha_i}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j+\sum_{i=n+1}^{n+d}\frac{\alpha_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j=0 \tag{12}$$

$$\frac{\partial L}{\partial\boldsymbol{w}_h}=0\Rightarrow\boldsymbol{w}_h-\lambda(\boldsymbol{w}_c-\boldsymbol{w}_h)-\sum_{i=1}^{n}\alpha_i^hy_ix_i=0 \tag{13}$$

$$\frac{\partial L}{\partial b_c}=0\Rightarrow\lambda(b_c-b_h)=\sum_{i=n+1}^{n+d}(\alpha_i-\alpha_i^*) \tag{14}$$

$$\frac{\partial L}{\partial b_h}=0\Rightarrow-\lambda(b_c-b_h)=\sum_{i=1}^{n}\alpha_i^hy_i \tag{15}$$

将式(10)~(15)回代到式(9)并化简后可得其对偶问题如下：

$$\min_{\alpha^h,\alpha,\alpha^*}\frac{1+\lambda}{2(1+2\lambda)}\Big(\sum_{i=n+1}^{n+d}\sum_{j=n+1}^{n+d}\frac{(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)}{|S_i||S_j|}$$

$$\sum_{i'\in S_i}\sum_{j'\in S_j}\boldsymbol{x}_{i'}^T\boldsymbol{x}_{j'}+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i^h\alpha_j^hy_iy_j\boldsymbol{x}_i^T\boldsymbol{x}_j\Big)$$

$$+\frac{\lambda}{2(1+2\lambda)}\Big(\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\alpha_iy_i\frac{(\alpha_j-\alpha_j^*)}{|S_j|}\sum_{k\in S_j}\boldsymbol{x}_i^T\boldsymbol{x}_k$$

$$+\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\alpha_i^hy_i\frac{(\alpha_j-\alpha_j^*)}{|S_j|}\sum_{k\in S_j}\boldsymbol{x}_k^T\boldsymbol{x}_i\Big)$$

$$+\frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i^h\alpha_j^h+\sum_{i=n+1}^{n+d}\alpha_i(\varepsilon_i-y_i)+\sum_{i=n+1}^{n+d}\alpha_i^*(\varepsilon_i+y_i),$$

$$\text{s.t.}\ \alpha_i^h\in[0,C_h],\alpha_i,\alpha_i^*\in[0,C_c],$$

$$\sum_{i=1}^{n}\alpha_i^hy_i+\sum_{i=n+1}^{n+d}(\alpha_i-\alpha_i^*)=0. \tag{16}$$

为将式(16)化为标准的二次规划形式，此处令 $\boldsymbol{\beta}=[\boldsymbol{\alpha}^h,\boldsymbol{\alpha},\boldsymbol{\alpha}^*]^T$，

$$0\leqslant\boldsymbol{\beta}\leqslant[\underbrace{C_h,\cdots,C_h}_{n}\ \underbrace{C_c,\cdots,C_c}_{d}\ \underbrace{C_c,\cdots,C_c}_{d}],$$

$\boldsymbol{f}^T=[y_1,\cdots,y_n,\underbrace{1,\cdots,1}_{d},\underbrace{-1,\cdots,-1}_{d}]$，$\{y_i\}_{i=1}^n$为历史域中训练样本的标签，

$$\tilde{\boldsymbol{e}}=[\underbrace{0,\cdots,0}_{n},\varepsilon-y,\varepsilon+y],$$

$$\tilde{\boldsymbol{K}}=\begin{bmatrix}\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{h,h}+\frac{1}{\lambda} & \frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c} & -\frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}\\ \frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}^T & \frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c} & -\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c}\\ -\frac{\lambda}{1+2\lambda}\boldsymbol{K}_{h,c}^T & -\frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c} & \frac{1+\lambda}{1+2\lambda}\boldsymbol{K}_{c,c}\end{bmatrix}_{(n+2d)\times(n+2d)},$$

$$\boldsymbol{K}_{h,h}=(y_iy_j\boldsymbol{x}_i^T\boldsymbol{x}_j)_{i,j=1,\cdots,n},$$

$$\boldsymbol{K}_{h,c}=\Big(\frac{y_i}{|S_k|}\sum_{j\in S_k}\boldsymbol{x}_i^T\boldsymbol{x}_j\Big)_{i=1,\cdots,n,k=1,\cdots,d},$$

$$\boldsymbol{K}_{c,c}=\Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}\boldsymbol{x}_{i'}^T\boldsymbol{x}_{j'}\Big)_{i,j=1,\cdots,d}.$$

化简后可得标准的二次规划形式，如式(17)所示：

$$\min_{\beta}\frac{1}{2}\boldsymbol{\beta}^T\tilde{\boldsymbol{K}}\boldsymbol{\beta}+\tilde{e}^T\boldsymbol{\beta} \tag{17}$$

$$\text{s.t.}\ \boldsymbol{f}^T\boldsymbol{\beta}=0$$

一般地，真实样本空间很难做到准确划分，为此需要进行核化，其实质是找到一个合适的映射 $\varphi:\boldsymbol{x}_i\in R^d\rightarrow\varphi(x_i)\in R^D(d<<D)$，并用核函数 $k(\mu,v)$表示映射后的内积 $\varphi(\mu)^T\varphi(v)$，令

$$\boldsymbol{K}_{h,h}=(y_iy_jk(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,\cdots,n}$$

$$\boldsymbol{K}_{h,c}=\Big(\frac{y_i}{|S_k|}\sum_{j\in S_k}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\Big)_{i=1,\cdots,n,k=1,\cdots,d}$$

$$\boldsymbol{K}_{c,c}=\Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}k(\boldsymbol{x}_{i'},\boldsymbol{x}_{j'})\Big)_{i,j=1,\cdots,d}$$

式(17)核化后可得式(8). (证毕)

定理 2 [19]：式(8)所示对偶问题所转化的二次规划问题为凸二次规划问题.

证明 式(8)中 $\tilde{\boldsymbol{K}}$ 可以表示成如下形式：

$$\tilde{\boldsymbol{K}}=\tilde{\boldsymbol{K}}_1+\tilde{\boldsymbol{K}}_2+\tilde{\boldsymbol{K}}_3+\frac{1}{\lambda}I,I=[1]_{n\times n},$$

$$\tilde{\boldsymbol{K}}_1=\frac{\lambda}{1+2\lambda}\begin{bmatrix}\boldsymbol{K}_{h,h} & \boldsymbol{K}_{h,c} & -\boldsymbol{K}_{h,c}\\ \boldsymbol{K}_{h,c}^T & \boldsymbol{K}_{c,c} & -\boldsymbol{K}_{c,c}\\ -\boldsymbol{K}_{h,c}^T & -\boldsymbol{K}_{c,c} & \boldsymbol{K}_{c,c}\end{bmatrix}_{(n+2d)\times(n+2d)},$$

$$\tilde{\boldsymbol{K}}_2=\frac{1}{1+2\lambda}\begin{bmatrix}\boldsymbol{K}_{h,h} & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}_{(n+2d)\times(n+2d)},$$

$$\tilde{\boldsymbol{K}}_3=\frac{1}{1+2\lambda}\begin{bmatrix}0 & 0 & 0\\ 0 & \boldsymbol{K}_{c,c} & -\boldsymbol{K}_{c,c}\\ 0 & -\boldsymbol{K}_{c,c} & \boldsymbol{K}_{c,c}\end{bmatrix}_{(n+2d)\times(n+2d)}.$$

令

$$\boldsymbol{Q}_1=\sqrt{\frac{\lambda}{1+2\lambda}}\Big(y_1\boldsymbol{x}_1,\cdots,y_n\boldsymbol{x}_n,\frac{1}{|S_1|}\sum_{i\in S_1}x_i,\cdots,$$

$$\frac{1}{|S_d|}\sum_{i\in S_d}\boldsymbol{x}_i,-\frac{1}{|S_1|}\sum_{i\in S_1}\boldsymbol{x}_i,\cdots,-\frac{1}{|S_d|}\sum_{i\in S_d}\boldsymbol{x}_i\Big),$$

则易见 $\tilde{\boldsymbol{K}}_1=\boldsymbol{Q}_1^T\boldsymbol{Q}_1$，所以 $\tilde{\boldsymbol{K}}_1$ 是半正定矩阵，同理可知 $\tilde{\boldsymbol{K}}_2$ 和 $\tilde{\boldsymbol{K}}_3$ 也是半正定矩阵，所以 $\tilde{\boldsymbol{K}}$ 为半正定矩阵，由此

得证式(8)所示二次规划为凸二次规划.(证毕)

定理 3 [19]:求解式(8)的二次规划问题得到的解为全局最优解.

证明　因为式(8)的二次规划为凸二次规划,则KKT条件也是充分条件,因此得到的二次规划的解为全局最优解.(证毕)

定理 4 [19]:设 $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\alpha}}^h, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}^*)^{\mathrm{T}}$ 是对偶问题式(8)的解,则式(7)所示 TGPLM 的原始优化问题对于 $\boldsymbol{w}_c$ 和 b_c 的解存在全局最优解,并可表示为:

$$\boldsymbol{w}_c^* = \frac{\lambda}{1+2\lambda}\sum_{i=1}^{n}\tilde{\alpha}_i^h y_i \boldsymbol{x}_i + \frac{1+\lambda}{1+2\lambda}\sum_{i=n+1}^{n+d}\frac{\tilde{\alpha}_i - \tilde{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j \tag{18}$$

$$b_c^* = y_i - \frac{\lambda}{1+2\lambda}\sum_{j=1}^{n}\frac{\tilde{\alpha}_j^h y_j}{|S_i|}\sum_{k\in S_i}k(\boldsymbol{x}_j,\boldsymbol{x}_k) - \frac{1+\lambda}{1+2\lambda}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_j - \tilde{\alpha}_j^*}{|S_j||S_i|}\sum_{l\in S_j}\sum_{k\in S_i}k(\boldsymbol{x}_l,\boldsymbol{x}_k)) \tag{19}$$

证明　根据定理 2 以及定理 3 的证明,可知式(8)为凸二次规划,而又根据定理 3 的满足条件可知该二次规划的解为全局最优解.

因此,若 $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\alpha}}^h, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}^*)^{\mathrm{T}}$ 为式(8)的解,那么根据式(12)和(13)可解得:

$$\boldsymbol{w}_c^* = \frac{\lambda}{1+2\lambda}\sum_{i=1}^{n}\tilde{\alpha}_i^h y_i \boldsymbol{x}_i + \frac{1+\lambda}{1+2\lambda}\sum_{i=n+1}^{n+d}\frac{\tilde{\alpha}_i - \tilde{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j,$$

选取 n 个 α_j^h 位于开区间 $(0, C_h)$,d 个 α_j 和 α_j^* 位于开区间 $(0, C_c)$ 的分量 $(\tilde{a}_j^h, \tilde{a}_j, \tilde{a}_j^*)^T$,可以计算得出

$$b_c^* = y_i - \frac{\lambda}{1+2\lambda}\sum_{j=1}^{n}\frac{\tilde{\alpha}_j^h y_j}{|S_i|}\sum_{k\in S_i}k(\boldsymbol{x}_j,\boldsymbol{x}_k) - \frac{1+\lambda}{1+2\lambda}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_j - \tilde{\alpha}_j^*}{|S_j||S_i|}\sum_{l\in S_j}\sum_{k\in S_i}k(\boldsymbol{x}_l,\boldsymbol{x}_k)),$$

由此得到的 $\boldsymbol{w}_c^*$ 和 b_c^* 则为原始问题(7)的全局最优解.(证毕)

这里值得指出的是,对于式(18)和(19)所给出的最优解同时包含了从当前领域和历史领域的信息,如 $\boldsymbol{w}_c^*$ 中 $\frac{\lambda}{1+2\lambda}\sum_{i=1}^{n}\tilde{\alpha}_i^h y_i \boldsymbol{x}_i$ 部分为从历史领域中学习得到的知识,$\frac{1+\lambda}{1+2\lambda}\sum_{i=n+1}^{n+d}\frac{\tilde{\alpha}_i - \tilde{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j$ 部分则为从当前领域中学习获取的知识.

3.3　TGPLM 算法流程

由上述分析可得 TGPLM 方法的具体步骤如表 1 所示.

3.4　TGPLM 的问题复杂度分析

TGPLM 的训练复杂度主要由其对应的二次规划问题决定,以经典的二次规划问题解法为例[15],所提方法空间复杂度为($O(N^2)$),时间复杂度为($O(N^3)$),$N = n + d$,其中 n 为历史领域样本个数,d 为当前领域样本分组数.

表 1　TGPLM 算法流程

步骤	迁移组概率学习机(TGPLM)
Input	n 个有标号的历史领域样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$,$m$ 个无标号的当前领域样本 $\{\boldsymbol{x}_j\}_{j=n+1}^{n+m}$,$d$ 个当前领域组概率 $\{(S_k, p_k\}_{k=1}^d$.
Output	分类决策函数 $g(\boldsymbol{x})$.
Step 1	根据公式(3)和 $\{(S_k, p_k\}_{k=1}^d$ 计算当前领域数据的分组的反向标定输出 $y_k, k = 1, \ldots, d$,进而结合 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$,$\{\boldsymbol{x}_j\}_{j=n+1}^{n+m}$ 计算 $\boldsymbol{K}_{h,h}, \boldsymbol{K}_{h,c}, \boldsymbol{K}_{c,c}$,最后求出矩阵 $\tilde{\boldsymbol{K}}$;
Step 2	根据定理 1 构造矩阵 $\tilde{\boldsymbol{K}}$,求解拉格朗日系数 β;
Step 3	根据公式(18)计算决策超平面法向量 $\boldsymbol{w}_c$;
Step 4	根据公式(19)计算偏移量 b_c;
Step 5	输出分类决策函数 $f(\boldsymbol{x}) = \boldsymbol{w}_c^T\varphi(\boldsymbol{x}) + b_c$.

4　实验结果与分析

本节将在几种不同类型的数据集上进行实验:(1)人工二维团状随机高斯型数据集;(2)不同领域真实数据集(包括 20Newsgroup[20,21], Reuters[20,21],垃圾邮件检测数据集[9~22]和入侵检测分类数据集[12]);(3)人脸图像分类数据集 PIE[23].

对测试人造数据集中主要引入 SVM[24], IC-SVM[2] 两种算法进行比较;在测试真实数据集的实验中主要引入 SVM, TSVM[25], TrSVM[11], LWE[12], LMPROJ[13] 五类算法进行比较.

本文方法与其他方法进行学习能力比较时,以当前领域数据分类的精度为所提方法评价指标,具体的指标为:Accuracy = $\frac{|\{x \mid x_c \in D_c \cap f(x_i) = y_c\}|}{|\{x \mid x_c \in D_c\}|}$,其中 D_c 表示当前领域数据集,y_c 表示 x_c 的真实标签类别,$f(x)$ 为使用学习所得分类器对 x_c 进行分类所得结果.

本文所有实验均通过网格搜索的方式来确定优化的实验参数.判定参数性能的标准采用文献[2]相同的策略,在训练集上以 10 倍交叉验证时所得分类精度为评价指标.在核函数选择上均采用高斯核函数,核宽度参数 $2\sigma^2$ 以源领域样本的平均 2 范数的平方 s 为基准,并在网格 $\{\frac{s}{64}, \frac{s}{32}, \frac{s}{16}, \frac{s}{8}, \frac{s}{4}, \frac{s}{2}, s, 2s, 4s, 8s, 16s, 32s, 64s\}$ 中搜索最优值;TGPLM 的正则化参数 C_h 和 C_c 在网格 $\{2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$ 中搜索最优值;平衡参数 λ 在区间 $\{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}\}$ 中搜索最优值.所有实验均在 Intel Core2, 2.0GHz 主频,2G RAM, Windows XP 系统下执行,SVM 算法由 Libsvm[26]软件实现,其他算法均在 Matlab R2009A 环境下实现.

对于所有数据集,历史领域和当前领域数据均具有标签信息,但当前领域标签信息仅用于学习方法分类性能的客观量化评价.

4.1 人工数据集

如图 2 所示,本人工生成两个分布服从不同团状高斯分布的二类 2-D 样本集,分别代表历史领域(HD)和当前领域(CD),HD 和 CD 中正负样本点数均为 50. HD 中样本均值为[1.2923 1.3959],方差为[2.9059 2.7528],CD 中样本均值为[1.2516 2.1858],方差为[2.3416 0.9897],其中 HD + ,HD - 和 CD + ,CD - 分布代表源领域和目标领域中正类和负类样本. 实验中当前领域分组中数据个数 K 分别取 2,4,8,16,每组中正类样本所占比例作为其组概率信息. 表 2 给出了不同分组下的实验结果,同时图 2 给出了 $K=8$ 时 TGPLM 的分类超平面和其他两种方法的分类超平面.

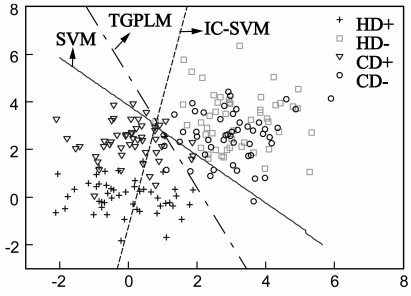


图2 SVM, IC-SVM和TGPLM的分类超平面

根据表 2 和图 2,我们给出如下观察:

(1)由表 2 可以看出:当每组内样本个数为 2 时,IC-SVM 与所提方法具有相同的分类性能,这说明在组概率信息较为丰富时,迁移学习机制已不能带来有益帮助;随组内数据量的增加,所提方法因采用迁移学习机制相对于非迁移学习的 IC-SVM 方法的优势逐渐明显.

(2)根据图 2 所得之实验结果可以看出:SVM 仅考虑了历史领域样本分类效果达到最优,导致领域适应性能最差;IC-SVM 忽视了历史领域的信息来辅助学习,分类效果也不理想;所提方法 TGPLM 则继承了 SVM 和 IC-SVM 方法的优点,既利用了当前领域的类标签概率信息,还充分考虑了领域间的相似性,从而获得了较好的决策分割线,性能在一定程度上优于 SVM 和 IC-SVM 方法.

表 2 团状随机高斯数据分类精度(粗体表示最优结果)

方法 \ K	2	4	8	16
SVM	66.57	66.57	66.57	66.57
IC-SVM	**92.06**	90.84	85.12	70.28
TGPLM	**92.06**	**91.60**	**90.01**	**85.89**

4.2 真实数据集

根据 4.1 小节中人工数据实验分析可知,若组内数据个数过少,数据的隐私保护性能降低;而组内数据过多则可利用的信息量急剧减少,组概率学习机不能有效的工作,综合考虑如上因素,下面各实验均将当前领域数据集每组数据的个数设置为 8,每组正类样本的比例作为该组的组概率信息. 实验所采用的数据集及各种算法的分类效果如下:

(1)不同领域真实数据集. 包括跨领域文本数据集 Reuters 和 20Newsgroups,垃圾邮件过滤数据集和入侵检测数据集. 各数据集预处理参见对应参考文献,详细信息见表 3 所示,相应的实验结果见表 4 所示.

表 3 各不同领域真实数据集详细描述

	学习任务	数据集	历史领域数据		当前领域数据	
			正类	负类	正类	负类
1	Reuters 20Newsgroups	Orgs vs. People	588	649	587	621
2		Orgs vs. Place	428	588	456	587
3		People vs. Place	428	649	456	621
4		Comp vs. Sci	996	998	997	999
5		Rec vs. Talk	998	1000	997	998
6		Rec vs. Sci	998	998	998	999
7		Sci vs. Talk	998	1000	999	998
8		Comp vs. Rec	992	998	998	997
9		Comp vs. Talk	992	1000	998	1000
10	Email spam filtering	User1 vs. User2	User1's email		User2's email	
11		User2 vs. User3	User'2s email		User3's email	
12		User3 vs. User1	User3's email		User1's email	
13	Intrusion	Dos&R2L vs. Probing	Dos&R2L&Normal		Probing &Normal	
14		Probing &Dos vs. R2L	Probing &Dos &Normal		R2L&Normal	
15		R2L&Probing vs. Dos	R2L&Probing &Normal		Dos &Normal	

表 4　各种方法在不同数据集下的精度比较结果

学习任务	SVM	TSVM	TrSVM	LWE	LMPROJ	TGPLM
1	69.31	72.84	73.14	79.61	**80.63**	79.29
2	69.95	69.80	71.16	72.87	75.20	**76.79**
3	56.93	63.00	67.32	68.49	70.80	**70.45**
4	72.26	76.23	71.23	83.11	82.12	**83.27**
5	70.10	73.40	79.66	76.60	79.30	**81.85**
6	78.33	83.90	85.14	**88.41**	86.68	87.79
7	75.90	81.20	85.09	80.68	84.21	**85.87**
8	83.77	83.24	84.34	85.40	86.78	**90.81**
9	90.60	90.74	93.56	94.43	**95.43**	93.93
10	95.98	95.13	95.08	93.22	93.62	**97.15**
11	96.69	96.00	95.45	**98.53**	94.01	97.90
12	90.17	89.73	89.34	88.68	88.65	**93.68**
13	89.02	88.71	83.83	**96.36**	89.56	95.97
14	92.75	84.25	87.52	96.23	88.51	**96.88**
15	90.18	89.21	77.83	80.24	87.95	**93.41**

(2)PIE 人脸数据集. PIE 数据库包含 68 个人的 41368 幅人脸灰度图像,随机选取 1 名男性和 1 名女性,各 170 幅人脸图像构成一个二类数据集进行实验. 分别进行逆时针旋转 10 度、30 度、50 度,以形成变化的当前领域图像数据集. 实验前,对上述图像集进行预处理,使得其缩放到 32 × 32 像素大小,且每个像素为 256 灰度级,则在图像空间,每幅图像由一个 1024 维的向量表示. 图 3(*a*)和图 3(*b*)分别显示了旋转前后的部分图像,对应的男性标签为 1 号,女性标签为 35 号. 表 5 给出了不同算法在 PIE 数据集上的实验结果.

(*a*) 历史领域人脸样本

(*b*) 逆时针旋转10度当前领域人脸样本

图3　基于PIE人脸数据库构造的历史领域及当前领域样本

根据上述真实数据集上的实验结果可得如下结论:

(1)基线算法 SVM 在几乎所有数据集内分类性能低于其他迁移学习方法,另外一个分类方法 TSVM 的分类效果也普遍不佳;

(2)由于充分考虑了当前领域数据的类标签概率信息及历史领域样本的辅助信息,TGPLM 方法在上述各种种数据集上的分类精度绝大部分都优于其它几类迁移学习方法. 同时上述的实验结果也进一步地说明了利用类标签概率的迁移学习分类方法的有效性.

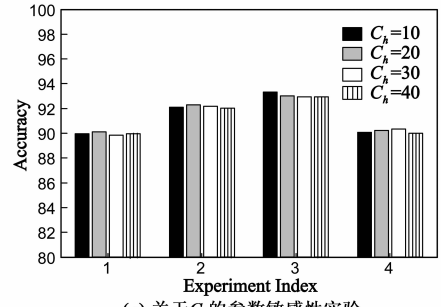
表 5　PIE 数据集上各算法的平均精度

PIE Dataset	SVM	TSVM	TrSVM	LWE	LMPROJ	TGPLM
10 度	73.59	72.84	74.24	75.16	74.63	**80.19**
30 度	69.31	68.80	70.17	72.87	73.20	**79.09**
50 度	64.66	63.80	64.72	67.61	68.80	**76.45**

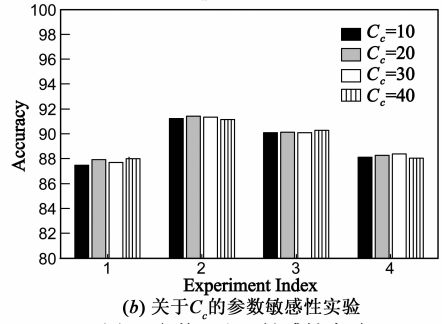
4.3　参数敏感实验

4.3.1　参数敏感实验 1

本节将考察历史领域和当前领域的结构风险正则化参数 C_h 和 C_c 相关取值对所提方法性能的影响. 选用 20Newsgroup 数据集中的 Comp vs. Talk 子数据集,具体的实施方案如下:首先,固定 $C_h = 20$,分别取 C_c 为 10、20、30 和 40,实验结果如图 4(*a*)所示;其次,同样地固定 $C_c = 20$,此时 C_h 则分别取 10、20、30 和 40,实验结果如图 4(*b*)所示.



(*a*) 关于C_h的参数敏感性实验



(*b*) 关于C_c的参数敏感性实验

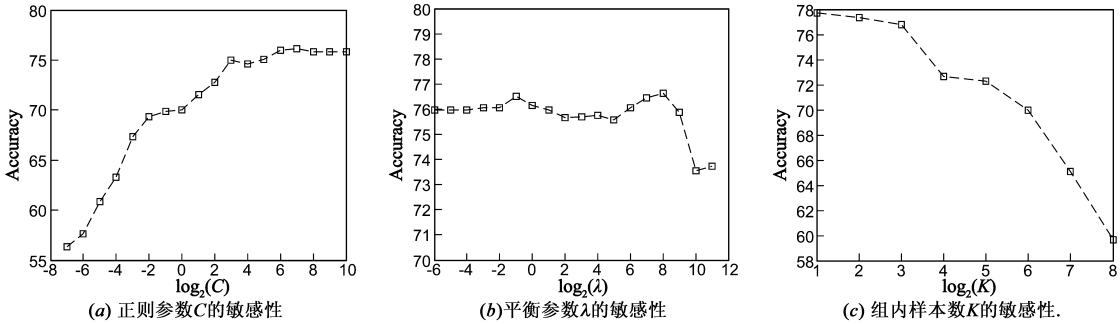
图4　参数C_h和C_c敏感性实验

实验结果说明随参数 C_c 和 C_h 之间的比例变化,所提算法分类性能变化不大,也表明了所提方法对于历史领域和当前领域正则化参数的微小变化存在鲁棒性.

4.3.2　参数敏感实验 2

本节考察当前领域数据所分子集样本个数 K 与可

调参数 C、λ 对算法性能的影响.以表 3 第 2 个 Reuters 数据集作为实验数据,采用两种实验策略:(1)在 $K=8$ 时,分别对 C 与 λ 进行实验分析,具体的结果如图 5(a)和(b)所示;(2)对实际应用参数 K 取不同值时所提方法最佳性能变化规律进行分析,具体如图 5(c) 所示.



(a) 正则参数C的敏感性 (b)平衡参数λ的敏感性 (c) 组内样本数K的敏感性.

图5 参数C, λ和K的敏感性实验

由所得实验结果可得如下结论:

(1)由于所提方法在原理上是基于结构风险最小化学习模型而构造的方法,所以正则化参数 C 对分类效果有较大程度上的影响,这说明了对参数 C 协调的重要性.

(2)随领域间惩罚参数 λ 取值逐渐增加,TGPLM 方法分类性能先缓慢上升后急剧下降.这是因为在 λ 取值很小时,所提方法几乎完全依赖当前领域的类标签概率信息进行知识学习,所以精度不高;随着 λ 取值的增大,所提方法在借鉴当前领域的类标签概率信息的同时学习了历史领域的辅助知识,从而获得了最优分类效果;当 λ 取值很很大时,所提方法过分依赖了历史领域的辅助知识,忽视了历史领域和当前领域数据分布的差异,TGPLM 方法与 SVM 方法所得的决策超平面被强制性趋同,导致了负迁移,使得新方法的分类性能下降.

(3)随着当前领域中数据分组内数据数目 K 的增加,TGPLM 方法分类性能呈现逐渐下降的趋势.这是因为伴随 K 的增加,整个当前领域的分组数目将会减少,类标签概率信息也会随之减少,所以导致所提方法的分类性能下降.

5 结论

本文从迁移学习角度对当前领域仅已知类标签概率信息的模式分类问题进行了探讨,将历史领域的样本信息和当前领域的类标签概率信息同时纳入目标决策函数的构造中,提出一种迁移组概率学习机 TGPLM.在人工和真实数据集实验结果表明,所提 TGPLM 方法不仅具备了 SVM 算法易实现的优点,还具有相似领域的知识迁移学习和目标领域类标签信息保护的功能.对于本文方法今后依然在如下方面值得进一步探讨:(1)多分类问题方面的扩展.由于所采用的反向标定技术仅适用二分类问题,这使得本文方法还不适于多类分类.今后我们将致力于研究一种适用于多类分类的迁移组概率学习机;(2)大数据集方面的扩展.在历史领域样本较大或当前领域组数较多的情况下,如何进行更为快速有效地进行迁移组概率学习机模型的构建也是我们今后尚需研究的内容.

参考文献

[1] Stolpe M, Morik K. Learning from Label Proportions by Optimizing Cluster Model Selection[A]. ECML PKDD 2011[C]. Berlin, Heidelberg, 2011, Part III, Vol.6913, 349 – 364.

[2] Rüping S. SVM classifier estimation from group probabilities [A]. Proceedings of 27th ICML[C]. Haifa, 2010: 911 – 918.

[3] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[A]. Proceedings of 25th ICML[C]. Omnipress, 2008. 776 – 783.

[4] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[J]. Journal of Machine Learning Research, 2009, (10): 2349 – 2374.

[5] 韩建民,于娟,虞慧群,贾▮.面向敏感值的个性化隐私保护[J].电子学报,2010,38(7):1723 – 1728.
Han J M, Yu J, Yu H Q, Jia J. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinica, 2010, 38(7): 1723 – 1728. (in Chinese)

[6] 胡文军,王士同.隐私保护的 SVM 快速分类方法[J].电子学报,2012,40(2):280 – 286.
HU W J, WANG S T. Fast classification approach of support vector machine with privacy preservation[J]. Acta Electronica Sinica. 2012, 40(2): 280 – 286. (in Chinese)

[7] 张战成,王士同,钟富礼.具有隐私保护功能的协作式分类机制[J].计算机研究与发展,2011,48(06):1018 – 1029.
Zhang Z C, Wang S T, Fu L C. Collaborative classification mechanism for privacy-preserving[J]. Journal of Computer Research and Development. 2011, 48(6): 1018 – 1028. (in Chinese).

[8] 蒋亦樟,邓赵红,王士同.ML 型迁移学习模糊系统[J].自动化学报,2012,38(9):1393 – 1409.
Jiang Y Z, Deng Z H, Wang S T. Mamdani-larsen type transfer learning fuzzy system[J]. Acta Automatica Sinica, 2012, 38(9):1393 – 1409. (in Chinese)

[9] Tao J W, Chung F L, Wang S T, On Minimum distribution discrepancy support vector machine for domain adaptation[J]. Pattern Recognition, 2012, 45(11):3962 – 3984.

[10] 于重重,田蕊,谭励,涂序彦.非平衡样本分类的集成迁移学习算法[J].电子学报,2012,40(7):1358 – 1363.
Yu C C, Tian R, Tan L, Tu X Y. Integrated transfer learning algorithmic for unbalanced samples classification[J]. Acta Electronica Sinica, 2012, 40(7):1358 – 1363. (in Chinese)

[11] 洪佳明,印鉴,黄云,等.一种基于领域相似性的迁移学习算法[J].计算机研究与发展,2011,48(10):1823 – 1830.
Hong J M, Yin J, Huang Y, et al. TrSVM: A transfer learning algorithm using domain similarity[J]. Journal of Computer Research and Development, 2011, 48(10):1823 – 1830. (in Chinese)

[12] Gao J, Fan W, Jiang J, Han J W. Knowledge transfer via multiple model local structure mapping[A]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York, USA: ACM, 2008. 283 – 291.

[13] Quanz B, Huan J. Large margin transductive transfer learning[A]. Proceedings of the 18th ACM conference on Information and knowledge management[C]. New York, USA: ACM, 2009. 1327 – 1336.

[14] Platt J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[A]. Advances in Large Margin Classifiers[C]. Cambridge: MIT Press, 1999. 61 – 74.

[15] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995. 123 – 167.

[16] Caruana R and Niculescu M A. Predicting good probabilities with supervised learning[A]. Proceedings of the 22nd International Conference on Machine Learning[C]. Bonn, Germany, 2005. 625 – 632.

[17] Sun S L. Multi-view Laplacian Support Vector Machines[A]. Proceedings of the 7th international conference on Advanced Data Mining and Applications[C]. Beijing, China, 2011: 209 – 222.

[18] Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem[A]. Proceedings of Conference on Learning Theory[C]. Amsterdam: Springer Press, 2001. 416 – 426.

[19] 邓乃杨,田英杰.数据挖掘的新方法——支持向量机[M].北京:科学出版杜,2004.
Deng N Y, Tian Y J. New Method in Data Mining: Support Vector Machine[M]. Beijing. China: Science Press, 2004. (in Chinese)

[20] Xiang E W, Cao B, Hu D H, Yang Q. Bridging domains using world wide knowledge for transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(6):770 – 783.

[21] Bruzzone L, Marconcini M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5):770 – 787.

[22] Bickel S. ECML-PKDD Discovery Challenge 2006 Overview[A]. Proceedings. ECML/PKDD Discovery Challenge Workshop[C]. Berlin, Germany, 2006.

[23] He X F, Cai D, Partha N. Laplacian score for feature selection[A]. Advances in Neural Information Processing Systems18[C]. M A: MIT Press, 2006. 507 – 514.

[24] Vapnik V. Statistical Learning Theory[M]. John Wiley and Sons, 1998.

[25] Joachims T. Transductive inference for text classification using support vector machines[A]. Proceedings of 16th International Conference on Machine Learning[C]. San Francisco, CA: Morgan Kaufmann Publishers, 1999. 200 – 209.

[26] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

作者简介

倪彤光(通信作者)　男,1978 年生于河北邢台.分别在 2000 年、2005 年于江南大学获得工学学士、硕士学位.2011 年进入江南大学攻读博士学位,主要从事模式识别、人工智能等方面的研究.
E-mail: hbxtntg – 12@163.com

王士同　男,1964 年生于江苏扬州.教授、博士生导师、中国计算机学会高级会员.1984 年、1987 年在南京航空航天大学获得工学学士、硕士学位.主要从事人工智能、模式识别、模糊系统、医学图像处理和生物信息学等方面的研究工作.
E-mail: wxwangst@yahoo.com.cn