共享隐空间迁移 SVM

董爱美 [1, 2] 王士同 [1]

摘 要 在机器学习中, 迁移学习被证明能有效使用一个领域信息提高另一个领域中受训模型的分类精度. 迁移学习总是假设相关领域间共享某些隐含因素, 但在当前的迁移学习方法中, 该部分隐含因素依然未得到充分探讨. 本研究引入低维共享隐空间的迁移学习方法, 基于经典支持向量机 (Support vector machine, SVM) 分类模型得到融入共享隐空间的迁移支持向量机, 该模型较以往相关方法能更好地利用隐空间这一有效信息, 从而提高所得分类器的泛化性能. 相关实验结果亦验证了所提方法的有效性.

关键词 迁移学习, 大间隔分类器, 隐空间, 支持向量机

引用格式 董爱美, 王士同. 共享隐空间迁移 SVM. 自动化学报, 2014, **40**(10): 2276−2287

DOI 10.3724/SP.J.1004.2014.02276

A Shared Latent Subspace Transfer Learning Algorithm Using SVM

DONG Ai-Mei[1, 2] WANG Shi-Tong[1]

Abstract In machine learning, transfer learning is proved to be able to efficiently use the information of one domain to enhance the classification accuracy in another domain. Transfer learning always assumes that the related domains share some correlated latent factors. But the existing transfer learning methods cannot make full use of these latent factors. So a low dimensionality shared latent subspace transfer learning method is proposed. Specifically, a shared latent subspace transfer support vector machine model is proposed in contrast to the classical support vector machine (SVM) model. Compared with traditional transfer learning methods, the proposed method can make better use of the latent subspace, therefore it can enhance the transfer classification performance. Experimental results confirm the efficiency of the proposed method.

Key words Transfer learning, large margin classifier, latent space, support vector machine (SVM)

Citation Dong Ai-Mei, Wang Shi-Tong. A shared latent subspace transfer learning algorithm using SVM. *Acta Automatica Sinica*, 2014, **40**(10): 2276−2287

传统机器学习的一个基本假设是训练数据和测试数据分布相同, 其学习方法在监督学习、半监督学习和无监督学习中已经取得很大成功. 但目前机器学习领域出现了一些新应用: 与测试数据分布相同的训练数据数量非常少, 采用传统的机器学习方法不足以训练一个可靠的学习模型. 但在相关领域中可能有大量的、相近分布的训练数据, 在新的领域如何有效利用相关领域的大量数据信息是非常重要的, 这就是迁移学习所探讨的工作, 即迁移学习旨在将从一个环境中学到的知识用来帮助新环境中的学习任务, 其放松了训练数据和测试数据必须满足相同分布的限制, 且能有效弥补当前学习场景有监督数据不足的缺陷. 由于具有良好的灵活性和广泛的适应性等特点, 迁移学习已经成为机器学习领域的研究热点[1−3]. 近几年, 众多学者从不同角度提出很多迁移学习方法. 从领域数据使用的角度, 迁移学习大致分为基于实例的迁移学习方法和基于特征的迁移学习方法[4]. 基于实例的迁移学习方法对领域间数据非常相近的迁移问题十分有效. 当领域间数据差别较大时, 该方法往往很难在实例层面找到领域间数据的共性信息. 基于特征的迁移学习方法认为尽管领域间数据存在较大差异, 但会共享部分的特征集合.

基于特征的迁移学习方法从特征结构上建立领域间数据的联系, 通过特征的变换重构发现潜在的共同子空间, 利用特征层面上的交集进行迁移学习. 过去几十年中, 子空间学习[5−11] 在机器学习和模式识别领域得到广泛而深入地探讨. 2005 年, Ando 等[5] 采用结构学习方法学习多任务中共同特征表示,

收稿日期 2013-05-07 录用日期 2014-04-10
Manuscript received May 7, 2013; accepted April 10, 2014
国家自然科学基金 (61170122, 61202311), 江苏省自然科学基金 (BK 2012552), 山东省高等学校科技计划项目 (J14LN05) 资助
Supported by National Natural Science Foundation of China (61170122, 61202311), Natural Science Foundation of Jiangsu Province (BK2012552), and the Project of Shandong Province Higher Educational Science and Technology Program (J14LN05)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 江南大学数字媒体学院 无锡 214122 2. 齐鲁工业大学信息学院 济南 250353
1. School of Digital Media, Jiangnan University, Wuxi 214122 2. School of Information, Qilu University of Technology, Jinan 250353

提出了交替结构优化 (Alternate structure optimization, ASO) 算法; 2008 年, Zheng 等[6] 提出了多任务学习隐空间算法, 并应用到多设备室内信号定位模型中; Pan 等[7] 采用降维方法实现迁移, 首先学习一个低维的隐特征空间, 然后通过在此空间中增强源域和目标域数据的均值中心一致性, 减小两个领域的差异, 提出了基于流形结构的 MMDE (Maximum mean discrepancy embedding) 算法; 2010 年, Si 等[8] 通过最小化目标域和辅助领域数据分布之间的 Bregman 散度, 提出了迁移子空间学习框架; 2012 年, Shao 等[9] 使用低秩表示学习子空间技术提出了低秩迁移子空间学习 (Low-rank transfer subspace learning, LRTSL) 算法; Yang 等[10] 针对目标域中稀疏实例问题提出了迁移稀疏子空间学习 (Transfer sparse subspace learning, TSSL) 算法; 2013 年, Gupta 等[11] 用非负矩阵分解方法提出了正则化的共享子空间学习框架, 并指出该方法适用于迁移学习, 能在一定程度上提高学习性能.

尽管上述基于特征的子空间迁移学习方法取得一定成效, 但由于它们拆分了领域间数据原始特征空间, 仅选择部分共同特征进行迁移学习, 忽略了领域原始特征空间的判别信息. 为此, 本文从基于特征的迁移技术和特征变换的角度出发, 充分考虑领域原始特征空间和领域间共享低维隐空间的约束, 提出一种基于大间隔支持向量机的领域共享隐空间的特征迁移方法. 该方法引入一个特征变换参数矩阵作为隐变量, 利用该隐变量源域和目标域数据被映射到一个公共的低维子特征空间上. 进一步地, 基于原始特征空间和公共低维子特征隐空间构造联合决策函数; 基于间隔最大化分类框架, 把领域原始特征空间和领域间公共低维子特征隐空间同时嵌入到支持向量机 (Support vector machine, SVM) 的训练中, 从而学习到一个在目标域中泛化性能更好的分类器.

本文内容包括: 第 1 节介绍提出的基于 SVM 的隐空间迁移学习方法原理及算法; 第 2 节给出实验结果及分析; 第 3 节进行总结.

1 基于 SVM 的隐空间迁移学习机

1.1 相关概念与问题描述

为便于讨论相关迁移学习算法，首先对一些相关概念和符号进行统一说明.

1) 领域 (Domain): 领域常用 $D = \{(\boldsymbol{x}_i, y_i) \in X \times Y : i = 1, 2, \cdots, n\}$ 表示. 其中 X 表示领域实例的特征空间, Y 表示实例的标签空间, $(\boldsymbol{x}_i, y_i)$ 表示领域 D 的一个实例.

2) $P(D) = P(X, Y)$ 表示领域 D 中数据的分布. 有两个领域 D_s 和 D_t, 若 $P(D_s) \neq P(D_t)$, 则 D_s 和 D_t 是两个不同的领域.

对于模式分类问题, 设数据领域为 D, 领域数据概率分布为 $P(D) = P(X, Y)$, 分类器为一个映射函数 $f(\boldsymbol{x}) : X \to Y$, 其将实例 $\boldsymbol{x} \in X$ 映射为相应的类标签 $y \in Y$. 线性分类器形式为 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$, 其中 $\boldsymbol{w}$ 表示实例 $\boldsymbol{x}$ 所在空间的分类面参数.

基于以上定义, 本文研究迁移学习中的源域和目标域, 可表示为

源域:

$$D_s \left\{ (\boldsymbol{x}_i^s, y_i^s) \in X^s \times Y^s : i = 1, 2, \cdots, n_s \right\}$$

目标域:

$$D_t = \left\{ (\boldsymbol{x}_j^t, y_j^t) \in X^t \times Y^t : j = 1, 2, \cdots, n_t \right\}$$

其中, $n_s \gg n_t$, 上标表示实例所在领域, 下标表示实例在所在领域中的索引位置. 本文研究基于如下假设:

1) 所研究对象仅包含一个源域 D_s 和一个目标域 D_t.

2) 领域间数据分布不同但存在相似性, 两个不同领域通过特征变换重构共享一个低维的子特征隐空间.

3) 源域中有大量数据且带标签信息; 目标域有少量带标签的数据. 学习任务是充分利用源域信息为目标域训练一个具有更好泛化性能的分类器.

1.2 基于 SVM 的共享隐空间迁移学习机原理和目标函数

假设给定源域 D_s 和目标域 D_t, 根据传统监督学习方法, 基于正则经验风险最小化原则, 通过最小化式 (1) 所示函数来优化得到源域和目标域的分类决策函数:

$$\begin{aligned} R(\{f_s\}\{f_t\}) = {} & \lambda(\Omega(f_s) + \Omega(f_t)) + \\ & C_s \sum_{i=1}^{n_s} L(f_s(\boldsymbol{x}_i^s), y_i^s) + C_t \sum_{j=1}^{n_t} L(f_t(\boldsymbol{x}_j^t), y_j^t) \end{aligned} \tag{1}$$

其中, $n_s \gg n_t$, f_s 是源域分类决策函数, f_t 是目标域分类决策函数, $\Omega(f_s)$ 和 $\Omega(f_t)$ 是正则化项, $\lambda > 0$ 是正则化参数; 第 3 项和第 4 项是源域和目标域的经验风险项.

对源域来说, 因其数据量比较多, 根据式 (1) 足以训练性能较好的分类决策函数; 而对目标域来说, 其数据量非常少, 根据式 (1) 得到的分类决策函数性能很低. 因此本研究目的在于充分借助大量源域信息学习到源域和目标域数据的共同特征表示, 用

其减小领域间的差异, 利用特征结构信息对迁移过程进行约束. 具体来说本文从特征变换角度出发, 通过对源域和目标域数据特征空间实施变换来寻找一个低维的公共隐空间, 在此隐空间尽可能利用源域和目标域的共性信息, 结合新的优化方式从而达到迁移学习之目的. 基于上述思路, 构造源域和目标域的新分类决策函数由两部分组成: 领域原始数据空间部分和领域间共享的低维子特征隐空间部分, 具体表示为

$$f_s(\boldsymbol{x}) = \boldsymbol{w}_s{}^{\mathrm{T}}\boldsymbol{x}^s + \boldsymbol{v}_s^{\mathrm{T}}\Theta\boldsymbol{x}^s \tag{2}$$

$$f_t(\boldsymbol{x}) = \boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t + \boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t \tag{3}$$

其中, $\boldsymbol{x}^s \in \mathbf{R}^d$, $\boldsymbol{x}^t \in \mathbf{R}^d$ 分别是源域和目标域中实例; $\boldsymbol{w}_s \in \mathbf{R}^d$, $\boldsymbol{w}_t \in \mathbf{R}^d$ 分别是源域和目标域的对应实例在原始数据空间中的分类面参数; $\Theta \in \mathbf{R}^{r\times d}$ 是隐空间变换参数矩阵, r 是领域间共享的低维子空间的维数, 源域和目标域实例经过 Θ 变换后映射到一个公共的低维子特征隐空间上, 对应实例可视为 $\Theta\boldsymbol{x}^s$ 和 $\Theta\boldsymbol{x}^t$; $\boldsymbol{v}_s \in \mathbf{R}^r$, $\boldsymbol{v}_t \in \mathbf{R}^r$ 分别是实例 $\Theta\boldsymbol{x}^s$, $\Theta\boldsymbol{x}^t$ 在共享低维子空间中的分类面参数; 在此隐空间上, 由于拥有共同的变换参数矩阵 Θ, 期望源域和目标域数据拥有尽可能多的共性信息.

遵循式 (1), 结合本研究目的, 从特征变换的角度出发, 本文通过最小化式 (4) 所示函数同时学习得到源域和目标域的分类决策函数. 需要说明的是, 本研究假定仅有源域数据和目标域数据来挖掘源域和目标域的共享信息来实现迁移, 所以得到源域决策函数仅是一个手段, 得到目标域决策函数是本研究的最终目的.

$$\begin{aligned} & C_s\sum_{i=1}^{n_s} L\left(\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i^s + \boldsymbol{v}_s^{\mathrm{T}}\Theta\boldsymbol{x}_i^s, y_i^s\right) + \\ & \quad C_t\sum_{j=1}^{n_t} L\left(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j^t + \boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}_j^t, y_j^t\right) + \\ & \quad \frac{1}{2}\left(\|\boldsymbol{w}_s\|^2 + \|\boldsymbol{w}_t\|^2\right) + \\ & \quad \alpha\left(\|\boldsymbol{w}_s + \Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2 + \|\boldsymbol{w}_t + \Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2\right) \\ & \text{s.t.} \quad \Theta\Theta^{\mathrm{T}} = I_{r\times r} \end{aligned} \tag{4}$$

针对式 (4), 说明如下:

1) 式中第 1 项和 2 项表示源域和目标域的损失函数, 第 3 项和 4 项是正则化项, 通过前两项和后两项的平衡期望达到最小的结构化风险.

2) 式中第 1 项和 2 项涉及的源域和目标域决策函数分别表示为 $f_s(\boldsymbol{x}) = \boldsymbol{w}_s{}^{\mathrm{T}}\boldsymbol{x}^s + \boldsymbol{v}_s^{\mathrm{T}}\Theta\boldsymbol{x}^s$ 和 $f_t(\boldsymbol{x}) = \boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t + \boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t$. 以目标域 $f_t(\boldsymbol{x})$ 为例, 其包含两项 $\boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t$ 和 $\boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t$, 对于 $\boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t$ 而言, $\boldsymbol{w}_t$ 为 $\boldsymbol{x}^t$ 实例对应的目标域原始空间的分类面参数, 而对于 $\boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t$, 其功能首先是通过隐空间变换参数矩阵 Θ 将实例 $\boldsymbol{x}^t$ 由 d 维空间映射到 r 维的隐空间内形成隐空间内的实例 $\Theta\boldsymbol{x}^t$, 然后令 $\boldsymbol{v}_t$ 为实例 $\Theta\boldsymbol{x}^t$ 对应的隐空间的分类面参数.

3) 从另一个角度观察 $f_t(\boldsymbol{x}) = \boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t + \boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t$, 此时隐空间变换参数矩阵 Θ 所起作用可理解为将隐空间分类面参数 $\boldsymbol{v}_t$ 通过 Θ 映射回原始空间内, 即分类面 $\Theta^{\mathrm{T}}\boldsymbol{v}_t$ 与 $\boldsymbol{w}_t$ 处于同空间内. 其作用是用 $\Theta^{\mathrm{T}}\boldsymbol{v}_t$ (由隐空间映射而来的原始空间的分类面参数) 对 $\boldsymbol{w}_t$ 进行适当校正形成一个原始空间内的组合分类面参数 $\tilde{\boldsymbol{w}}_t$, 即 $\tilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \Theta^{\mathrm{T}}\boldsymbol{v}_t$.

4) 观察源域和目标域的组合分类面参数 $\tilde{\boldsymbol{w}}_s$ ($\tilde{\boldsymbol{w}}_s = (\boldsymbol{w}_s + \Theta^{\mathrm{T}}\boldsymbol{v}_s)$) 和 $\tilde{\boldsymbol{w}}_t$ ($\tilde{\boldsymbol{w}}_t = (\boldsymbol{w}_t + \Theta^{\mathrm{T}}\boldsymbol{v}_t)$) 可知, $\boldsymbol{w}_s$, $\boldsymbol{w}_t$ 和 $\boldsymbol{v}_s$, $\boldsymbol{v}_t$ 分别是源域和目标域在原始空间和隐空间中的分类面参数, Θ 是源域和目标域的公共隐空间变换参数矩阵, 通过 Θ 源域和目标域在某种程度上实现不同域之间的信息共享, 从而有效实现信息迁移, 最终源域的信息可用来增强目标域的模型性能.

5) 需要指出, 作为一种理论极端可能性, 可以通过 $\boldsymbol{v}_s = \boldsymbol{v}_t = 0$ 这个平凡解使得源域与目标域相互独立, 即此时式 (4) 的约束不起作用. 但该极端可能性在有共享隐空间时, 并不一定保证式 (4) 达到最小. 另外, $\boldsymbol{v}_s$, $\boldsymbol{v}_t$ 是在隐空间约束下为求式 (4) 达到最小值时而得. 因此, 即使在此空间下 $\boldsymbol{v}_s, \boldsymbol{v}_t$ 真的同时为零矢量, 也可解释为在此依赖于 r 的共享隐空间下, 源域与目标域没有迁移价值 (注意即使二者独立, 并不表明没有迁移价值). 该种特殊情形根据本文实验并不多见 (也未见). 因此, 式 (4) 是一个可行的迁移学习模型.

以经典的 SVM[12] 分类器为基础, 本文在传统迁移学习分类领域中引入隐空间学习策略得到低维共享子特征隐空间的迁移学习方法, 称为共享隐空间迁移 SVM (Shared latent space transfer SVM, SLSTSVM), 算法原理如图 1 所示.

虽然源域和目标域数据在实例层面上可能共享知识很少, 但在特征层面上通过隐空间变换参数矩阵 Θ 转换, 可以挖掘领域间更多的共享知识.

结合式 (4), 给出如下基于 SVM 的共享隐空间迁移学习的目标函数

$$\begin{aligned} \arg\min_{\boldsymbol{w}_s,\boldsymbol{w}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta} J_1 = \arg\min_{\boldsymbol{w}_s,\boldsymbol{w}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta} \Bigg\{ & \frac{1}{2}(\|\boldsymbol{w}_s\|^2 + \\ & \|\boldsymbol{w}_t\|^2) + C_s\sum_{i=1}^{n_s}\xi_i^s + C_t\sum_{j=1}^{n_t}\xi_j^t + \\ & \alpha\left(\|\boldsymbol{w}_s + \Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2 + \|\boldsymbol{w}_t + \Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2\right) \Bigg\} \end{aligned}$$

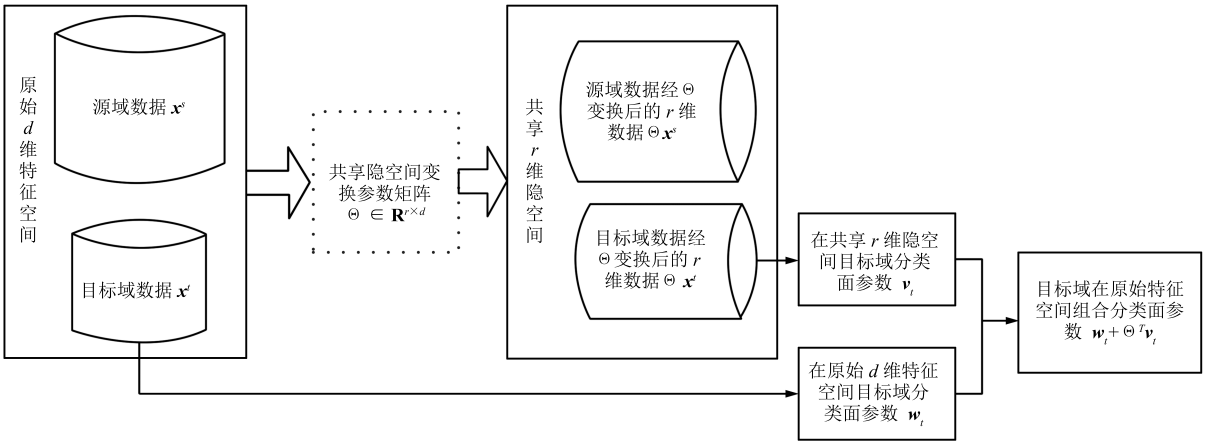


图 1　共享隐空间迁移 SVM 原理图

Fig. 1　Schematic of SLSTSVM method

$$\begin{aligned}\text{s.t.}\quad & y_i^s\left(\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_i^s+\boldsymbol{v}_s^{\mathrm{T}}\Theta\boldsymbol{x}_i^s\right)\geq 1-\xi_i^s,\ \ \xi_i^s\geq 0\\ & y_j^t\left(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j^t+\boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}_j^t\right)\geq 1-\xi_j^t,\ \ \xi_j^t\geq 0\\ & i=1,2,\cdots,n_s,\ \ j=1,2,\cdots,n_t,\ \ n_s\gg n_t\\ & \Theta\Theta^{\mathrm{T}}=I_{r\times r}\end{aligned}\tag{5}$$

其中, ξ_i^s 和 ξ_j^t 分别表示源域和目标域的松弛变量; C_s 和 C_t 是大于 0 的正则化参数; 第 1 项和第 4 项是结构风险项, 正则化参数 $\alpha>0$. 运用上述模型求解得到 $\boldsymbol{w}_t,\boldsymbol{v}_t,\Theta$ 后, 可根据决策函数 $f_t(\boldsymbol{x})=\boldsymbol{w}_t{}^{\mathrm{T}}\boldsymbol{x}^t+\boldsymbol{v}_t^{\mathrm{T}}\Theta\boldsymbol{x}^t$ 对目标域的测试样本或将来的新数据进行决策.

为便于求解, 引入辅助变量 $\boldsymbol{u}_s$, $\boldsymbol{u}_t$, 其中 $\boldsymbol{u}_s=\boldsymbol{w}_s+\Theta^{\mathrm{T}}\boldsymbol{v}_s$, $\boldsymbol{u}_t=\boldsymbol{w}_t+\Theta^{\mathrm{T}}\boldsymbol{v}_t$, 则式 (5) 化为

$$\begin{aligned}\arg\min_{\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta} J_2=\arg\min_{\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta}\Bigg\{&\frac{1}{2}(\|\boldsymbol{u}_s-\\ &\Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2+\|\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2)+\\ &C_s\sum_{i=1}^{n_s}\xi_i^s+C_t\sum_{j=1}^{n_t}\xi_j^t+\\ &\alpha\left(\|\boldsymbol{u}_s\|^2+\|\boldsymbol{u}_t\|^2\right)\Bigg\}\end{aligned}$$

$$\begin{aligned}\text{s.t.}\quad & y_i^s\boldsymbol{u}_s^{\mathrm{T}}\boldsymbol{x}_i^s\geq 1-\xi_i^s,\ \ \xi_i^s\geq 0\\ & y_j^t\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{x}_j^t\geq 1-\xi_j^t,\ \ \xi_j^t\geq 0\\ & i=1,2,\cdots,n_s,\ \ j=1,2,\cdots,n_t,\ \ n_s\gg n_t\\ & \Theta\Theta^{\mathrm{T}}=I_{r\times r}\end{aligned}\tag{6}$$

式 (6) 中, 有 5 个变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta$ 需要同时优化, 本文把 5 个变量分为 3 组: 1) $\boldsymbol{u}_s$ 和 $\boldsymbol{u}_t$; 2) $\boldsymbol{v}_s$ 和 $\boldsymbol{v}_t$; 3) Θ. 由于直接对这些参数优化比较繁琐困难, 本文采用模糊聚类、模糊神经网络等技术中经常采用的交替迭代策略对式 (6) 进行参数优化[13−18]. 该迭代过程包含两步: 1) 固定 $\{\Theta,\boldsymbol{v}_s,\boldsymbol{v}_t\}$, 通过优化目标函数 (6) 学习得到 $\{\boldsymbol{u}_s,\boldsymbol{u}_t\}$; 2) 固定 $\{\boldsymbol{u}_s,\boldsymbol{u}_t\}$, 通过优化目标函数 (6) 学习得到 $\{\Theta,\boldsymbol{v}_s,\boldsymbol{v}_t\}$. 循环迭代上述步骤直到目标函数收敛.

1.3　目标函数参数学习规则

定理 1. 假定变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta$ 的定义形式如式 (6) 所述, 则式 (6) 中变量 $\boldsymbol{v}_s,\boldsymbol{v}_t$ 的最优值可表示为 $\boldsymbol{v}_s=\Theta\boldsymbol{u}_s$, $\boldsymbol{v}_t=\Theta\boldsymbol{u}_t$.

证明. 假设变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\Theta$ 已知, 根据相关矩阵理论[19], $\|A\|_F^2=\mathrm{tr}(A^{\mathrm{T}}A)$, 式 (6) 变为

$$\begin{aligned}\arg\min_{\boldsymbol{v}_s,\boldsymbol{v}_t} J_3=&\arg\min_{\boldsymbol{v}_s,\boldsymbol{v}_t}\left\|\boldsymbol{u}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\right\|^2+\\ &\left\|\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\right\|^2=\arg\min_{\boldsymbol{v}_s,\boldsymbol{v}_t}\mathrm{tr}\Big(\left(\boldsymbol{u}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\right)^{\mathrm{T}}\times\\ &\left(\boldsymbol{u}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\right)+\left(\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\right)^{\mathrm{T}}\left(\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\right)\Big)=\\ &\arg\min_{\boldsymbol{v}_s,\boldsymbol{v}_t}\mathrm{tr}\left(\boldsymbol{u}_s^{\mathrm{T}}\boldsymbol{u}_s+\boldsymbol{v}_s^{\mathrm{T}}\Theta\Theta^{\mathrm{T}}\boldsymbol{v}_s-2\boldsymbol{u}_s^{\mathrm{T}}\Theta^{\mathrm{T}}\boldsymbol{v}_s+\right.\\ &\left.\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_t+\boldsymbol{v}_t^{\mathrm{T}}\Theta\Theta^{\mathrm{T}}\boldsymbol{v}_t-2\boldsymbol{u}_t^{\mathrm{T}}\Theta^{\mathrm{T}}\boldsymbol{v}_t\right)\end{aligned}\tag{7}$$

式 (7) 对变量 $\boldsymbol{v}_s,\boldsymbol{v}_t$ 分别求偏导数且令其值为 0. 则得:

$$\begin{aligned}\boldsymbol{v}_s&=\Theta\boldsymbol{u}_s\\ \boldsymbol{v}_t&=\Theta\boldsymbol{u}_t\end{aligned}\tag{8}$$

□

定理 2. 假定变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta$ 的定义形式如式 (6) 所述, 则式 (6) 中变量 $\boldsymbol{u}_s,\boldsymbol{u}_t$ 的最优值可表示为

$$\boldsymbol{u}_s=\frac{\sum\limits_{i=1}^{n_s}\eta_i^s y_i^s\boldsymbol{x}_i^s+\Theta^{\mathrm{T}}\boldsymbol{v}_s}{1+\alpha}$$

$$\boldsymbol{u}_t=\frac{\sum\limits_{j=1}^{n_t}\eta_j^t y_j^t\boldsymbol{x}_j^t+\Theta^{\mathrm{T}}\boldsymbol{v}_t}{1+\alpha} \tag{9}$$

证明. 假设已知变量 $\Theta,\boldsymbol{v}_s,\boldsymbol{v}_t$, 通过优化式 (6) 得到 $\boldsymbol{u}_s,\boldsymbol{u}_t$, 此时式 (6) 对应的是一个二次规划问题. 构造式 (6) 的拉格朗日函数

$$\begin{aligned}L=&\frac{1}{2}\left(\|\boldsymbol{u}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2+\|\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2\right)+\\&C_s\sum_{i=1}^{n_s}\xi_i^s+C_t\sum_{j=1}^{n_t}\xi_j^t+\alpha\left(\|\boldsymbol{u}_s\|^2+\|\boldsymbol{u}_t\|^2\right)+\\&\sum_{i=1}^{n_s}\eta_i^s\left(1-\xi_i^s-y_i^s\boldsymbol{u}_s^{\mathrm{T}}\boldsymbol{x}_i^s\right)-\sum_{i=1}^{n_s}\gamma_i^s\xi_i^s+\\&\sum_{j=1}^{n_t}\eta_j^t\left(1-\xi_j^t-y_j^t\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{x}_j^t\right)-\sum_{j=1}^{n_t}\gamma_j^t\xi_j^t\end{aligned} \tag{10}$$

令

$$\begin{aligned}L_s=&\frac{1}{2}\|\boldsymbol{u}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2+C_s\sum_{i=1}^{n_s}\xi_i^s+\alpha\|\boldsymbol{u}_s\|^2+\\&\sum_{i=1}^{n_s}\eta_i^s\left(1-\xi_i^s-y_i^s\boldsymbol{u}_s^{\mathrm{T}}\boldsymbol{x}_i^s\right)-\sum_{i=1}^{n_s}\gamma_i^s\xi_i^s\end{aligned} \tag{11}$$

$$\begin{aligned}L_t=&\frac{1}{2}\|\boldsymbol{u}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2+C_t\sum_{j=1}^{n_t}\xi_j^t+\alpha\|\boldsymbol{u}_t\|^2+\\&\sum_{j=1}^{n_t}\eta_j^t\left(1-\xi_j^t-y_j^t\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{x}_j^t\right)-\sum_{j=1}^{n_t}\gamma_j^t\xi_j^t\end{aligned} \tag{12}$$

这里拉格朗日乘子 $\eta_i^s>0$, $\eta_j^t>0$, $\gamma_i^s>0$, $\gamma_j^t>0$. 式 (11) 对 $\boldsymbol{u}_s$, ξ_i^s $(i=1,2,\cdots,n_s)$ 分别求偏导, 由 KKT 条件可得:

$$\boldsymbol{u}_s=\frac{\sum\limits_{i=1}^{n_s}\eta_i^s y_i^s\boldsymbol{x}_i^s+\Theta^{\mathrm{T}}\boldsymbol{v}_s}{1+\alpha} \tag{13}$$

$$C_s=\eta_i^s+\gamma_i^s \tag{14}$$

把式 (13) 和式 (14) 代回式 (11), 得:

$$\begin{aligned}\arg\min_{\eta_i^s}J_4=&\arg\min_{\eta_i^s}\frac{1}{2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\eta_i^s\eta_j^s y_i^s y_j^s\times\\&{\boldsymbol{x}_i^s}^{\mathrm{T}}\boldsymbol{x}_j^s+\sum_{i=1}^{n_s}\eta_i^s\left(y_i^s{\boldsymbol{x}_i^s}^{\mathrm{T}}\Theta^{\mathrm{T}}\boldsymbol{v}_s-1-\alpha\right)\\\text{s.t.}\quad&0\le\eta_i^s\le C_s\end{aligned} \tag{15}$$

同理可得:

$$\begin{aligned}\arg\min_{\eta_i^t}J_5=&\arg\min_{\eta_i^t}\frac{1}{2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\eta_i^t\eta_j^t y_i^t y_j^t\times\\&{\boldsymbol{x}_i^t}^{\mathrm{T}}\boldsymbol{x}_j^t+\sum_{i=1}^{n_t}\eta_i^t\left(y_i^t{\boldsymbol{x}_i^t}^{\mathrm{T}}\Theta^{\mathrm{T}}\boldsymbol{v}_t-1-\alpha\right)\\\text{s.t.}\quad&0\le\eta_i^t\le C_t\end{aligned} \tag{16}$$

此时采用经典的二次规划算法即可得到 $\eta_i^s\eta_j^t$ 的最优解, 从而得到当前 $\boldsymbol{u}_s\boldsymbol{u}_t$ 最优解, 具体表达为式 (9). □

定理 3. 假定变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t,\Theta$ 的定义形式如前所述, 式 (6) 在变量 $\boldsymbol{u}_s,\boldsymbol{u}_t,\boldsymbol{v}_s,\boldsymbol{v}_t$ 取得最优值的条件下, 变量 Θ 的求解可以转化为矩阵 $[\boldsymbol{u}_s,\boldsymbol{u}_t]$ 的奇异值分解问题.

证明. 假设已知变量 $\boldsymbol{u}_s,\boldsymbol{u}_t$ 且对于式 (6) 均取得最优值, 则式 (6) 的最优化问题变为式 (17)

$$\begin{aligned}&\left[\hat{\boldsymbol{v}}_s,\hat{\boldsymbol{v}}_t,\hat{\Theta}\right]=\\&\quad\arg\min\left(\|\hat{\boldsymbol{u}}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2+\|\hat{\boldsymbol{u}}_t-\Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2\right)\\&\text{s.t}\quad\Theta\Theta^{\mathrm{T}}=I_{r\times r}\end{aligned} \tag{17}$$

假设变量 $\boldsymbol{\Theta}$ 固定, 式 (17) 等价于

$$\begin{aligned}[\hat{\boldsymbol{v}}_s,\hat{\boldsymbol{v}}_t]=&\arg\min\left(\|\hat{\boldsymbol{u}}_s-\Theta^{\mathrm{T}}\boldsymbol{v}_s\|^2+\|\hat{\boldsymbol{u}}_t-\right.\\&\left.\Theta^{\mathrm{T}}\boldsymbol{v}_t\|^2\right)=\arg\min\left((\|\hat{\boldsymbol{u}}_s\|^2-\|\Theta\hat{\boldsymbol{u}}_s\|^2)+\right.\\&\left.(\|\hat{\boldsymbol{u}}_t\|^2-\|\Theta\hat{\boldsymbol{u}}_t\|^2)\right)\end{aligned} \tag{18}$$

且式 (18) 在 $\hat{\boldsymbol{v}}_s=\Theta\hat{\boldsymbol{u}}_s$, $\hat{\boldsymbol{v}}_t=\Theta\hat{\boldsymbol{u}}_t$ 取得最优值. 进而式 (17) 可写成式 (19)

$$\begin{aligned}&\widehat{\Theta}=\arg\max\left(\|\Theta\hat{\boldsymbol{u}}_s\|^2+\|\Theta\hat{\boldsymbol{u}}_t\|^2\right)\\&\text{s.t.}\quad\Theta\Theta^{\mathrm{T}}=I_{r\times r}\end{aligned} \tag{19}$$

令 $U=[\boldsymbol{u}_s,\boldsymbol{u}_t]\in\mathbf{R}^{d\times 2}$, 则有

$$\begin{aligned}&\widehat{\Theta}=\arg\max\operatorname{tr}\left(\Theta UU^{\mathrm{T}}\Theta^{\mathrm{T}}\right)\\&\text{s.t.}\quad\Theta\Theta^{\mathrm{T}}=I_{r\times r}\end{aligned} \tag{20}$$

由矩阵理论知识[19] 知, 式 (20) 的求解可通过对矩阵 U 实施奇异值分解得到. 令 $U=V_1DV_2^{\mathrm{T}}$, 且对角矩阵 D 的对角线元素按降序排列, 那么矩阵 V_1^{T} 的前 r 行就是变量 Θ 的最优解. □

1.4 算法描述

根据上述章节分析和推导, 给出共享隐空间迁移 SVM 算法描述.

参数. 共享低维隐空间维数 r, 正则化参数 α, C_s, C_t.

输入. 源域数据 $\{(\boldsymbol{x}_i^s,y_i^s)|i=1,2,\cdots,n_s\}$, 目标域数据 $\{(\boldsymbol{x}_j^t,y_j^t)|j=1,2,\cdots,n_t\}$, 其中, $n_s\gg n_t$.

输出. 隐空间变换参数矩阵 $\Theta \in \mathbf{R}^{r\times d}$, 原始空间分类面参数 $\boldsymbol{w}_s, \boldsymbol{w}_t$ 共享低维隐空间分类面参数 $\boldsymbol{v}_s, \boldsymbol{v}_t$ ($\boldsymbol{w}_s = \boldsymbol{u}_s - \Theta^{\mathrm{T}}\boldsymbol{v}_s, \boldsymbol{w}_t = \boldsymbol{u}_t - \Theta^{\mathrm{T}}\boldsymbol{v}_t$).

初始化. $\boldsymbol{u}_s^0 = 0_{d\times 1}$, $\boldsymbol{u}_t^0 = 0_{d\times 1}$, $\Theta_{r\times d}^0$, 当前迭代值 $iter = 0$, 设置最大迭代次数 $iterm$ 和误差阈值 ε.

Repeat

1) 利用式 (8) 更新 $\boldsymbol{v}_s^{iter}$, $\boldsymbol{v}_t^{iter}$;

2) $iter = iter + 1$;

3) 利用式 (9) 更新 $\boldsymbol{u}_s^{iter}$, $\boldsymbol{u}_t^{iter}$;

4) 设 $U = [\boldsymbol{u}_s^{iter}, \boldsymbol{u}_t^{iter}]$, 对 U 实施奇异值分解操作 $U = V_1 D V_2^{\mathrm{T}}$, 且对角矩阵 D 的对角线元素按降序排列, 那么矩阵 V_1^{T} 的前 r 行就是变量 Θ^{iter} 的最优解.

Until

$\|\boldsymbol{u}_t^{iter} - \boldsymbol{u}_t^{iter-1}\| < \varepsilon$ 或 $\|\boldsymbol{v}_t^{iter} - \boldsymbol{v}_t^{iter-1}\| < \varepsilon$ 或 $\|\Theta^{iter} - \Theta^{iter-1}\| < \varepsilon$ 或 $iter > iterm$.

1.5 算法收敛问题说明

对于给出的共享隐空间迁移 SVM 算法, 以第 $iter+1$ 步的迭代学习为例, 对其收敛性分析如下:

1) 在第 $iter+1$ 步, 固定 $\boldsymbol{v}_s^{iter}$, $\boldsymbol{v}_t^{iter}$, Θ^{iter} 不变, 由式 (9) 可得:

$$\boldsymbol{u}_s^{iter+1} = \frac{\sum\limits_{i=1}^{n_s} \eta_i^{s^{iter}} y_i^s \boldsymbol{x}_i^s + \Theta^{iter^{\mathrm{T}}} \boldsymbol{v}_s^{iter}}{1+\alpha}$$

$$\boldsymbol{u}_t^{iter+1} = \frac{\sum\limits_{j=1}^{n_t} \eta_j^{t^{iter}} y_j^t x_j^t + \Theta^{iter^{\mathrm{T}}} \boldsymbol{v}_t^{iter}}{1+\alpha}$$

而 $\eta_i^{s^{iter}}$, $\eta_j^{t^{iter}}$ 为式 (15) 和式 (16) 的最优解. 式 (15) 和式 (16) 是典型的二次规划问题, 由优化理论可知, 式 (15) 和式 (16) 存在全局最优解, 从而式 (9) 取得全局最优解.

2) 在第 $iter+1$ 步, 固定 $\boldsymbol{u}_s^{iter}$, $\boldsymbol{u}_t^{iter}$ 不变, 令 $U = [\boldsymbol{u}_s^{iter}, \boldsymbol{u}_t^{iter}]$, 对 U 实施奇异值分解操作 $U = V_1 D V_2^{\mathrm{T}}$, 且对角矩阵 D 的对角线元素按降序排列, 那么矩阵 V_1^{T} 的前 r 行就是变量 Θ 的最优解. 由于 $\boldsymbol{u}_s^{iter}$, $\boldsymbol{u}_t^{iter}$ 在其当前迭代步骤中取得全局最优解, 从而可以确保变量 Θ 在第 $iter+1$ 步取得全局最优解.

3) 在第 $iter+1$ 步, 固定 $\boldsymbol{u}_s^{iter}$, $\boldsymbol{u}_t^{iter}$, Θ^{iter} 不变, 由定理 1 的证明过程可知, $\boldsymbol{v}_s^{iter+1} = \Theta^{iter}\boldsymbol{u}_s^{iter}$ $\boldsymbol{v}_t^{iter+1} = \Theta^{irer}\boldsymbol{u}_t^{iter}$ 是目标函数 J_3 取极值的必要但非充分条件. 此时, $\boldsymbol{v}_s^{iter+1} = \Theta^{iter}\boldsymbol{u}_s^{iter}$ $\boldsymbol{v}_t^{iter+1} = \Theta^{irer}\boldsymbol{u}_t^{iter}$ 是 J_3 的极值或鞍点. 故变量 $\boldsymbol{v}_s, \boldsymbol{v}_t$ 的更新规则不能保证问题严格收敛.

据上述分析可知, SLSTSVM 算法并不能保证严格收敛. 值得指出的是交替迭代优化技术在智能模型学习方法中[13−18] 经常被使用, 如经典的 FCM (Fuzzy C-means) 等聚类算法, 其通常能收敛于某个局部最优解或鞍点. 虽然此类算法目前在理论上不能保证是严格收敛或是条件收敛, 但已有的采用交替迭代优化技术的迭代算法表明此优化技术在大多数场合是非常简单而有效的[13−18]. 在最新的该类学习方法, 如文献 [5] 即选择了该种优化方法来解决多任务间共享子空间模型参数的优化问题, 并得到了不错的效果. 正如文献 [18] 指出, 此类交替迭代的优化方法其收敛性仍是一个有待进一步深入探讨的开放性问题, 保证该类方法的严格收敛是一个值得深入研究的课题. 另外, 由于初始化等不同因素的影响, 目前许多交叉迭代算法多次执行后最终可能收敛于不同的局部最优解. 针对此问题, 一个可行的解决方案是探讨新的优化算法来求解给定的优化目标函数. 如近年来具有更好寻优能力的进化计算优化技术 (遗传算法、粒子群算法等) 已受到较多关注并被尝试应用于优化不同的建模模型.

1.6 与相关工作的比较分析

文献 [5] 的方法和本文工作有一定相似之处, 本节对二者的联系和区别做简单分析讨论. 文献 [5] 中, Ando 和 Zhang 从多任务中通过子空间学习多个任务之间的共享预测结构, 提出一个相似的算法, 其优化问题目标函数为

$$\min_{\{\boldsymbol{u}_l, \boldsymbol{v}_l\}, \Theta} \sum_{l=1}^{m} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} L\left(\boldsymbol{u}_l^{\mathrm{T}} \boldsymbol{x}_i^l, \boldsymbol{y}_i^l\right) \right) + \alpha \left\| \boldsymbol{u}_l - \Theta^{\mathrm{T}} \boldsymbol{v}_l \right\|^2$$

$$\text{s.t.} \quad \Theta\Theta^{\mathrm{T}} = I \tag{21}$$

其中, $\boldsymbol{x}_i^l$ 表示第 l 个任务中的第 i 个实例, n_l 表示第 l 个任务中实例的数量. 文献 [5] 采用以奇异值分解为基础的交替结构优化 (SVD-based alternating structure optimization, SVD-ASO) 方法来求解得到一个局部最优解.

本文和文献 [5] 的相似之处是二者都在不同的数据特征空间上寻找一个共享的子特征隐空间, 并且采用迭代方法求解得到一个局部最优解. 不同之处在于, 文献 [5] 研究的是多任务学习场景, 各任务间的实例数量是接近的, 均衡考虑最终所得分类器在各任务上的分类误差, 其最终模型仅平衡各任务在各自原始特征空间上的泛化能力; 而对于本文研究的迁移场景, 考虑的是单任务中实例分布不同且包含大量源域实例和极少量目标域实例的问题. 为了更加适应本文研究场景, 我们的目标函数得到的分类器不关注最终分类器在源域的学习性能. 用于

决策的组合特征向量可实现不同域之间的信息共享, 有效实现了信息迁移. 最终利用源域的信息增强目标域的模型性能.

2 实验分析

为验证本文算法的有效性, 在三个文本数据集 (20NewsGroup[20]、Reuters-21578[20] 和 EmailSpam[21]) 上进行实验. 用于比较的方法有 SVM、AuxSVM[22] 和 LMPROJ[23]. 对于 SVM, 从三个方面进行比较: 1) 训练数据为源域数据、测试数据为目标域中的所有数据 (SVM); 2) 训练数据为目标域中的带标签数据、测试数据为目标域中的所有数据 (SVM(t)); 3) 训练数据为源域数据和目标域中的带标签数据、测试数据为目标域中的所有数据 (SVMt). 比较算法的参数均根据具体应用通过网格搜索的方式获得, 所有算法中均采用线性核函数.

采用 SVM 算法是为验证源域数据和目标域数据分布不同; 采用 SVM(t) 算法是为验证目标域中带标签数据太少不足以训练一个足够好的分类器; 采用 SVMt 算法是为验证当训练数据和测试数据分布不同时, 迁移学习方法和传统机器学习方法相比表现出更好的性能; 采用 AuxSVM 算法是为验证在相同的场景设置下, 本文所提算法表现出的可比较的性能; 采用 LMPROJ 算法是为验证半监督的迁移学习方法和无监督的迁移学习方法的性能优劣.

对于 SVM 来说, 可由 libsvm[24] 软件实现. 正则化参数 C 取值范围为 $\{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$, 采取 5 重交叉验证法来选取最优值. 其他算法都在 Matlab (R2009a) 环境下实现, 且均通过网格搜索的方式来确定优化的模型参数. 对于本文提出的方法, 源域和目标域共享的低维隐空间维数 r 需要人工确定, 在当前的基于隐空间的学习方法中一般均用实验验证或网格寻优的方法确定, 目前常用的方法主要有 Ando 方法[5]、Zheng 方法[6] 和 Ji 方法[25]. 本文参考 Ando 方法, 对于实验中的三个文本数据集, 只关心共享空间维数在 $5 \sim 100$ 之间范围, 因此 r 的取值范围设为 $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100\}$; 参数 C_s, C_t 取值范围为 $\{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$; 参数 α 的取值范围为 $\{0, 1\text{E}{-6}, 1\text{E}{-5}, 1\text{E}{-4}, 1\text{E}{-3}, 1\text{E}{-2}, 1\text{E}{-1}, 1\}$. 参数 r, C_s, C_t, α 均通过网格搜索的方式确定最优值.

2.1 数据集与评价准则

本节采用文本数据集 20NewsGroup、Reuters-21578 和 EmailSpam 为实验数据集, 因都不是为迁移学习设计, 所以必须采用特殊的数据生成策略, 使数据符合我们提出问题的要求. 20NewsGroup 和 Reuters-21578 数据集具有层次结构. 前者包含近 20 000 新闻组, 大约分成 20 类, 包含 6 个顶层类别, 每个顶层类别下分别包含若干个子类别, 将其中子类别为空的顶层类别删掉, 剩余 4 个大的类别 rec, talk, comp 和 sci; 后者包含近 21 000 个文档集, 按大的类别将其中的文档描述为 orgs, people 和 places 共 3 个大类, 每个类别下面包含相应的子类别. 本文把父类作为分类目标, 然后对源域和目标域数据按照子类进行划分. 如图 2 所示, A 和 B 是两个父类. 问题是要对 A 和 B 进行分类. 在 A 下面, 有两个子类 $A1$ 和 $A2$, 而 $B1$ 和 $B2$ 是 B 下面的两个子类. $A1$ 和 $B1$ 是源域数据, $A2$ 和 $B2$ 是目标域数据.

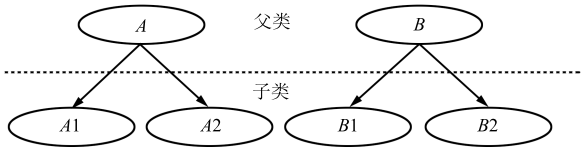


图 2 数据生成示例

Fig. 2 The example of data generation

本文采用与文献 [23] 相同的实验设置, 对于 20Newsgroup 和 Reuters-21578 数据集, 分别从顶层大类中抽取 4 个和 3 个大类构建学习数据集, 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, 不同的子类认为不同的领域. 产生迁移数据集如表 1 和表 2 所示. EmailSpam 数据集有 3 个 Email 子集 (表示为 User1, User2 和 User3), 代表 3 个不同用户. 学习任务是划分出 spam 邮件和非 spam 邮件. 由于数据集中不同用户的 spam 邮件和非 spam 邮件是不同的, 因此 3 个 email 数据集的数据分布是不同但相关的. 每个数据集包含 2 500 封邮件, 其中一半邮件为非 spam 邮件 (类标签为 1), 另一半为 spam 邮件 (类标签为 -1). 根据文献 [2] 的设置, 构造了三个迁移数据集如表 3 所示.

按照上述数据生成策略生成的三个文本数据集被应用到了大量有关迁移学习的实验[2, 23, 25−29] 中, 证明是能有效体现迁移学习思想的数据集.

为与其他方法对比及分析参数对分类结果的影响, 本文采用正确率作为衡量分类器整体性能的指标; 为使实验结果尽量公正, 在一组实验中选定最优参数并 10 次随机运行后, 用均值统计分类精度.

2.2 实验结果分析

在目标域中带标签数据和目标域总数据不同比例下, 4 个分类器 SVM(t), SVMt, AuxSVM 和 SLSTSVM, 在不同数据集上的分类精度曲线如图 3

表 1 迁移学习文本分类数据 20NewsGroup

Table 1 The transfer learning text classification datasets 20NewsGroup

编号	数据集	源域/目标域	正例	负例
NG1	comp vs. sci	源域	comp.graphics	sci.crypt
			com.os.ms-windows.misc	sci.electronics
		目标域	comp.sys.ibm.pc.hardware	sci.med
			comp.sys.mac.hardware	sci.space
			comp.windows.x	
NG2	rec vs. talk	源域	rec.autos	talk.politics.guns
			rec.motorcycles	talk.politics.mideast
		目标域	rec.sport.baseball	talk.politics.misc
			rec.sport.hockey	talk.religion.misc
NG3	comp vs. talk	源域	comp.graphics	talk.politics.mideast
			comp.sys.mac.hardware	talk.religion.misc
			comp.windows.x	
		目标域	comp.os.ma-windows.misc	talk.politics.guns
			comp.sys.ibm.pa.hardware	talk.politics.misc
NG4	sci vs. talk	源域	sci.electronicis	talk.politics.misc
			sci.med	talk.religion.misc
		目标域	sci.crypt	talk.politics.guns
			sci.space	talk.politics.mideast
NG5	comp vs. rec	源域	comp.graphics	rec.motorcycles
			comp.sys.ibm.pc.hardware	rec.sport.hockey
			comp.sys.mac.hardware	
		目标域	comp.os.ms-windows.misc	rec.autos
			comp.windows.x	rec.sport.baseball
NG6	rec vs. sci	源域	rec.autos	sci.med
			rec.sport.baseball	sci.space
		目标域	rec.motorcycles	sci.crypt
			rec.sport.hockey	sci.electronics

表 2 迁移学习文本分类数据 Reuters-21578

Table 2 The transfer learning text classification datasets Reuters-21578

编号	数据集	源域	目标域
Rt1	orgs vs. people	orgs.*, people.*	orgs.*, people.*
Rt2	orgs vs. places	orgs.*, places.*	orgs.*, places.*
Rt3	people vs. places	people.*, places.*	people.*, places.*

∼6 所示. 在目标域带标签数据和目标域总数据比例为 0.3 时, 分类精度如表 4 所示. 实验结论如下:

1) 基线方法 SVM 因把源域数据训练得到的模型直接应用于目标域测试数据, 故在几乎所有数据集上的分类性能均低于其他方法, 验证了源域数据和目标域数据分布是不同的, 从而进一步验证了传统的机器学习方法只适用于训练数据和测试数据分布相同的情况; 基线方法 SVM(t) 在目标域少量带标签数据上训练得到的模型应用于目标域数据上, 数据分布相同, 但由于带标签数据数量少, 不足以训

表 3 迁移学习文本分类数据 EmailSpam

Table 3 The transfer learning text classification datasets EmailSpam

编号	数据集	源域	目标域
Spam1	User1 vs. User2	User1′s emails	User2′s emails
Spam2	User2 vs. User3	User2′s emails	User3′s emails
Spam3	User3 vs. User1	User3′s emails	User1′s emails

表 4 文本分类精度比较

Table 4 Classification accuracy comparison on text datasets

方法	NG1	NG2	NG3	NG4	NG5	NG6	Rt1	Rt2	Rt3	Spam1	Spam2	Spam3
SVM	0.6642	0.5924	0.7123	0.7080	0.7203	0.8220	0.7302	0.6661	0.6124	0.6581	0.7560	0.8401
SVM(t)	0.7513	0.8160	0.8161	0.7975	0.8013	0.8613	0.7980	0.7880	0.7640	0.7688	0.8388	0.8113
SVMt	0.7650	0.7460	0.8080	0.7450	0.8388	0.8639	0.7150	0.6601	0.7020	0.7675	0.8938	0.7988
AuxSVM	0.8540	0.8880	0.8952	0.8280	0.8680	0.8976	0.7130	0.6670	0.7430	0.8095	0.8762	0.8120
LMPROJ	0.8201	0.7674	0.9402	0.7203	0.8504	0.8702	0.8463	0.8020	0.7080	0.6764	0.7890	0.6980
SLSTSVM	0.9398	0.9362	0.9102	0.9240	0.9070	0.9182	0.8860	0.8900	0.8980	0.9763	0.9313	0.9013

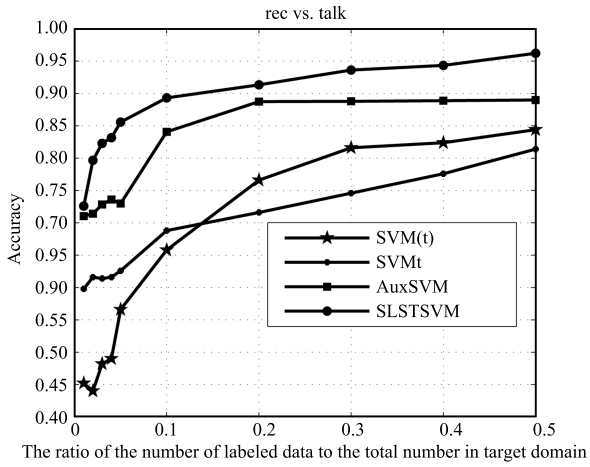


图 3 不同分类器在 NG2 数据集上的分类精度曲线图

Fig. 3 The accuracy curves on NG2 dataset for different classifiers

练高性能的模型, 故在部分数据集上的分类性能高于其他迁移方法, 而在部分数据集上的分类性能低于其他迁移方法; 基线方法 SVMt 在源域数据和目标域带标签数据上训练模型, 因为目标域带标签测试数据较少, 直接应用到目标域数据, 故在大部分数据集上其分类性能高于其他两个基线方法, 而低于其他 3 个迁移方法.

2) 在 3 种迁移算法中, SLSTSVM 算法均高于其他迁移算法, 原因为: a) SLSTSVM 算法中目标域中部分数据带标签, 是半监督迁移学习的情况, 而 LMPROJ 算法目标域数据不带标签, 是无监督迁移学习的情况; b) SLSTSVM 算法是基于特征隐空间迁移, 而其他两种算法是基于实例的迁移算法, 当源域和目标域数据分布差别比较大时, 基于实例的方法很难在实例层面上找到一些共享的知识, 而基于特征的方法可在隐含的特征层面上找到一些交集作为迁移的知识.

3) 从图 3 可看出, 目标域中带标签数据占目标域总数据的比例在 [0.01, 0.5] 范围内变化时, 算法 SLSTSVM 的分类精度均高于算法 SVM(t) 和 SVMt; 目标域中带标签数据占目标域总数据的比例大于 0.1 时, SVM(t) 算法性能提高的速度明显高于算法 SLSTSVM、SVMt 和 AuxSVM, 但其总体性能略低于其他两个迁移算法.

4) 从图 4 和图 5 可看出, 目标域中带标签数据占目标域总数据的比例大于 0.2 时, 算法 SLSTSVM 性能提升的速度高于算法 AuxSVM 算法提升的速度, 说明在同样场景设置下, 基于特征的迁移学习方法比基于实例的迁移学习方法在一定程度上更容易学习到共同的知识.

5) 从图 6 可看出, 目标域中带标签数据占目标域总数据的比例大于 0.4 时, 算法 SVM(t) 性能和算法 SLSTSVM 性能相当, 略高于算法 AuxSVM, 说明在数据集中随着目标域中带标签数据比例的提高, 传统的分类算法相比迁移算法显示出较明显的优势; 而算法 SVMt 的性能高于算法 AuxSVM, 低于算法 SLSTSVM, 说明虽然目标域数据已经较充分, SLSTSVM依然能够从共享结构中学习到有益的

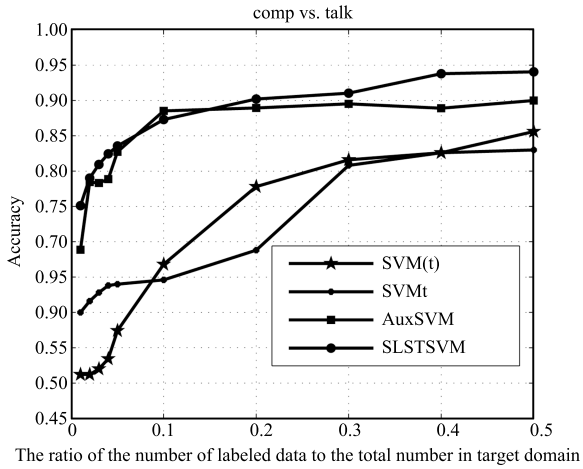


图 4 不同分类器在 NG3 数据集上的分类精度曲线图

Fig. 4 The accuracy curves on NG3 dataset for different classifiers

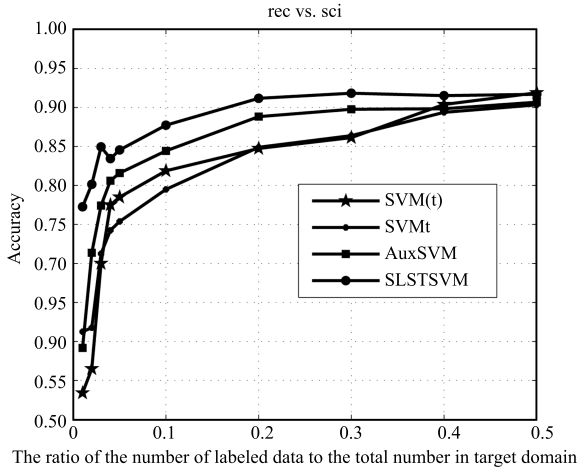


图 5 不同分类器在 NG6 数据集上的分类精度曲线图

Fig. 5 The accuracy curves on NG6 dataset for different classifiers

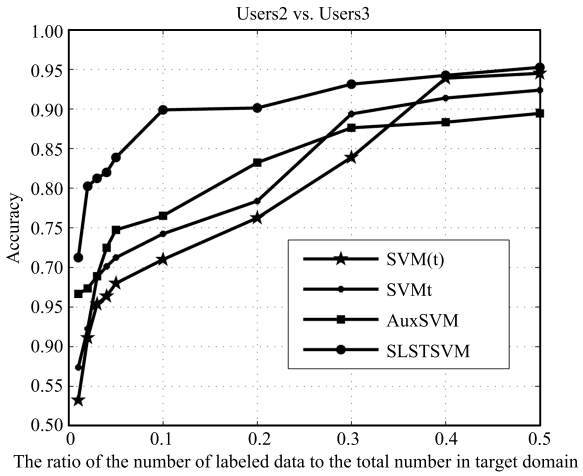


图 6 不同分类器在 Spam2 数据集上的分类精度曲线图

Fig. 6 The accuracy curves on Spam2 dataset for different classifiers

知识. 此时 AuxSVM 由于是基于实例进行迁移, 故其带来的更多的是负迁移, 即噪声样本带来的负作用.

2.3 参数敏感性实验

本节讨论 SLSTSVM 算法参数对分类性能的影响. 在目标域带标签数据占目标域总数据比例为 0.01 和 0.3 两种情况下, 共享隐空间维数对分类精度的影响曲线如图 7 和图 8 所示. 正则化参数 α 对分类精度的影响曲线如图 9 所示. 正则化参数 C_s, C_t 对分类精度的影响曲线如图 10 和图 11 所示. 从图 10 和图 11 看出, 正则化参数 C_s, C_t 在一定范围内的取值对分类精度有一定的影响. 因此本文所提算法在 $\{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$ 内寻优.

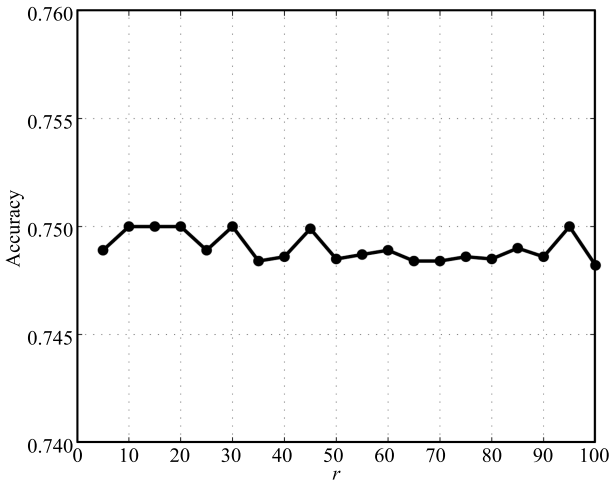


图 7 共享隐空间维数 r 在数据集 NG1 上对分类精度的影响曲线图: rate = 0.01

Fig. 7 The accuracy curve of the dimensionality of shared latent space on NG1 dataset: rate = 0.01

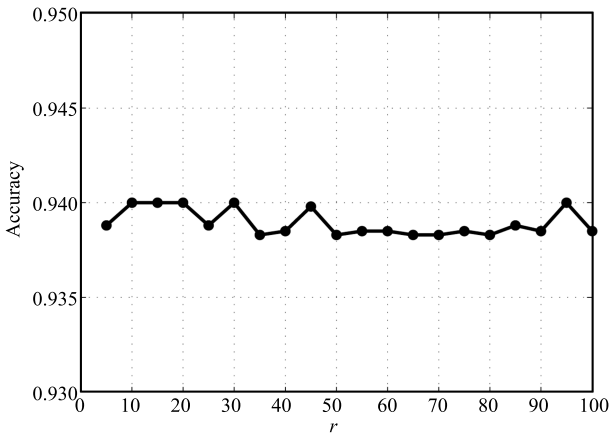


图 8 共享隐空间维数 r 在数据集 NG1 上对分类精度的影响曲线图: rate = 0.3

Fig. 8 The accuracy curve of the dimensionality of shared latent space on NG1 dataset: rate = 0.3

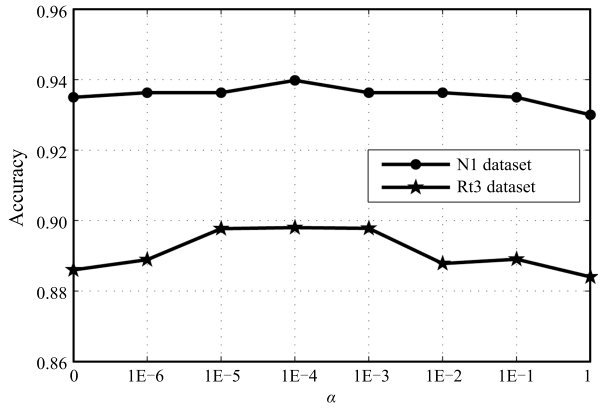


图 9 参数 α 在 NG1 和 Rt3 数据集上对分类精度的影响曲线图

Fig. 9 The accuracy curves of parameter α on NG1 and Rt3 dataset

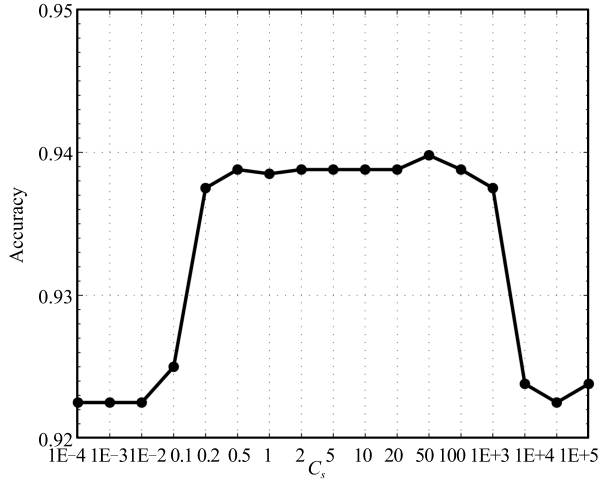


图 10 参数 C_s 在 NG1 数据集上对分类精度的影响曲线图

Fig. 10 The accuracy curve of parameter C_s on NG1 dataset

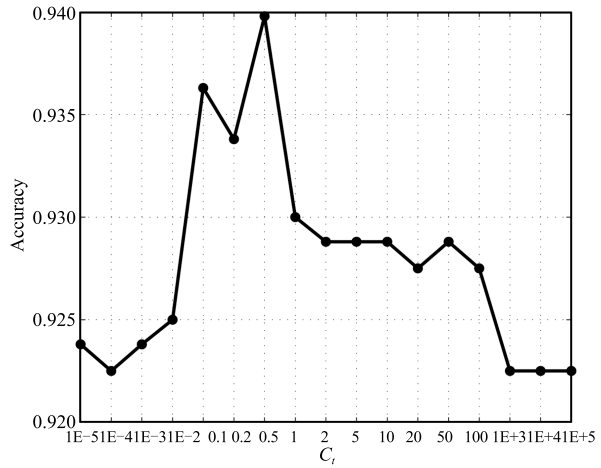


图 11 参数 C_t 在 NG1 数据集上对分类精度的影响曲线图

Fig. 11 The accuracy curve of parameter C_t on NG1 dataset

3 结论

迁移学习的关键技术是如何寻找源域和目标域之间公共的 “知识” 表示, 本文针对归纳迁移学习方法中的特征迁移提出了一种新的方法. 在源域中包含有大量带标签数据, 而目标域中只含有少量带标签数据的场景下, 本文基于传统支持向量机的分类框架, 在最大化源域和目标域数据分类间隔和控制源域和目标域分类模型复杂度的基础上, 通过寻求一个特征变换, 使得源域和目标域在此特征变换下共享某些 “知识”. 在文本分类数据集上的系列实验显示了所提方法的有效性. 虽然本文方法展现了较好性能, 其依然面临一些进一步需要探讨的问题. 如: 怎样把本文提出的方法融合到多源域数据的迁移学习中还需进一步深入探讨; 此外, 有关隐空间维数的选取策略在当前的基于隐空间的学习方法中一般均为实验验证或网格寻优的方法确定, 如何理论性地给出隐空间维数的确定方法是一项具有挑战性的工作, 也是今后研究的重点之一.

References

1 Evgeniou T, Micchelli C A, Pontil M. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 2005, **6**(4): 615−637

2 Duan L X, Tsang I W, Xu D. Domains transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(3): 465−479

3 Tu W T, Sun S L. A subject transfer framework for egg classification. *Neurocomputing*, 2012, **82**: 109−116

4 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

5 Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal Machine Learning Research*, 2005, **6**: 1817−1853

6 Zheng V W, Pan J L, Yang Q, Pan J F. Transferring multi-device localization models using latent multi-task learning. In: Proceedings of the 23th International Conference on Artificial Intelligence. Chicago, USA: ACM, 2008. 1427−1432

7 Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction. In: Proceedings of the 23th International Conference on Artificial Intelligence. Chicago, USA: ACM 2008. 677−682

8 Si S, Tao D C, Geng B. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(7): 929−942

9 Shao M, Castillo C, Gu Z H, Fu Y. Low-rank transfer subspace learning. In: Proceedings of the 12th International Conference on Data Mining. Brussels, Belgium: IEEE 2012. 1104−1109

10 Yang S Z, Lin M, Hou C P, Zhang C S, Wu Y. A general framework for transfer sparse subspace learning. *Neural Computing and Applications*, 2012, **21**(7): 1801−1817

11 Gupta S K, Phung D, Adams B, Adams B, Venkatesh S. Regularized nonnegative shared subspace learning. *Data mining and knowledge discovery*, 2013, **26**(1): 57−97

12 Vapnik V. *Statistical Learning Theory*. New Jersey: Wiley-Interscience Press, 1998.

13 Domeniconi C, Gunopulos D, Ma S, Yan B J, Al-Razgan M, Papadopoulos D. Locally adaptive metrics for custering high dimensional data. *Data Mining and Knowledge Discovery*, 2007, **14**(1): 63−97

14 Wu K L, Yu J, Yang M S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests. *Pattern Recognition Letters*, 2005, **26**(5): 639−652

15 Deng Z H, Choi K S, Chung F L, Wang S T. Enhanced soft subspace clustering integrating within-cluster and between-cluster information. *Pattern Recognition*, 2010, **43**(3): 767−781

16 Yu J, Cheng Q S, Huang H K. Analysis of the weighting exponent in the FCM. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2004, **34**(1): 634−639

17 Yang S, Yan S C, Zhang C, Tang X O. Bilinear analysis for kernel selection and nonlinear feature extraction. *IEEE Transactions on Neural Networks*, 2007, **18**(5): 1442−1452

18 Jiang Yi-Zhang, Deng Zhao-Hong, Wang Shi-Tong. Mamdani-Larsen type transfer learning fuzzy system. *Acta Automatica Sinica*, 2012, **38**(9): 1393−1409
(蒋亦樟, 邓赵红, 王士同. ML 型迁移学习模糊系统. 自动化学报, 2012, **38**(9): 1393−1409)

19 Golub G H, Van Loan C F. *Matrix Computations (3rd Edition)*. Baltimore: The Johns Hopkins University Press, 1996.

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

21 Discovery Challenge [Online], available: http://www.ecmlpkdd2006.org/challenge.html, January 3, 2013

22 Wu P, Dietterich T G. Improving SVM accuracy by training on auxiliary data sources. In: Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004. 110−117

23 Quanz B, Huan J. Large margin transductive transfer learning. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2009. 1327−1336

24 Chang C C, Lin C J. LIBSVM: A library for support vector machines [Online], available: http://www.csie.ntu.edu.tw/~cjlin/libsvm, October 14, 2012

25 Ji S W, Tang L, Yu S P, Ye J P. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery From Data*, 2010, **4**(2), Article No.8, DOI: 10.1145/1754428.1754431

26 Tao Jian-Wen, Wang Shi-Tong. Kernel distribution consistency based local domain adaptation learning. *Acta Automatica Sinica*, 2013, **39**(8): 1295−1309
(陶剑文, 王士同. 核分布一致局部领域适应学习. 自动化学报, 2013, **39**(8): 1295−1309)

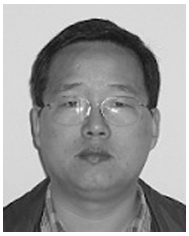
27 Dai W Y, Yang Q, Xue G R, Yu Y. Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 193−200

28 Gu Xin, Wang Shi-Tong, Xu Min. A new cross-multidomain classification algorithm and its fast version for large datasets. *Acta Automatica Sinica*, 2014, **40**(3): 531−547
(顾鑫, 王士同, 许敏. 基于多源的跨领域数据分类快速新算法. 自动化学报, 2014, **40**(3): 531−547)

29 Tao Jian-Wen, Wang Shi-Tong. Domain adaptation kernel support vector machine. *Acta Automatica Sinica*, 2012, **38**(5): 797−811
(陶剑文, 王士同. 领域适应核支持向量机. 自动化学报, 2012, **38**(5): 797−811)



董爱美 江南大学数字媒体学院博士研究生, 齐鲁工业大学信息学院讲师. 主要研究方向为人工智能, 模式识别, 机器学习. 本文通信作者.
E-mail: amdong@163.com
(**DONG Ai-Mei** Ph. D. candidate at the School of Digital Media, Jiangnan University, and lecturer at Qilu University of Technology. Her research interest covers artificial intelligence, pattern recognition, and machine learning. Corresponding author of this paper.)



王士同 江南大学数字媒体学院教授. 主要研究方向为人工智能和机器学习.
E-mail: wxwangst@yahoo.com.cn
(**WANG Shi-Tong** Professor at the School of Digital Media, Jiangnan University. His research interest covers artificial intelligence and machine learning.)