文章编号: 1001-0920 (2014) 10-1751-07 **DOI:** 10.13195/j.kzyjc.2013.0790

适用于不确定类标签数据学习的迁移支持向量机

倪彤光[1,2], 王士同[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: 为了解决包含不确定信息的分类学习问题, 提出一种新的适用于不确定类标签数据的迁移支持向量机. 该方法基于结构风险最小化模型, 同时将源领域中所学知识、领域间的共享数据、目标领域中已标定的和不确定的数据纳入学习框架中, 进而实现了源领域和目标领域的知识迁移. 在多种真实数据集上的实验结果表明了所提出方法的有效性.

关键词: 迁移学习；分类；支持向量机；共享数据；不确定数据

中图分类号: TP181 **文献标志码:** A

Transfer support vector machine for learning from data with uncertain labels

NI Tong-guang[1,2], *WANG Shi-tong*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Information Science and Technology，Changzhou University，Changzhou 213164，China. Correspondent：NI Tong-guang，E-mail：hbxtntg-12@163.com)

Abstract: To address the learning problems which include unlabeled samples, a novel transfer support vector machine for learning from data with uncertain labels(TSVM-UL) is proposed based on the structure risk minimization model. This method takes knowledge of source domain, the common data between different domains, labeled samples and probabilities of unlabeled samples of target domain into account, and knowledge transferring between the source domain and the target domain is realized. Experiment results on several real-world datasets show the effectiveness of the proposed method.

Key words: transfer learning；classification；support vector machine；common data；uncertain data

0 引 言

随着科学技术的飞速发展, 人们可以获得的信息越来越多, 如何有效地从信息中获取知识变得越来越重要. 众多学者针对此问题展开了广泛的研究, 发现上述问题的根本原因是标记数据的代价较为昂贵, 且在很多应用场景中无法对数据进行定性的标注. 如文献 [1] 对于病患者癌症肿瘤图片进行研究的应用场景中, 即使是治疗肿瘤的专家也很难在短时间内对一个新病例作出精确判断, 往往是先作出一个估计判断. 文献 [2] 指出上述例子中肿瘤专家最初的估计判断作为一种样本的后验概率信息数据, 也可以作为一种训练数据指导分类模型进行模式识别, 并针对分类器训练集中标记数据不足的情况, 提出了利用定性数据 (标签信息) 和定量信息 (后验概率估计) 同时作为模式分类器的可训练数据, 以提高算法的效果. 上述研究仅限于传统的机器学习领域, 即要求训练数据和测试数据满足同分布的要求. 但是, 针对新的兴趣领域, 完全丢弃这些大量的、在不同分布下的训练数据而重新构建训练数据是非常浪费的. 迁移学习即是为了更好地利用这些已有的数据而提出的新的研究方向[3], 它使用历史总结的知识对新实例的学习提供一个有益的补充, 从而使得当前的学习过程更为快速有效[4-5]. 众多学者已就迁移学习的相关问题展开了深入研究, Brian 等[6]提出了一种基于特征空间的大间隔直推式迁移学习方法 (LMPROJ), 该方法通过寻求一个特征变换使得源领域数据与目标领域数据之间的分布距离最小化来实现跨领域学习; Gao 等[7]提出了局部加权嵌入学习算法 (LWE), 依据不同领域训练样

收稿日期: 2013-06-14；**修回日期:** 2013-10-08.

基金项目: 国家自然科学基金项目(61272210, 61170122)；江苏省自然科学基金项目(BK2012552).

作者简介: 倪彤光(1978−), 男, 讲师, 博士生, 从事模式识别、人工智能的研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、人工智能等研究.

本与待测领域样本间的联系, 赋予不同的权重, 进而实现领域间知识的迁移; 洪佳明等[8]提出了基于领域相似性的迁移学习方法(TrSVM), 首先定义了领域弱相似性的概念, 然后将相似性约束与目标分类器联系起来, 达到了利用源领域的大量数据和少量目标领域数据进行有效迁移学习的目的. 这些研究充分表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.

针对上述包含不确定类标签数据信息的机器学习问题, 本文从迁移学习的角度在两方面作出改进: 1) 将迁移学习框架纳入学习任务中, 因为迁移学习作为一种有效利用历史数据的新型学习策略, 同时考虑原有的、相关的实例或已训练的源领域相关模型参数(历史知识)来指导新实例的学习过程, 从而使得在目标领域的学习过程更为准确; 2) 拥有交叉的数据使得系统间的迁移更加容易[9], 即源领域和目标领域中的交叉部分可以提高迁移学习系统的泛化性能.

基于上述算法改进, 本文提出一种新的适用于不确定类标签数据学习的迁移支持向量机方法(TSVM-UL). 鉴于SVM在解决模式分类问题方面的高效性, 将其作为基本框架模型构造迁移支持向量机. 在此基础上, 引入迁移学习机制, 将源领域中的知识引入到目标领域分类学习中, 目标领域已标记样本和概率标记样本同时作为训练数据, 并考虑共享数据对分类决策的影响, 构造出新的目标函数分类器, 证明得到的新分类器的求解过程仍然是一个二次规划(QP)问题. 由于在进行迁移学习时同时利用了已标注信息和不确定信息, 在新场景不断涌现的当今世界, 所提出的方法具有较高的应用价值.

1 相关工作

关于如何利用不确定样本信息进行有效分类学习, 已有许多学者作了深入的研究[1-2,10-11], 其中文献[2]提出了一种可同时利用已标定数据和不确定数据的支持向量机方法(QQSVM). 下面首先对QQSVM方法作简要评述.

考虑线性可分的分类问题, 训练集为 $T=\{(\boldsymbol{x}_1, l_1),(\boldsymbol{x}_2,l_2),\cdots,(\boldsymbol{x}_m,l_m)\}$, 其中 $(\boldsymbol{x}_i)_{i,i+1,\cdots,m}\in X$, X 为特征空间. 类标签信息分为两部分: $(l_i)_{i,i+1,\cdots,n}=y_i\in\{-1,+1\}$ 为已标定的样本; $(l_i)_{n+1,n+2,\cdots,m}=p_i\in[0,1]$ 为不确定样本. 为了利用不确定样本信息, 文献[2]定义不确定样本为正的后验概率, 即

$$p_i=p(\boldsymbol{x}_i)=P(Y_i=1|X_i=\boldsymbol{x}_i). \tag{1}$$

则QQSVM的原始优化目标函数为

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i+\tilde{C}\sum_{i=n+1}^{m}(\xi_i^-+\xi_i^+).$$

$$\text{s.t. } y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)-1+\xi_i\geqslant 0,\ \xi_i\geqslant 0,\ i=1,2,\cdots,n;$$
$$z_i^++\xi_i^+\geqslant\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\geqslant z_i^--\xi_i^-;$$
$$\xi_i^-,\xi_i^+\geqslant 0,\ i=n+1,n+2,\cdots,m. \tag{2}$$

其中: C 和 $\tilde{C}$ 为正则化参数; ξ_i,ξ_i^-,ξ_i^+ 为松弛变量; z_i^+ 和 z_i^- 为依赖于 p_i 的类标签概率边界.

由式(2)可见, QQSVM仍然是一种传统的支持向量机方法, 可以转化为二次规划的形式进行求解. 虽然其在利用已标定信息之外还利用了不确定样本信息进行训练, 但只有在训练数据和待测数据满足独立同分布的前提下才能得到令人满意的效果.

2 适用于不确定类标签数据学习的迁移支持向量机

2.1 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. 鉴于支持向量机[12]在分类学习上的诸多优点, 选用L2-SVMs[13]作为所提出算法的基本模型. 对于支持向量机而言, 一类有用的知识可以描述为该支持向量分类机对应的分类超平面参数 $(\boldsymbol{w},b)$, 因此对于某源领域数据受训得到的支持向量机模型, 其对应的 $(\boldsymbol{w}_s,b_s)$ 既可以作为已有的可用源领域知识, 也可以作为相似领域差异的一种度量[14]. 为了使源领域知识和共享数据进行有效的知识迁移, 构造如下适用于不确定标签数据学习的迁移学习目标函数:

$$\min_{\boldsymbol{w}_t,b_t}\frac{1}{2}\|\boldsymbol{w}_t\|^2+\frac{C_0}{2}\sum_{i=1}^{l}\eta_i^2+\frac{C_1}{2}\sum_{i=1}^{m}\xi_i^2+$$
$$\frac{C_2}{2}\sum_{i=1}^{n}((\xi_i^-)^2+(\xi_i{}^+)^2)+\frac{\lambda}{2}\|\boldsymbol{w}_t-\boldsymbol{w}_s\|^2.$$
$$\text{s.t. } \boldsymbol{w}_t^{\mathrm{T}}\tilde{\boldsymbol{x}}_i+b_t=\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i+b_s-\eta_i,\ i=1,2,\cdots,l;$$
$$y_i(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j+b_t)=1-\xi_i,\ i=1,2,\cdots,m;$$
$$z_i^--\xi_i^-\leqslant\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j+b_t\leqslant z_i^++\xi_i^+,\ i=1,2,\cdots,n. \tag{3}$$

其中: $(\tilde{\boldsymbol{x}}_i,\tilde{y}_i)(i=1,2,\cdots,l)$ 为共享数据; $(\boldsymbol{x}_i,y_i)(i=1,2,\cdots,m)$ 为目标领域内的已标定样本; $(\boldsymbol{x}_i,y_i)(i=1,2,\cdots,n)$ 为目标领域内的不确定类标签的样本; $z_i^-,z_i^+(i=1,2,\cdots,n)$ 对应于每一个不确定数据的边界值; $\eta=[\eta_1,\eta_2,\cdots,\eta_n]^{\mathrm{T}}$ 为领域间共享数据的松弛向量; $\boldsymbol{\xi}=[\xi_1,\xi_2,\cdots,\xi_m]^{\mathrm{T}}$, $\boldsymbol{\xi}^-=[\xi_i{}^-,\xi_{i+1}{}^-,\cdots,\xi_n^-]^{\mathrm{T}}$ 和 $\boldsymbol{\xi}^+=[\xi_i{}^+,\xi_{i+1}{}^+,\cdots,\xi_n^+]^{\mathrm{T}}$ 分别为目标域数据中已标定数据和不确定数据的松弛向量; C_0,C_1 和 C_2 分别为共享数据、目标领域中已标定数据和不确定数据的正则化参数(惩罚误差程度).

对于上述优化目标函数, 为了进一步阐述其机理, 给出如下分析和说明:

1) 目标函数中的前4项分别表示目标领域数据

的结构风险项和对应于共享数据、目标领域中已标定数据、目标领域中不确定数据的经验风险项.

2) $\lambda(\|\boldsymbol{w}_t-\boldsymbol{w}_s\|^2)/2$ 部分为目标领域和源领域间的差异项, 其大小反映两个相似领域数据分布的差异程度, 数值越大表示分类器间的差异越大, 反之则差异越小. 惩罚的程度通过参数 λ 进行控制, 若 λ 较大, 则源领域和目标领域的分类超平面非常接近; 若 λ 较小, 则源领域和目标领域的分类超平面相对独立. 注意到, 此差异项不包含分类面的参数 b, 这是因为第 1 约束项中同时包含参数 $\boldsymbol{w}$ 和 b, 所以本优化目标函数可以使得源领域和目标领域分类面的参数 b 达到接近的效果, 而且这样构造所带来的另一个好处是简化了目标函数表达式求解的复杂性.

3) 第 1 约束项是为了保证对于源领域和目标领域所共享那部分数据的分类结果应该尽可能相同, 第 2 约束项表示目标领域中已知标签的样本应保证分类正确, 第 3 约束项表示在目标领域中不确定样本的分类结果应与其对应标签的概率相近.

4) 第 3 约束项中 z_i 为目标领域数据中不确定样本的类标签边界. 此处本文采取与文献 [2] 相同的方法, 使用

$$z_i^-=-\frac{1}{a}\ln\left(\frac{1}{p_i-\tau}-1\right),\ z_i^+=-\frac{1}{a}\ln\left(\frac{1}{p_i-\tau}-1\right)$$

进行计算. 其中: $a=\ln(1/\tau-1)$, $\tau=\varepsilon+\delta$, ε 和 δ 分别为目标领域数据中不确定数据的分类精度和置信度, p_i 为样本是正类的后验概率[2].

2.2　相关定理推导和证明

根据相关优化理论, 式 (3) 的原始问题可以转化为如下的对偶问题进行求解.

定理 1　TSVM-UL 原始优化问题 (3) 的对偶问题为

$$\begin{aligned}&\min_{\boldsymbol{\Gamma}}\ \frac{1}{2}\boldsymbol{\Gamma}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\Gamma}+\tilde{\boldsymbol{e}}\boldsymbol{\Gamma},\\&\text{s.t. }\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\Gamma}=0.\end{aligned}\tag{4}$$

其中

$$\boldsymbol{\Gamma}=[\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}^-,\boldsymbol{\gamma}^+]^{\mathrm{T}};$$

$$0\leqslant\boldsymbol{\beta}\leqslant[\underbrace{C_0,\cdots,C_0}_{l}\underbrace{C_1,\cdots,C_1}_{m}\underbrace{C_2,\cdots,C_2}_{n}\underbrace{C_2,\cdots,C_2}_{n}];$$

$$\boldsymbol{f}^{\mathrm{T}}=[\underbrace{1,\cdots,1}_{l},\boldsymbol{y}^{\mathrm{T}},\underbrace{1,\cdots,1}_{n},\underbrace{1,\cdots,1}_{n}];$$

$$\tilde{\boldsymbol{e}}=[\boldsymbol{h}-\boldsymbol{b_s},\boldsymbol{v},\boldsymbol{z}^--\boldsymbol{g},\boldsymbol{g}-\boldsymbol{z}^+];$$

$$h_i=\frac{\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}{1+\lambda},\ i=1,2,\cdots,l;$$

$$v_i=\frac{\lambda y_i\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i}{1+\lambda},\ i=1,2,\cdots,m;$$

$$g_i=\frac{\lambda\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i}{1+\lambda},\ i=1,2,\cdots,n;$$

$$\tilde{\boldsymbol{K}}=\frac{1}{2(1+\lambda)}\times\begin{bmatrix}\boldsymbol{K}_{1,1}&\boldsymbol{K}_{1,2}&\boldsymbol{K}_{1,3}&-\boldsymbol{K}_{1,3}\\\boldsymbol{K}_{1,2}^{\mathrm{T}}&\boldsymbol{K}_{2,2}&\boldsymbol{K}_{2,3}&-\boldsymbol{K}_{2,3}\\\boldsymbol{K}_{1,3}^{\mathrm{T}}&\boldsymbol{K}_{2,3}^{\mathrm{T}}&-\boldsymbol{K}_{3,3}&\boldsymbol{K}_{3,3}\\-\boldsymbol{K}_{1,3}^{\mathrm{T}}&-\boldsymbol{K}_{2,3}^{\mathrm{T}}&\boldsymbol{K}_{3,3}&-\boldsymbol{K}_{1,3}\end{bmatrix}_{(l+m+2n)\times(l+m+2n)}+\begin{bmatrix}\frac{\delta_{ij}}{C_0}&0&0&0\\0&\frac{\delta_{ij}}{C_1}&0&0\\0&0&\frac{\delta_{ij}}{C_2}&0\\0&0&0&v\frac{\delta_{ij}}{C_2}\end{bmatrix}_{(l+m+2n)\times(l+m+2n)};$$

$$\boldsymbol{K}_{1,1}=(k(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j))_{i,j=1,2,\cdots,l};$$

$$\boldsymbol{K}_{2,2}=(y_iy_jk(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,2,\cdots,m};$$

$$\boldsymbol{K}_{3,3}=(k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,2,\cdots,n};$$

$$\boldsymbol{K}_{1,2}=(k(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_j)y_j)_{i=1,2,\cdots,l,\ j=1,2,\cdots,m};$$

$$\boldsymbol{K}_{1,3}=(k(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_j))_{i=1,2,\cdots,l,\ j=1,2,\cdots,n};$$

$$\boldsymbol{K}_{2,3}=(y_ik(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i=1,2,\cdots,m,\ j=1,2,\cdots,n};$$

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

证明　最小值问题 (3) 的拉格朗日函数为

$$\begin{aligned}&L(\boldsymbol{w}_t,b_t,\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}^-,\boldsymbol{\gamma}^+)=\\&\frac{1}{2}\|\boldsymbol{w}_t\|^2+\frac{C_0}{2}\sum_{i=1}^{l}\eta_i^2+\frac{C_1}{2}\sum_{i=1}^{m}\xi_i^2+\frac{C_2}{2}\sum_{i=1}^{n}((\xi_i{}^-)^2+(\xi_i{}^+)^2)+\frac{\lambda}{2}\|\boldsymbol{w}_t-\boldsymbol{w}_s\|^2-\\&\sum_{i=1}^{l}\alpha_i(\boldsymbol{w}_t^{\mathrm{T}}\tilde{\boldsymbol{x}}_i+b_t-\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i-b_s+\eta_i)-\\&\sum_{i=1}^{m}\beta_i(y_i(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_i+b_t)-1+\xi_i)-\\&\sum_{i=1}^{n}\gamma_i^-((\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j+b_t)-z_i^-+\xi_i^-)-\\&\sum_{i=1}^{n}\gamma_i^+(-(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j+b_t)+z_i^++\xi_i^+),\end{aligned}\tag{5}$$

其中 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)$、$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_m)$、$\boldsymbol{\gamma}^-=(\gamma_1^-,\gamma_2^-,\cdots,\gamma_n^-)$ 和 $\boldsymbol{\gamma}^+=(\gamma_1^+,\gamma_2^+,\cdots,\gamma_n^+)$ 为拉格朗日乘子.

根据 Karush-Kuhn-Tucker (KKT)[15]条件, 有

$$\frac{\partial L}{\partial\eta_i}=0\Rightarrow C_0\eta_i=\alpha_i,\tag{6}$$

$$\frac{\partial L}{\partial\xi_i}=0\Rightarrow C_0\xi_i=\beta_i,\tag{7}$$

$$\frac{\partial L}{\partial\xi_i^-}=0\Rightarrow C_2\xi_i{}^-=\gamma_i{}^-,\tag{8}$$

$$\frac{\partial L}{\partial {\xi_i}^+}=0\Rightarrow C_2{\xi_i}^+={\gamma_i}^+, \tag{9}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_t}=0\Rightarrow \boldsymbol{w}_t+\lambda(\boldsymbol{w}_t-\boldsymbol{w}_s)=\sum_{i=1}^{l}\alpha_i\tilde{\boldsymbol{x}}_i+\sum_{i=1}^{m}\beta_i y_i\boldsymbol{x}_i+\sum_{i=1}^{n}(\gamma_i^- -{\gamma_i}^+)\boldsymbol{x}_i, \tag{10}$$

$$\frac{\partial L}{\partial b_t}=0\Rightarrow \sum_{i=1}^{l}\alpha_i+\sum_{i=1}^{m}\beta_i y_i+\sum_{i=1}^{n}(\gamma_i^- -{\gamma_i}^+)=0. \tag{11}$$

将式(6)～(11)代回(5), 化简后可得到对偶问题为

$$\min_{\boldsymbol{\Gamma}}\frac{1}{2}\boldsymbol{\Gamma}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\Gamma}+\tilde{\boldsymbol{e}}\boldsymbol{\Gamma},\quad \text{s.t. } \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\Gamma}=0. \tag{12}$$

其中

$$\boldsymbol{\Gamma}=[\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}^-,\boldsymbol{\gamma}^+]^{\mathrm{T}};$$

$$0\leqslant\boldsymbol{\beta}\leqslant[\underbrace{C_0,\cdots,C_0}_{l}\underbrace{C_1,\cdots,C_1}_{m}\underbrace{C_2,\cdots,C_2}_{n}\underbrace{C_2,\cdots,C_2}_{n}];$$

$$\boldsymbol{f}^{\mathrm{T}}=[\underbrace{1,\cdots,1}_{l},\boldsymbol{y}^{\mathrm{T}},\underbrace{1,\cdots,1}_{n},\underbrace{1,\cdots,1}_{n}];$$

$$\tilde{\boldsymbol{e}}=[\boldsymbol{h}-\boldsymbol{b_s},\boldsymbol{v},\boldsymbol{z}^- -\boldsymbol{g},\boldsymbol{g}-\boldsymbol{z}^+];$$

$$h_i=\frac{\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}{1+\lambda},\ i=1,2,\cdots,l;$$

$$v_i=\frac{\lambda y_i\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i}{1+\lambda},\ i=1,2,\cdots,m;$$

$$g_i=\frac{\lambda\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i}{1+\lambda},\ i=1,2,\cdots,n;$$

$$\tilde{\boldsymbol{K}}=\frac{1}{2(1+\lambda)}\times\begin{bmatrix}\boldsymbol{K}_{1,1}&\boldsymbol{K}_{1,2}&\boldsymbol{K}_{1,3}&-\boldsymbol{K}_{1,3}\\ \boldsymbol{K}_{1,2}^{\mathrm{T}}&\boldsymbol{K}_{2,2}&\boldsymbol{K}_{2,3}&-\boldsymbol{K}_{2,3}\\ \boldsymbol{K}_{1,3}^{\mathrm{T}}&\boldsymbol{K}_{2,3}^{\mathrm{T}}&-\boldsymbol{K}_{3,3}&\boldsymbol{K}_{3,3}\\ -\boldsymbol{K}_{1,3}^{\mathrm{T}}&-\boldsymbol{K}_{2,3}^{\mathrm{T}}&\boldsymbol{K}_{3,3}&-\boldsymbol{K}_{1,3}\end{bmatrix}_{(l+m+2n)\times(l+m+2n)}+\begin{bmatrix}\delta_{ij}/C_0&0&0&0\\ 0&\delta_{ij}/C_1&0&0\\ 0&0&\delta_{ij}/C_2&0\\ 0&0&0&\delta_{ij}/C_2\end{bmatrix}_{(l+m+2n)\times(l+m+2n)};$$

$$\boldsymbol{K}_{1,1}=(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j)_{i,j=1,2,\cdots,l};$$

$$\boldsymbol{K}_{2,2}=(y_iy_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)_{i,j=1,2,\cdots,m};$$

$$\boldsymbol{K}_{3,3}=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)_{i,j=1,2,\cdots,n};$$

$$\boldsymbol{K}_{1,2}=(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{x}_jy_j)_{i=1,2,\cdots,l,\ j=1,2,\cdots,m};$$

$$\boldsymbol{K}_{1,3}=(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{x}_j)_{i=1,2,\cdots,l,\ j=1,2,\cdots,n};$$

$$\boldsymbol{K}_{2,3}=(y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)_{i=1,2,\cdots,m,\ j=1,2,\cdots,n};$$

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

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

$$\boldsymbol{K}_{1,1}=(k(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j))_{i,j=1,2,\cdots,l},$$

$$\boldsymbol{K}_{2,2}=(y_iy_jk(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,2,\cdots,m},$$

$$\boldsymbol{K}_{3,3}=(k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1,2,\cdots,n},$$

$$\boldsymbol{K}_{1,2}=(k(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_j)y_j)_{i=1,2,\cdots,l,\ j=1,2,\cdots,m},$$

$$\boldsymbol{K}_{1,3}=(k(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_j))_{i=1,2,\cdots,l,\ j=1,2,\cdots,n},$$

$$\boldsymbol{K}_{2,3}=(y_ik(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i=1,2,\cdots,m,\ j=1,2,\cdots,n}.$$

式(12)核化后可得到(4). □

与SVM问题求解类似, TSVM-UL的原始优化问题(3)对于 $\boldsymbol{w}$ 和 b 的解存在且唯一, 为全局最优解, 可以表示为

$$\boldsymbol{w}_t^*=\frac{1}{1+\lambda}\Big(\sum_{i=1}^{l}\tilde{\alpha}_i\varphi(\tilde{\boldsymbol{x}}_i)+\sum_{i=1}^{m}\tilde{\beta}_iy_i\varphi(\boldsymbol{x}_i)+\sum_{i=1}^{n}(\tilde{\gamma}_i^- -\tilde{\gamma}_i^+)\varphi(\boldsymbol{x}_i)+\lambda\boldsymbol{w}_s\Big), \tag{13}$$

$$b_t{}^*=y_j-\frac{1}{1+\lambda}\boldsymbol{w}_s^{\mathrm{T}}\varphi(\boldsymbol{x}_j)-\frac{1}{1+\lambda}\sum_{i=1}^{l}\tilde{\alpha}_ik(\tilde{\boldsymbol{x}}_i^{\mathrm{T}},\boldsymbol{x}_j)-\frac{1}{1+\lambda}\sum_{i=1}^{m}\tilde{\beta}_iy_ik(\boldsymbol{x}_i^{\mathrm{T}},\boldsymbol{x}_j)-\frac{1}{1+\lambda}\sum_{j=1}^{n}(\tilde{\gamma}_j^- -\tilde{\gamma}_j^+)k(\boldsymbol{x}_i^{\mathrm{T}},\boldsymbol{x}_j), \tag{14}$$

其中 $(\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}}^-,\tilde{\boldsymbol{\gamma}}^+)$ 为式(12)的解.

注意到, 式(13)和(14)给出的最优解同时包含了源领域和历史领域的信息. 如: ${\boldsymbol{w}_t}^*$ 中 $\frac{\lambda}{1+\lambda}\boldsymbol{w}_s$ 部分为从源领域中学习得到的知识; $\frac{1}{1+\lambda}\sum_{i=1}^{l}\tilde{\alpha}_i\varphi(\tilde{\boldsymbol{x}}_i)$ 部分为从共享数据中学习得到的知识; $\frac{1}{1+\lambda}\sum_{i=1}^{m}\tilde{\beta}_iy_i\varphi(\boldsymbol{x}_i)$ 和 $\frac{1}{1+\lambda}\sum_{i=1}^{n}(\tilde{\gamma}_i^- -\tilde{\gamma}_i^+)\varphi(\boldsymbol{x}_i)$ 部分为从目标领域中已标定数据和不确定数据中获取的知识.

2.3 TSVM-UL算法流程

由第2.2节的推导和分析, 可以得到TSVM-UL方法的具体步骤如下.

输入: N 个有标号的源领域样本 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$, 目标领域样本包含 m 个已标定的样本 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$, n 个不确定样本 $\{(\boldsymbol{x}_i,p_i)\}_{i=1}^n$, 源领域和目标领域共有的 l 个有标号数据 $\{(\tilde{\boldsymbol{x}}_i,\tilde{y}_i)\}_{i=1}^l$;

输出: 目标领域的分类决策函数 $f(\boldsymbol{x})$.

源领域知识获取阶段.

Step 1: 设置核函数带宽 σ_s 和正则参数 C_s;

Step 2: 利用 SVM 得到源领域数据的分类模型;

Step 3: 利用二次规划原理求解源领域拉格朗日向量 $\boldsymbol{\alpha}^s$ 和偏移量 b_s.

目标领域迁移学习阶段.

Step 4: 根据 $\{(\boldsymbol{x}_i, p_i)\}_{i=1}^n$ 计算当前领域数据的不确定样本的边界 $z_i^+,\ z_i^-,\ i=1,2,\cdots,n$;

Step 5: 根据 $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^l$、$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ 和 $\{(\boldsymbol{x}_i, p_i)\}_{i=1}^n$ 计算核矩阵, $\boldsymbol{K}_{1,1}, \boldsymbol{K}_{1,2}, \boldsymbol{K}_{2,2}, \boldsymbol{K}_{1,3}, \boldsymbol{K}_{2,3}, \boldsymbol{K}_{3,3}$;

Step 6: 根据定理 1 构造矩阵 $\tilde{\boldsymbol{K}}$, 求解拉格朗日向量 $\boldsymbol{\Gamma}$.

目标领域决策函数生成阶段.

Step 7: 根据式 (13) 计算决策超平面法向量 $\boldsymbol{w}_t$;

Step 8: 根据式 (14) 计算偏移量 b_t;

Step 9: 输出分类决策函数 $f(\boldsymbol{x}) = {\boldsymbol{w}_t}^{\mathrm{T}}\varphi(\boldsymbol{x}) + b_t$.

3　实验分析

为了表明 TSVM-UL 方法在不同领域数据集分类学习问题上的有效性, 在不同类型的真实数据集上进行实验: 1) 人脸图像分类数据集 PIE[16]; 2) 跨领域文本分类的真实数据集 20Newsgroup[7]; 3) 3 个常用的 UCI 数据集[17].

实验中, 主要引入 SVM[18]、LMPROJ[6]、LWE[7]、TrSVM[8]和 QQSVM[2]5 种算法进行比较: SVM 用来验证所提出方法在迁移学习问题上与传统基于独立同分布假设的支持向量机分类方法的优势所在; TrSVM、LWE 和 LMPROJ 均为迁移学习分类方法, 用来表明在目标领域类标签保护的前提下, 所提出方法与其他领域自适应方法具有可比性; QQSVM 为基于传统支持向量机框架的可同时利用已标定样本和不确定样本的分类方法, 用来表明所提出方法在优化目标函数中通过融合迁移学习机制和共享数据而带来的性能提升.

将本文方法与其他方法进行学习能力比较时, 以目标域数据分类的精度作为评价指标, 即

$$\text{Accuracy} = \frac{|\{x|x_t \in D_t \bigcap f(x_i) = y_t\}|}{|\{x|x_t \in D_t\}|}.$$

其中: D_t 为目标领域数据集, y_t 为 x_t 的真实标签类别, $f(x)$ 为使用学习所得分类器对 x_t 进行分类得到的结果.

如无特别说明, 则所有实验均通过网格搜索的方式确定优化的实验参数. 采用高斯核函数 $k(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\dfrac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right)$, 核函数中 $2\sigma^2$ 的选择以源领域样本平均 2 范数的平方 s 为基准, 在网格 $\{s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s, 16s, 32s, 64s\}$ 中搜索直至最优. TSVM-UL 的正则化参数 C_0、C_1 和 C_2 在网格 $\{2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$ 中搜索直至最优. 平衡参数 λ 在区间 $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$ 中搜索直至最优. 不失一般性, 为了模拟真实场景, 将目标领域数据中 50% 的数据设定为不确定数据, 不确定信息的概率由 Platt 标定算法[19]计算, 并加入 10% 的高斯噪声, 这是因为真实场景中所得概率信息一般总是有误差的. 实验中每个学习任务重复进行 10 次, 取其平均精度作为算法实验结果. 所有实验均在 Intel Core 2, 2.0 GHz 主频, 3G RAM, Windows XP 系统, Matlab 2009a 平台实现. SVM 算法由 Libsvm[20]软件实现, 其他算法均在 Matlab R2009A 环境下实现.

对于所有数据集, 源领域和目标领域数据均具有标签信息, 但目标域标签信息仅用于学习方法分类性能的客观量化评价.

3.1　PIE 人脸数据集

PIE 数据库包含 68 个人的 41 368 幅人脸灰度图像. 首先构造迁移学习任务, 随机选取 2 人, 每人 170 幅人脸图像进行实验. 原始人脸数据样本构成目标领域数据集, 再加入不同程度的服从正态分布的噪声构成源领域数据集和领域间共享数据, 共享数据占总数据量的 5%. 由于每个像素点灰度值的区间为 [0, 255], 在构造源领域数据集时需要注意, 若加入噪声后某个像素点的灰度值超出上下界时, 则应将该像素点的灰度值设置为 255 或 0. 图 1 为基于 PIE 人脸数据库构造的源领域和目标领域样本, 其中图 1(a) 的噪声服从正态分布 $\mu = 50,\ \delta = 10$. 表 1 为 6 类算法在 PIE 数据集上的分类精度.

表 1　6 类算法在 PIE 数据集上的分类精度

学习	领域说明		算法精度/%					
任务	源领域数据	共享数据	SVM	TrSVM	LWE	LMPROJ	QQSVM	TSVM-UL
T1	$\mu=0, \delta=10$	$\mu=0, \delta=30$	87.81	95.90	95.13	92.91	96.91	**97.78**
T2	$\mu=0, \delta=50$	$\mu=0, \delta=30$	83.47	94.28	93.58	91.78	93.51	**96.89**
T3	$\mu=25, \delta=10$	$\mu=25, \delta=30$	79.50	90.60	91.00	90.20	90.33	**95.65**
T4	$\mu=25, \delta=50$	$\mu=25, \delta=30$	76.80	87.85	88.88	87.22	85.33	**92.32**
T5	$\mu=50, \delta=10$	$\mu=50, \delta=30$	71.51	85.60	83.12	83.32	83.10	**90.12**
T6	$\mu=50, \delta=50$	$\mu=50, \delta=30$	71.11	73.64	79.81	79.61	78.40	**89.56**

(a) 加服从正态分布噪声的源领域人脸样本

(b) 未加噪声的目标领域人脸样本

图 1 基于 PIE 人脸数据库构造的源领域和目标领域样本

3.2 跨领域文本数据集 20Newsgroups

抽取常用的文本分类数据集20Newsgroups评价所提出方法与相关方法的迁移学习性能. 对于20Newsgroups文本数据集, 分别从顶层大类中抽取6个大类以构建学习数据集, 其中每2个大类分别选作正类和负类, 数据基于子类进行分割, 认为不同的子类属于不同的领域. 本文所构造的20Newsgroups数据集各学习任务的详细信息如表2和表3所示, 各学习任务的相应实验结果如表4所示.

表 2 UCI 数据集描述

大类	小类编号	小类	样本数
Comp	1	comp.sys.ibm.pc.hardware	979
	2	comp.windows.x	982
	3	comp.sys.mac.hardware	958
Rec	4	sport.hockey	997
	5	motorcycles	993
	6	autos	987
Sci	7	crypt	989
	8	med	987
	9	space	985
Talk	10	politics.guns	909
	11	politics.mideast	940
	12	politics.misc	774

表 3 取自 20Newsgroups 的源领域和目标领域信息

学习任务	数据集	源领域数据 ([子类号, 占子类样本总量的百分比, 样本数])			目标领域数据 ([子类号, 占子类样本总量的百分比, 样本数])		
		正类	共享数据	负类	正类	共享数据	负类
T7	Comp vs. Sci	[1, 80%, 784]	[2, 9, 20%, 393]	[7, 80%, 792]	[3, 80%, 766]	[2, 9, 20%, 393]	[8, 80%, 790]
T8	Rec vs. Talk	[5, 80%, 794]	[4, 10, 20%, 381]	[11, 80%, 752]	[6, 80%, 790]	[4, 10, 20%, 381]	[12, 80%, 618]
T9	Rec vs. Sci	[5.80%, 794]	[4, 9, 20%, 396]	[7, 80%, 792]	[6, 80%, 790]	[4, 9, 20%, 396]	[8, 80%, 790]
T10	Sci vs. Talk	[7, 80%, 792]	[9, 10, 20%, 379]	[11, 80%, 752]	[8, 80%, 790]	[9, 10, 20%, 379]	[12, 80%, 618]
T11	Comp vs. Rec	[1, 80%, 784]	[2, 4, 20%, 395]	[5, 80%, 794]	[3, 80%, 766]	[2, 4, 20%, 395]	[6.80%, 790]
T12	Comp vs. Talk	[1, 80%, 784]	[2, 10, 20%, 378]	[11.80%, 752]	[3, 80%, 766]	[2, 10, 20%, 378]	[12, 80%, 618]

表 4 6 类算法在 20Newsgroups 数据集上的分类精度 %

学习任务	SVM	Tr-SVM	LWE	LMPROJ	QQSVM	TSVM-UL
T7	72.26	71.23	83.11	82.12	74.11	**84.21**
T8	70.18	79.66	76.60	79.30	72.20	**82.35**
T9	78.35	85.14	88.41	86.68	81.80	**88.90**
T10	76.00	85.09	80.68	84.21	80.09	**86.31**
T11	83.67	84.34	85.40	86.78	86.24	**90.90**
T12	90.90	93.56	94.43	95.43	90.74	**96.35**

3.3 UCI 数据集

所提出方法针对包含不确定类标签信息的学习问题, 因此不确定类标签数据在数据集中所占的比重是影响算法性能的重要因素之一. 本节选取Ionosphere (126个正类样本, 225个负类样本)、Sonar (97个正类样本, 111个负类样本) 和Spanbase (1 813个正类样本, 2 788个负类样本) 3个常用的UCI数据集在不同不确定信息比例下对所提出方法作进一步评价. 各数据集的样本组成和实验结果如表5所示.

3.4 实验结果分析

根据6类方法在真实数据集上的实验结果可以得到如下结论:

1) 传统的SVM无法将源领域的知识有效地迁移至目标领域以帮助学习, 因此在所有真实数据集上的学习任务所得到的分类精度均低于其余几类方法.

2) 由于充分考虑了共享数据和由源领域中获取的知识, TSVM-UL方法在所有迁移学习任务上的分类精度均优于其他几类领域自适应方法.

3) 目标领域中不确定信息所占比例会影响TSVM-UL方法的精度, 比例越大精确度越低.

表 5 TSVM-UL 算法在 UCI 数据集上的分类精度

学习任务	数据集	源领域数据 ([占样本总量百分比, 样本数])			目标领域数据 ([占样本总量百分比, 样本数])			比例/%	分类精度/%
		正类	共享域	负类	正类	共享域	负类		
T13	Ionosphere	[80%, 100]	[10%, 21]	[50%, 112]	[80%, 100]	[10%, 21]	[50%, 112]	30%	**95.35**
T14		[80%, 100]	[10%, 21]	[50%, 112]	[80%, 100]	[10%, 21]	[50%, 112]	50%	93.21
T15	Sonar	[60%, 58]	[10%, 11]	[50%, 56]	[60%, 58]	[10%, 11]	[50%, 56]	30%	**86.89**
T16		[60%, 58]	[10%, 11]	[50%, 56]	[60%, 58]	[10%, 11]	[50%, 56]	50%	85.02
T17	Spanbase	[30%, 543]	[10%, 110]	[20%, 557]	[30%, 543]	[10%, 110]	[20%, 557]	30%	**89.77**
T18		[30%, 543]	[10%, 110]	[20%, 557]	[30%, 543]	[10%, 110]	[20%, 557]	50%	87.67

4) 与 QQSVM 的比较表明, 在已标定数据较少的情况下, 虽然 QQSVM 使用不确定信息来弥补一定的分类精度下降造成的损失, 但仍然依赖于已标定数据和不确定信息的准确程度, 而 TSVM-UL 方法借鉴了目标领域中已标定信息和不确定信息同时利用源领域中所学知识和共享数据来帮助分类, 因此依然可以达到可利用的精度.

4 结　论

待研究领域包含概率信息的模式分类问题, 传统方法虽然利用概率信息帮助训练, 但忽视了与目标领域相关的源领域所包含的知识, 从而在具体的模式分类问题上存在一定的局限性. 鉴于此, 本文从迁移学习的角度出发, 将源领域的知识和共享数据同时纳入目标决策函数的构造, 并结合结构风险最小化模型, 提出了一种适用于不确定类标签信息学习的 TSVM-UL 方法. 实验结果表明, 所提出方法不仅具备 SVM 算法快速、容易实现的优点, 而且兼顾了相似领域的知识迁移学习. 下一步的研究方向是在多分类问题和大数据集两个方面进行深入探讨.

参考文献(References)

[1] Stempfel G, Ralaivola L. Learning SVMs from sloppily labeled data[C]. Proc of Artificial Neural Networks. Limassol: Springer, 2009: 884-893.

[2] Emilie N, Remi F, Carole L, et al. Handling Uncertains in SVM Classification[C]. IEEE Statistical Signal Processing Workshop(SSP). Nice: IEEE, 2011: 757-760.

[3] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[4] 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.)

[5] 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.

[6] Quanz B, Huan J. Large margin transductive transfer learning[C]. Proc of the 18th ACM Conf on Information and Knowledge Management. New York: ACM, 2009: 1327-1336.

[7] Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping[C]. Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 283-291.

[8] 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]. J of Computer Research and Development, 2011, 48(10): 1823-1830.)

[9] 周涛. 个性化推荐的十大挑战[J]. 中国计算机学会通讯, 2012, 8(7): 48-56.
(Zhou T. The ten major challenges of personalized recommendation[J]. Communications of the CCF, 2012, 8(7): 48-56.)

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

[11] Stolpe M, Morik K. Learning from label proportions by optimizing cluster model selection[C]. Proc of ECML PKDD. Berlin: Heidelberg, 2011: 349-364.

[12] Vapnik V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995: 123-167.

[13] Tsang I W, Kwok J T, Zurada J M. Generalized core vector machines[J]. IEEE Trans on Neural Network, 2006, 17(5): 1126-1140.

[14] Guillermo L G, Lucas C U, Alejandro C H, et al. Solving nonstationary classification problems with coupled support vector machines[J]. IEEE Trans on Knowledge on Neural Network, 2011, 22(1): 37-51.

[15] Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem[C]. Proc of Conf on Learning Theory. Amsterdam, 2001: 416-426.

[16] He X F, Cai D, Partha N. Laplacian score for feature selection[C]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2006: 507-514.

[17] Asuncion A, Newman D J. UCI machine learning repository[EB/OL]. (2008-08-10)[2008-11-01]. http://archive.ics.uci.edu/ml/.

[18] Vapnik V. Statistical learning theory[M]. John Wiley and Sons, 1998.

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

[20] Chang C C, Lin C J. LIBSVM: A library for support vector machines[EB/OL]. (2008-10-31)[2008-11-04]. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

(责任编辑: 郑晓蕾)