文章编号: 1001-0920 (2014) 01-0141-06 **DOI:** 10.13195/j.kzyjc.2012.1450

TL-SVM: 一种迁移学习算法

许 敏[1,2], 王士同[1], 顾 鑫[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 无锡职业技术学院 物联网技术学院，江苏 无锡 214121)

摘 要: 迁移学习旨在利用大量已标签源域数据解决相关但不相同的目标域问题. 当与某领域相关的新领域出现时, 若重新标注新领域, 则样本代价昂贵, 丢弃所有旧领域数据又十分浪费. 对此, 基于SVM算法提出一种新颖的迁移学习算法——TL-SVM, 通过使用目标域少量已标签数据和大量相关领域的旧数据来为目标域构建一个高质量的分类模型, 该方法既继承了基于经验风险最小化最大间隔SVM的优点, 又弥补了传统SVM不能进行知识迁移的缺陷. 实验结果验证了该算法的有效性.

关键词: 迁移学习；分类；支持向量机

中图分类号: TP273 **文献标志码:** A

TL-SVM: A transfer learning algorithm

XU Min[1,2], *WANG Shi-tong*[1], *GU Xin*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Internet of Things Engineering，Wuxi Institute of Technology，Wuxi 214121，China．Correspondent：XU Min，E-mail：xum@wxit.edu.cn)

Abstract: Transfer learning(TL) aims to solve related but different target domain problems by using plenty of labeled source domain data. When the task from one new domain comes, new domain samples are relabeled costly, and it would be a waste to discard all the old domain data. Therefore, an algorithm called TL-SVM based on the SVM algorithm is proposed, which uses a small amount of target domain tag data and a large number of source domain old data to build a high-quality classification model. The method inherits the advantages of the maximum interval SVM based on empirical risk minimization and makes up for the defects that traditional SVM can not migrate knowledge. Experimental results show the effectiveness of the proposed algorithm.

Key words: transfer learning；classification；support vector machine

0 引 言

传统的机器学习要求训练数据和测试数据分布一致, 而在实际应用中此假设未必成立. 若直接利用训练数据获得分类器对测试数据进行分类预测, 则因分布不同可能会导致分类性能变差; 若重新对测试集进行大量人工标注, 则获得类标签代价非常昂贵. 换言之, 若拥有大量相关领域的训练数据, 则尽管它们或多或少已过时, 但仍然包含一部分有用信息. 如何将相关领域知识“迁移”到新数据集的分类模型是一个全新的研究方向[1-3].

目前, 相关学者针对传统分类方法进行改进, 提出了若干迁移学习算法. 其主流算法有: 1) 基于特征的迁移学习, 旨在找出源域与目标域相同或相近的特征表示, 从而将源域知识传递给目标域. 如Pan等[4]提出通过降维方式找出新的特征空间, 建立了源域与目标域的桥梁; Xie等[5]提出了结合回归和SVD降维的LatentMap方法, 使两域分布差异减小. 2) 基于实例迁移学习, 使用源域存在的相关信息帮助目标域训练一个有效分类器. 如Tradaboost算法[6]利用Boosting技术建立一种自动调整权重的机制, 将重要的源域训练数据权重增加, 不重要的源域训练数据权重减小, 最终这些带权重的辅助训练数据将会作为额外训练数据, 与目标域已标签训练数据一起提高分类模型的可靠度; SemiTrBoost算法[7]通过扩展Tradaboost, 结合Hedge (β) 与半监督Boosting方法, 利用目标域已标记与无标记样本对源域样本的相

收稿日期: 2012-09-28；**修回日期:** 2013-04-09.

基金项目: 国家自然科学基金项目(61272210, 61170122)；江苏省研究生创新工程项目(CXZZ12-0759).

作者简介: 许敏(1980−), 女, 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、生物信息等研究.

关性做出更好的判断. 此外, 陈德品[8]就跨领域的排序学习算法进行了研究, 旨在利用源域标注数据帮助目标域进行排序学习.

本文提出一种具有迁移能力的TL-SVM算法, 利用已标签但可能过时的源域相关知识辅助只包含少量已标签样本的目标域建立分类模型. 其中, 源域大量已标签样本集(T_s)与目标域测试集(Test)分布相似, 目标域少量已标签样本集(T_t)与目标域测试集分布相同, 通过将T_s的知识$\boldsymbol{w}_s$"迁移"给T_t, 获得了分类模型f:$X \to Y$, 使f能对Test进行正确分类. 本文算法给出了一种适用于SVM的迁移学习方法, 继承了基于经验风险最小化框架最大间隔SVM的优点, 并从数学理论上严格证明了TL-SVM分类模型满足KKT条件.

1 TL-SVM算法

1.1 软间隔优化SVM

Vapnik[9]于1995年提出了支持向量机(SVM)概念, 在VC维理论和结构风险最小原理基础上, 根据有限样本信息在模型复杂性和学习能力之间寻求最佳折衷, 以期获得最好的泛化能力[10].

设$T_t = \{(\boldsymbol{x}_1^t, y_1^t), (\boldsymbol{x}_2^t, y_2^t), \cdots, (\boldsymbol{x}_n^t, y_n^t)\}$, $\boldsymbol{x}_i^t$为目标域第i个样本, y_i^t为第i个样本的类标签, n为训练集规模. 给出软间隔优化SVM模型最优化问题如下:

$$\begin{aligned}&\min_{\boldsymbol{w}_t, b_t, \xi^t} \frac{1}{2}||\boldsymbol{w}_t||^2 + C_t \sum_{i=1}^{n} \xi_i^t.\\ &\text{s.t. } y_i^t((\boldsymbol{w}_t \cdot \boldsymbol{x}_i) + b_t) \geqslant 1 - \xi_i^t,\\ &\quad i = 1, 2, \cdots, n;\\ &\quad \xi_i^t \geqslant 0, i = 1, 2, \cdots, n. \end{aligned} \tag{1}$$

其中: $\xi^t = (\xi_1^t, \xi_2^t, \cdots, \xi_n^t)^{\mathrm{T}}$, $C_t > 0$是一个惩罚参数.

传统SVM要求训练集与测试集分布相同, 若直接使用目标域训练集训练分类模型, 则会因目标域已标签样本过少而导致分类性能较低. 对此, 在原一阶范数软间隔优化SVM基础上, 本文提出一种具有迁移学习能力的TL-SVM方法.

1.2 TL-SVM算法

1.2.1 TL-SVM算法理论依据

SVM分类器由$(\boldsymbol{w}, b)$组成, 判别函数为$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$, 分类决策函数为$L(\boldsymbol{x}) = \mathrm{sign}(f(\boldsymbol{x}))$. 本文算法的理论依据是: 若两领域相关, 则两域分类器各自的$\boldsymbol{w}$值应相近. 通过在SVM目标式中增加$\mu||\boldsymbol{w}_t - \boldsymbol{w}_s||^2$项实现两域间迁移学习. 其中: $||\boldsymbol{w}_t - \boldsymbol{w}_s||^2$表示两域分类器的差异程度, 该值越大则分类器间差异越大, 反之越小; 参数μ控制惩罚程度. TL-SVM原理可用图1表示.

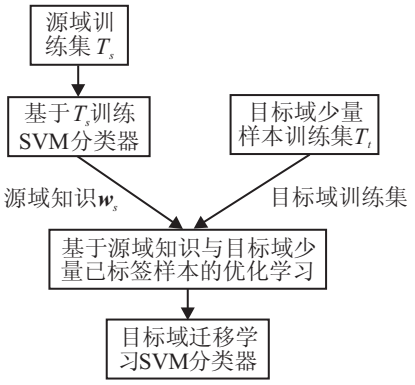


图1 TL-SVM原理

1.2.2 TL-SVM目标函数的构造

首先, 存在源域SVM分类器$(\boldsymbol{w}_s, b_s)$; 然后, 利用源域分类器知识$\boldsymbol{w}_s$对目标域进行迁移学习. 优化目标问题如下:

$$\begin{aligned}&\min_{\boldsymbol{w}_t, b_t} \frac{1}{2}||\boldsymbol{w}_t||^2 + C_t \sum_{i=1}^{n} \xi_i^t + \mu||\boldsymbol{w}_t - \boldsymbol{w}_s||^2.\\ &\text{s.t. } y_i^t((\boldsymbol{w}_t \cdot \boldsymbol{x}_i^t) + b_t) \geqslant 1 - \xi_i^t,\ i = 1, 2, \cdots, n;\\ &\quad \xi_i^t \geqslant 0,\ i = 1, 2, \cdots, n. \end{aligned} \tag{2}$$

其中: n为目标域样本数, C_t为控制目标域训练集惩罚误差程度. 式(2)体现了最大化间隔和最小化误差间的平衡, 同时, 用$\mu||\boldsymbol{w}_t - \boldsymbol{w}_s||^2$表示迁移学习项. $y_i^t((\boldsymbol{w}_t \cdot \boldsymbol{x}_i^t) + b_t) \geqslant 1 - \xi_i^t$保证目标域分类器对目标域已标签样本尽可能分类正确, 既体现了源域在训练过程的辅助作用, 又充分利用了目标域数据的分类特性. 式(2)优化问题对应的拉格朗日形式为

$$\begin{aligned}&L(\boldsymbol{w}_t, \xi_i^t, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) =\\ &\frac{1}{2}||\boldsymbol{w}_t||^2 + C_t \sum_{i=1}^{n} \xi_i^t + \mu||\boldsymbol{w}_t - \boldsymbol{w}_s||^2 +\\ &\sum_{i=1}^{n} \beta_i(1 - \xi_i^t - y_i^t((\boldsymbol{w}_t \cdot \boldsymbol{x}_i^t) + b_t)) - \sum_{i=1}^{n} \gamma_i \xi_i^t. \end{aligned} \tag{3}$$

其中: $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_n)^{\mathrm{T}}$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)^{\mathrm{T}}$是拉格朗日乘子列向量. 分别置方程$L(\boldsymbol{w}_t, \xi_i^t, \boldsymbol{\beta}, \boldsymbol{\gamma})$对原始变量$\boldsymbol{w}_t$、$\xi_i^t$和$b_t$的偏导数为0, 可得

$$\frac{\partial L}{\partial \xi_i^t} = C_t - \beta_i - \gamma_i = 0 \ \Rightarrow\ 0 \leqslant \beta_i \leqslant C_t, \tag{4}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_t} = 0 \ \Rightarrow\ \boldsymbol{w}_t = \frac{2\mu \boldsymbol{w}_s + \sum_{i=1}^{n} \beta_i (y_i^t \cdot \boldsymbol{x}_i^t)}{2\mu + 1}, \tag{5}$$

$$\frac{\partial L}{\partial b_t} = 0 \Rightarrow \sum_{i=1}^{n} \beta_i y_i^t = 0. \tag{6}$$

将式(5)、(6)代入目标函数(3), 可得原问题的对偶形式为

$$\min_{\boldsymbol{\beta}} \frac{1}{2(2\mu + 1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i^t y_j^t (\boldsymbol{x}_i^t \cdot \boldsymbol{x}_j^t) +$$

$$\sum_{i=1}^{n}\Big(\frac{2\mu y_i^t(\boldsymbol{x}_i^t\cdot\boldsymbol{w}_s)}{2\mu+1}-1\Big)\beta_i-\frac{\mu}{2\mu+1}||\boldsymbol{w}_s||^2.$$

$$\text{s.t. } 0\leqslant\beta_i\leqslant C_t,\ \sum_{i=1}^{n}\beta_i y_i^t=0, i=1,2,\cdots,n. \quad (7)$$

1.2.3 TL-SVM 算法分析

由式 (2) 可知, 源域与目标域间的迁移学习通过 $||\boldsymbol{w}_t-\boldsymbol{w}_s||^2$ 实现. 若 μ 值为 0, 则退化为目标域 SVM 训练; 若 μ 值趋向无穷大, 则有

$$\boldsymbol{w}_t=\lim_{\mu\to\infty}\frac{2\mu\boldsymbol{w}_s+\sum_{i=1}^{n}\beta_i(y_i^t\cdot\boldsymbol{x}_i^t)}{2\mu+1}=\boldsymbol{w}_s,$$

即 $\boldsymbol{w}_t^*$ 介于 $\boldsymbol{w}_t$ 和 $\boldsymbol{w}_s$ 之间.

式 (7) 中的核函数 $\boldsymbol{K}(\cdot)$ 只有保证 Mercer 核时, 才能保证其是二次凸规划, 所求的解才为全局最优解. 为了验证这一问题, 给出如下定理.

引理 1　定义在 $R^n\times R^n$ 上的对称函数 K 为 Mercer 核的充要条件是: 对于任意 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\in R^n$, K 关于 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 的 Gram 矩阵是半正定的[11].

定理 1　式 (7) 核函数是 Mercer 核.

证明　式 (7) 核矩阵

$$\tilde{\boldsymbol{K}}=[\tilde{k}_{ij}]_{n\times n}=\frac{1}{2(2\mu+1)}(y_i^t\boldsymbol{x}_i^t)^{\mathrm{T}}(y_j^t\boldsymbol{x}_j^t)$$

是实对称矩阵. 下面证明矩阵 $\tilde{\boldsymbol{K}}$ 是半正定的. 设任意 $c_1,c_2,\cdots,c_n\in R$, 则有

$$\sum_{i,j=1}^{n}c_ic_j\tilde{k}_{ij}=$$

$$\sum_{i,j=1}^{n}c_ic_j\Big(\frac{1}{2(2\mu+1)}(y_i^t\boldsymbol{x}_i^t)(y_j^t\boldsymbol{x}_j^t)\Big)=$$

$$\frac{1}{2(2\mu+1)}\Big(\sum_{i=1}^{n}c_i(y_i^t\boldsymbol{x}_i^t)\Big)\Big(\sum_{j=1}^{n}c_j(y_j^t\boldsymbol{x}_j^t)\Big)=$$

$$\frac{1}{2(2\mu+1)}\Big(\sum_{i=1}^{n}c_i(y_i^t\boldsymbol{x}_i^t)\Big)^2\geqslant 0, \quad (8)$$

因此 $\boldsymbol{c}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{c}\geqslant 0$, 即矩阵 $\tilde{\boldsymbol{K}}$ 半正定且为 Mercer 核函数. □

引理 2　假设二次规划中的 Gram 矩阵为半正定矩阵, 则该二次规划为凸二次规划[11].

引理 3　假设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[11].

定理 2　设 $\boldsymbol{\alpha}$ 是对偶问题 (7) 的解, 则原始问题 (3) 对于 $\boldsymbol{w}_t$ 的解为全局最优解, 并可表示为

$$\boldsymbol{w}_t=\frac{2\mu\boldsymbol{w}_s+\sum_{i=1}^{n}\beta_i(y_i^t\cdot\boldsymbol{x}_i^t)}{2\mu+1}. \quad (9)$$

证明　由引理 2 及定理 1 的证明可知, 式 (7) 为凸二次规划, 又根据引理 3 的满足条件可知, 该二次规划的解为全局最优解.

需要指出的是, 式 (9) 中, $\frac{2\mu\boldsymbol{w}_s}{2\mu+1}$ 为从源域学到的知识, $\sum_{i=1}^{n}\beta_i(y_i^t\cdot\boldsymbol{x}_i^t)\Big/(2\mu+1)$ 为从当前目标域数据中学到的新知识. □

1.2.4 TL-SVM 算法流程

迁移学习目标域分类器具体算法如下:

1) 获得源域知识 $\boldsymbol{w}_s$, 选择适当的惩罚参数 C_t, μ;

2) 构造式 (7) 凸二次规划问题, 得到解 $\beta^*=(\beta_1^*,\beta_2^*,\cdots,\beta_n^*)^{\mathrm{T}}$, 并按式 (5) 获得 $\boldsymbol{w}_t^*$;

3) 选取位于开区间 $(0,C_t)$ 中的 β^* 分量 β_j^*, 据此计算 $b_t^*=y_j^t-(\boldsymbol{w}_t^*\cdot\boldsymbol{x}_j^t)$;

4) 构造分划超平面 $(\boldsymbol{w}_t^*\cdot\boldsymbol{x}_t)+b_t^*=0$, 由此求得决策函数 $f(\boldsymbol{x}^e)=\mathrm{sign}(g(\boldsymbol{x}^e))$, 其中 $g(\boldsymbol{x}^e)=(\boldsymbol{w}_t^*\cdot\boldsymbol{x}^e)+b_t^*$.

2 实验结果与分析

为验证本文 TL-SVM 方法的有效性, 本节在不同数据集上对其性能进行评估, 所采用的数据集有两类: 1) 人造数据集; 2) 多个针对不同应用领域的真实数据集, 包括文本数据集 20Newsgroups、Reuters-21578 和 UCI 数据集 mushroom.

实验中将 TL-SVM 方法与相关方法进行性能比较, 以目标域测试集分类精度作为评价指标, 具体描述为

$$\text{Accuracy}=\frac{|\{\boldsymbol{x}|\boldsymbol{x}_t\in D_t\bigcap f(\boldsymbol{x}_i)=Y_t\}|}{|\{\boldsymbol{x}|\boldsymbol{x}_t\in D_t\}|}.$$

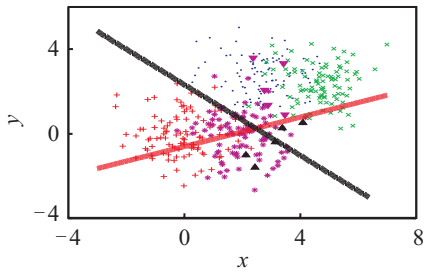
其中: D_t 表示目标域测试集, Y_t 表示 $\boldsymbol{x}_t$ 的真实类标签, $f(\boldsymbol{x}_t)$ 表示使用学习所得分类器对 $\boldsymbol{x}_t$ 进行分类得到的结果.

所有实验均通过网格搜索方式确定最优参数, 并采用线性核函数.

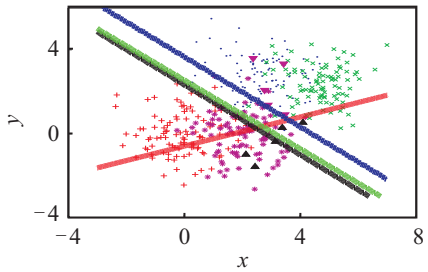
实验环境: Intel Core 2 2.40 GHz CPU, 2.39 GHz 1.94 GB RAM, Windows XP SP3, Matlab 7.1 等.

2.1 人造数据集实验

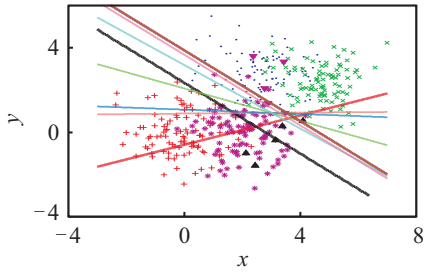
生成源域、目标域两类高斯分布数据样本各 100 个, 如图 2 所示. 其中: "+" 表示源域正类样本, "•" 表示源域负类样本, "*" 表示目标域正类样本, "×" 表示目标域负类样本. 源域两类样本方差均为 1, 均值分别为 0 和 2.8; 目标域两类样本方差均为 1, 均值分别为 0 和 2.5. 为了表示数据缓慢变化, 将目标域数据样本向 X 轴正向移动 2 个单位. 目标域样本集分为两部分, 图 2 中上三角、下三角分别为目标域已标签样本, 其余目标域样本作为测试集. 通过此方式生成两域数据既存在相关性, 又不完全相同.



(a) 两域SVM训练效果图



(b) 两域合并、迁移训练效果图



(c) 不同μ值迁移学习效果图

图 2 人造数据集实验

由图 2(a) 可知: 若用源域分类器对目标域测试集进行分类, 则会由于目标域数据集存在漂移, 分类精度不高, 仅为 0.865 0; 若用少量目标域已标签样本训练分类器, 则会由于样本过少不能反映目标域本身特性, 分类精度也不高, 为 0.735 0. 图 2(b) 中与源域分类器靠近的分类器表示两域合并训练所得分类器, 虽分类精度稍有提高, 为 0.88, 但因目标域样本规模与源域相比较小, 从而对分类器影响很小; 图 2(b) 中另一个分类器是使用 TL-SVM 算法所得, 由该图可知, 此法既考虑了目标域已标签样本本身的作用, 又学习了源域数据集的分类经验, 分类精度显著提升, 达到 0.97. 图 2(c) 为随 μ 值增大的迁移学习效果图, 可以看出, 随着 μ 值增大, 目标域 $\boldsymbol{w}_t$ 向源域 $\boldsymbol{w}_s$ 靠拢, 同时又保证了已标签样本的分类性能, 亦即既保留了目标域本身性质, 又学习了源域知识, 提升了目标域分类模型的性能.

2.2 真实数据集实验

为了评估本文算法性能, 在两个文本数据集 (20Newsgroups, Reuters-21578) 和一个非文本 UCI 数据集 (mushroom) 上进行实验验证.

2.2.1 数据预处理

1) 文本数据集的预处理.

20Newsgroups 是由 7 大类 20 小类组成的报文数据集, 本文选用其中 comp、rec、sci、talk 四大类进行实验, 将大类中子类分别作为源域和目标域, 如图 3 所示. comp.sys.ibm.pc.hardware 作为源域中的正类, rec.autos 作为源域中的负类; comp.sys.mac.hardware 作为目标域中的正类, rec.motorcycles 作为目标域中的负类. 按文献 [12] 对文本数据集进行预处理.

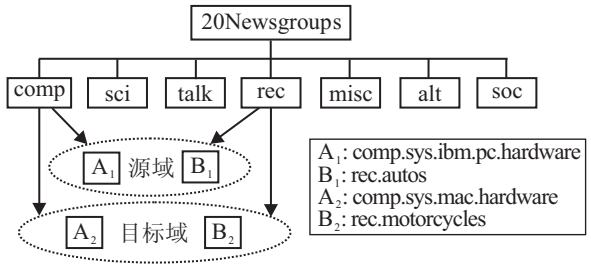


图 3 迁移数据集构造方式

2) UCI 数据集的预处理.

文献 [4] 对 mushroom 数据集进行如下处理: 基于 stalk-shape 属性将数据集分为两部分, 源域包含所有该属性值为 enlarging 的样本, 目标域包含所有该属性值为 tapering 的样本, 使两域分布不同.

2.2.2 算法比较研究

实验从两方面进行比较: 1) 使用源域、目标域已标签样本集、源域与目标域已标签样本集合并 3 种情况下的样本集作为 SVM 的训练集进行训练, 分别记为 SVM-S、SVM-T 和 SVM-ST; 2) 迁移学习方面, 与 Tradaboost 算法进行比较, 选用 SVM 作为 Tradaboost 基准分类器.

按文献 [6] 进行数据集设置: 将目标域数据集划分成测试集和训练集两部分. 在目标域中选取源域训练集规模的 1% 作为目标域训练集, 其余为目标域测试集. SVM 算法参数 C 在 (0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 500, 1 000, 2 000, 4 000, 8 000, 10 000, 100 000) 中选取; 因目标域参与训练的样本数较少, 应尽量使样本分类正确, 故参数 C_t 值不宜过大, 应在 (0.000 1, 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100) 中选取; TL-SVM 算法 μ 在 (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) 中选取. Tradaboost 算法按文献 [6] 设置参数.

1) 不同算法分类精度比较.

表 1 列出了各算法在 10 个数据集上的分类精度. 可以看出, TL-SVM 算法的分类效果较为理想, 在 10 个数据集上, 有 6 个数据集分类精度最高, 另外 4 个

分类精度也与最优值非常接近. 由表 1 还可知:

① 观察 SVM-T 列, 仅用目标域少量已标签样本进行分类, 除 comp.vs.talk 数据集外, 错误率均大于 30%, 说明少量目标域已标签样本不能很好反映两类特性, 分类效果不显著.

② 观察 SVM-S 与 SVM-ST 列, 分类效果较 SVM-T 大多有所提高, 说明源域对目标域的分类可起到一定作用, 但提高程度取决于两域相关性: 如 comp.vs.rec、comp.vs.sci 等数据集, 提高幅度较大, 说明两域相关性较高; peo.vs.pla 分类精度不升反降, 说明两域相关性较低或无相关性.

③ 观察最后两列实验结果, 分类精度大多优于前 3 列. SVM 直接利用监督学习算法进行训练, 没有考虑不同域间的迁移学习; 两域合并训练 (SVM-ST), 目标域样本数量有限对分类影响较小; 而迁移学习算法既能保证目标域训练集对建立分类器的主导作用, 又能学习源域已有知识. 因此, 迁移学习算法性能优于两域合并训练.

表 1　各方法实验结果比较

数据集	SVM-S	SVM-T	SVM-ST	TL-SVM	Tradaboost
comp.vs.rec	0.841 9	0.567 2	0.846 9	0.910 3	**0.913 3**
comp.vs.talk	0.899 2	0.701 6	0.907 8	**0.952 4**	0.950 9
comp.vs.sci	0.819 1	0.550 7	0.874 9	0.892 1	**0.901 2**
rec.vs.sci	0.670 5	0.540 0	0.782 4	0.776 8	**0.779 9**
rec.vs.talk	0.835 0	0.570 3	0.889 7	**0.904 9**	0.890 2
sci.vs.talk	0.559 2	0.623 5	0.678 1	**0.720 6**	0.657 4
org.vs.pla	0.683 8	0.652 1	0.684 3	0.722 9	**0.743 7**
org.vs.peo	0.668 9	0.680 3	0.708 8	**0.725 5**	0.705 7
peo.vs.pla	0.598 3	0.645 4	0.579 7	**0.614 4**	0.614 2
enlarge.vs.tap	0.751 8	0.624 3	0.751 0	**0.857 3**	0.851 8

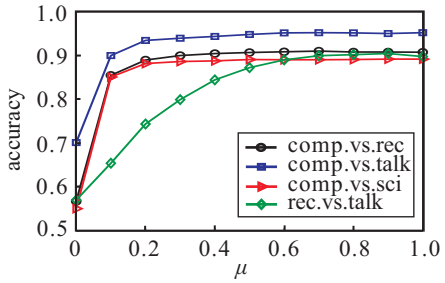
2) μ 值变化对分类精度的影响.

这里仅在文本数据集上讨论 μ 值对分类精度的影响 (见图 4). 由表 1 和图 4 可知, comp.vs.rec、comp.vs.talk、comp.vs.sci 和 rec.vs.talk 四组数据集源域分类器对目标域测试集的分类精度均大于 80%, 且两域训练集合并训练后, 测试精度有所提升, 表明两域相关性较大. 如图 4(a) 所示, 4 组数据集随着 μ 值增大, 分类精度随之提升; μ 值越大, 分类精度越趋向于稳定. 因此, 若两域相关性较大, 则可选取较大的 μ 值.

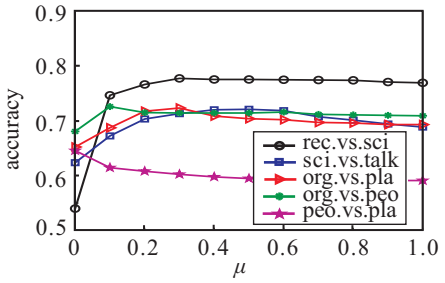
剩下 5 组数据集源域分类器对目标域测试集的分类精度均小于 70%, 且分为两种情况:

① peo.vs.pla 数据集. 该数据集两域合并训练后, 对目标域测试集分类精度不升反降, 表明两域无相关性, 故当两域不相关时, 迁移学习不能获得成功.

② 剩余 4 组数据集. 该 4 组数据集两域合并训练后, 目标域测试集分类精度略有提升, 表明两域有相关性, 但相关性不高. 如图 4(b) 所示, 最优 μ 值均小于 0.5. 因此, 当两域相关性较低时, 应选取较小的 μ 值.



(a) 领域相关性较高



(b) 领域相关性较低

图 4　μ 值变化对分类精度影响

3) 样本数的变化对不同算法分类精度的影响.

下面实验主要研究目标域训练集样本数增加对不同算法分类精度的影响. 实验在 rec.vs.talk 数据集上进行. 样本数依次取源域训练样本总数的 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40%, 50%, 比较 SVM-T、TL-SVM 两方法的分类精度, 如图 5 所示.

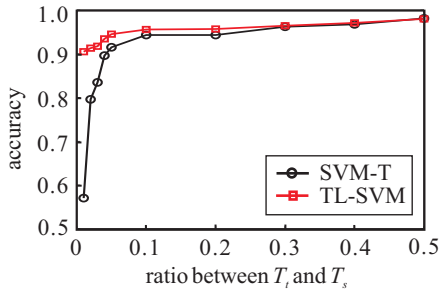


图 5　样本数变化对分类精度的影响

由图 5 可知:

① 随着目标域训练样本数量的增加, 有监督学习 SVM-T 信息量也增加, 分类精度呈上升趋势, 当比例达到 30% 时, 分类精度趋于稳定.

② 当目标域样本数量小于源域样本数量的 5% 时, 迁移算法效果显著. 但当比例大于 10% 时, 迁移算法效果与 SVM-T 相当; 若比例继续增大到 50%, 则迁移算法效果劣于 SVM-T. 原因在于源域训练集不仅含有有助于目标域训练的有用知识, 也包含噪音数据. 当目标域训练集数量过少不足以获得一个较好分类器时, 这些有用知识可以帮助目标域获得较好的分类器; 但当目标域训练集数量增大到足以获得一个较好

分类器时, 源域中的噪音数据将会起反作用, 故分类精度略低于SVM-T. 需要指出的是: 获得足够多的目标域训练样本以获得较好的分类器代价非常昂贵.

4) 抗噪实验.

本节在comp.vs.talk数据集上进行抗噪音实验. 在训练样本中加入均值为0、不同标准差的高斯噪音, 表2列出了50次实验后分类精度与噪音水平间的关系.

表 2 噪声水平与分类精度间的关系

噪声水平(标准差)	0.01	0.02	0.03	0.04	0.05
SVM-T	0.635 8	0.623 9	0.601 9	0.596 9	0.581 1
TL-SVM	0.948 4	0.943 4	0.936 0	0.896 3	0.864 2

由表2可知, 若无迁移学习, 则分类精度不高且随着噪音水平的增加而降低. 使用TL-SVM进行迁移学习, 虽然能够提高分类精度, 但当噪音水平较低时能保持较好的精度, 若加大噪音则精度呈下降趋势, 因此尽管TL-SVM算法具有一定的鲁棒能力, 但不适用于噪音水平较高的场合.

2.2.3 实验小结

由上面实验可知, TL-SVM将源域知识传递给目标域进行迁移学习, 在源领域与目标域相似性较大的情况下是有效的. 且因直接使用源域知识, 故其时间、空间复杂度与SVM相同, 分别为$O(n^3)$、$O(n^2)$, 其中n为目标域样本规模. 而Tradaboost算法, 将源域、目标域一起训练, 通过调整源领域训练数据权重, 最终带权重的辅助训练数据作为额外的训练数据, 其时间复杂度高于TL-SVM算法.

3 结 论

本文针对训练域与测试域分布不一致时, 传统支持向量机不能对测试集作出精确预测的问题, 提出了TL-SVM算法, 继承了基于经验风险最小化框架最大间隔SVM的优点, 只需少量新目标数据域已标签样本, 通过学习相似源领域知识, 达到获得目标域较好分类模型的目的. 实验结果表明, 该方法具有迁移学习能力, 优于传统的SVM分类方法且与Tradaboost迁移算法效果相当. 本算法的局限性在于适用于源域与目标域相似性较大的场景.

参考文献(References)

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

[2] Tao J W, Chung F L, Wang S T. A kernel learning framework for domain adaptation learning[J]. Science China Information Sciences, 2012, 55(9): 1983-2007.

[3] 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.

[4] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction[C]. Proc of the 23rd National Conf on Artificial Intelligence. Menlo Park: AAAI Press, 2008: 677-682.

[5] Xie S, Fan W, Peng J, et al. Latent space domain transfer between high dimensional overlapping distributions[C]. Proc of the 18th Int Conf on World Wide Web. New York: ACM Press, 2009: 91-100.

[6] Dai W, Yang Q, Xue G, et al. Boosting for transfer learning[C]. Proc of the 24th Int Conf on Machine Learning. New York: ACM Press, 2007: 193-200.

[7] 洪佳明, 陈炳超, 印鉴. 一种结合半监督Boosting方法的迁移学习算法[J]. 小型微型计算机系统, 2011, 32(11): 2169-2173.
(Hong J M, Chen B C, Yin J. Transfer learning via semi-supervised Boosting method[J]. J of Chinese Computer Systems, 2011, 32(11): 2169-2173.)

[8] 陈德品. 基于迁移学习的跨领域排序学习算法研究[D]. 合肥: 中国科学技术大学计算机科学与技术学院, 2010.
(Chen D P. Knowledge transfer for cross domain learning to rank[D]. Hefei: School of Computer Science and Technology, University of Science and Technology of China, 2010.)

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

[10] Pal M, Foody G M. Feature selection for classification of hyper spectral data by SVM[J]. IEEE Trans on Geoscience and Remote Sensing, 2010, 48(5): 2297-2307.

[11] 邓乃杨, 田英杰. 支持向量机——理论、算法与拓展[M]. 北京: 科学出版杜, 2009: 164-223.
(Deng N Y, Tian Y J. Support vector machine — Theory, algorithm and extension[M]. Beijing: Chinese Science Press, 2009: 164-223.)

[12] 王骏, 王士同, 王晓明. 基于特征加权距离的双指数模糊子空间聚类算法[J]. 控制与决策, 2010, 25(8): 1207-1210.
(Wang J, Wang S T, Wang X M. Double-indices fuzzy subspace clustering algorithm based on feature weighted distance[J]. Control and Decision, 2010, 25(8): 1207-1210.)

(责任编辑: 李君玲)