DOI:10.7544/issn1000-1239.2016.20148263
53(3):632-643,2016

基于迁移共享空间的分类新算法

董爱美[1,2] 毕安琪[1] 王士同[1]
[1](江南大学数字媒体学院 江苏无锡 214122)
[2](齐鲁工业大学信息学院 济南 250353)
(amdong@163.com)

A Classification Method Using Transferring Shared Subspace

Dong Aimei[1,2], Bi Anqi[1], and Wang Shitong[1]
[1](*School of Digital Media, Jiangnan University, Wuxi, Jiangsu* 214122)
[2](*School of Information, Qilu University of Technology, Jinan* 250353)

Abstract Transfer learning algorithms have been proved efficiently in pattern classification filed. The characteristic of transfer learning is to better use one domain information to improve the classification performance in different but related domains. In order to effectively solve the classification problems with a few labeled and abundant unlabeled data coming from different but related domains, a new algorithm named transferring shared subspace support vector machine (TS^3VM) is proposed in this paper. Firstly a shared subspace used as the common knowledge between source domain and target domain is built and then classical support vector machine method is introduced to the subspace for the labeled data, therefore the resulting classification model has the ability of transfer learning. Specifically, using the theory of transfer learning and the principal of large margin classifier, the proposed algorithm constructs a shared subspace between two domains by maximizing the joint probability distribution of the labeled and unlabeled data. Meanwhile, in order to fully consider the distribution of the few labeled data, the classification model is trained in the augmented feature space consisting of the original space and the shared subspace. Experimental results confirm the efficiency of the proposed method.

Key words shared subspace; transfer learning; support vector machine; joint probability distribution; large margin classifier

摘 要 为解决来自不同但相关领域的大量无标签数据和少量带标签数据的分类问题,首先构造一个联系源域到目标域的共享特征空间,并将该空间引入经典的支持向量机算法使其获得迁移能力,最终得到一种新的基于支持向量机的迁移共享空间的分类新算法,即迁移共享空间支持向量机.具体地,该方法以迁移学习理论为基础,结合分类器最大间隔原理,通过最大化无标签数据和带标签数据的联合概率分布来构建无标签数据和带标签数据的共享空间;为充分考虑少量带标签数据之数据分布,在其原始特征空间和共享空间组成的扩展空间中训练分类模型.相关实验结果验证了该迁移学习分类器的有效性.

关键词 共享空间;迁移学习;支持向量机;联合概率分布;大间隔分类器

中图法分类号 TP391.4

收稿日期:2014-11-17;**修回日期**:2015-06-03

基金项目:国家自然科学基金项目(61170122,61202311);山东省高等学校科技计划基金项目(J14LN05)
This work was supported by the National Natural Science Foundation of China (61170122,61202311) and the Project of Shandong Province Higher Educational Science and Technology Program (J14LN05).

随着信息技术的快速发展和互联网的日益普及,人们获得的信息呈现出越来越复杂的变化,例如绝大部分信息不带有标签,而少部分信息带有标签;并且大量的无标签信息和少量的带标签信息又来自不同但又有一定关系的问题领域.而传统半监督学习方法研究对象的特点是大量无标签数据和少量带标签数据来自相同的问题领域并且数据分布相同,那么如何充分利用这些来自不同但相关问题领域的大量无标签数据和少量带标签数据来训练模式分类模型是机器学习重点研究的问题之一.

为了解决上述问题,迁移学习得以提出,其旨在解决2个不同但相关领域的机器学习问题,且放松了训练数据和测试数据独立同分布之要求[1].迁移学习的关键是尽最大可能缩小源域和目标域数据间的差异,来寻找源域和目标域数据之间的"共同点"作为2者间的"桥梁",从而实现从源域到目标域的知识迁移.近年来,不同的学者主要从2个角度来寻找源域和目标域间的"共同知识"[2-11]:数据样本和数据特征.前者通过样本加权方法从源域寻找与目标域样本相似度最大的样本来作为迁移的知识,后者通过各种不同方式学习源域和目标域的共享特征表示来作为迁移的知识.学习源域和目标域的共享特征表示的方法主要有:Pan等人[2]在2008年通过降维方式学习得到领域间共享的特征表示从而实现迁移分类:首先通过最大均值差异嵌入(maximum mean discrepancy embedding, MMDE)进行降维得到一个低维的映射空间,并在此空间中使得源域数据和目标域数据的均值中心对齐,以减小2个领域的差异,经过这种处理后的数据可以直接利用传统分类器对数据进行训练和泛化;Xie等人[3]在2009年同样也是通过降维的方法研究了特征缺失下的迁移问题,先通过填补缺失值将领域间的条件分布变得一致,然后利用降维的方法将领域间的边缘分布变得一致,最后利用一些传统分类器实现了利用源域判别信息对目标域数据的分类;Quanz等人[4]在2009年以正则风险最小化思想为基础,结合最大均值差(maximum mean discrepancy, MMD)方法,提出一种基于特征空间的大间隔直推式迁移学习方法(large margin projected transductive support vector machine, LMPROJ),该方法是以经验风险正则化分类为框架,通过寻求一个特征变换来缩小源域和目标域之间的差异达到迁移分类之目的;Pan等人[5]在2010年提出将不同问题领域中的特征分为领域相关特征和领域独立特征,通过领域相关特征建立领域间联系实现领域间的迁移,再基于谱图划分对数据进行分类;Pan等人[6]在2011年进一步基于降维的思想,提出了快速特征提取算法TCA来学习一个低维的映射空间,对MMDE算法在计算复杂度方面进行了可观的改善;Zhuang等人[8]在2012年针对跨领域文本分类问题特有的共性和个性特征提出了CD-PLSA迁移学习算法,并且应用到多源域和多目标域问题中;Shao等人[9]在2012年使用低秩表示学习低维共享子空间技术,并提出了低秩迁移子空间学习算法;Gupta等人[10]在2013年采用非负矩阵分解思想提出了正则化的共享空间迁移学习框架,在一定程度上提高了模型的学习性能.

纵观基于寻找源域和目标域的共享特征表示的迁移学习分类方法,尽管其取得了一定效果,但也存在不足之处:1)共享特征表示是在少量带标签数据原始空间基础上得到的,那么如果这些少量带标签数据受到噪声干扰,则会直接影响其分布情况,从而导致学习到的共享特征表示有误,最终使得迁移学习分类模型的性能降低;2)现有的基于共享特征表示的迁移学习分类方法仅仅考虑了领域间的共同特征表示,没有考虑目标域数据的分布情况,由于源域不带标签数据数量极大从而对共享特征表示的学习起着决定性的作用,这就导致仅仅在共享特征空间学习到的分类模型对于目标域数据来说是不完善的.结合这2方面的考虑,本文提出了迁移共享空间支持向量机算法:以分类器最大间隔原理为指导思想,通过最大化源域大量无标签数据和目标域少量带标签数据的联合概率分布来构建源域和目标域间的共享特征空间;为充分考虑少量带标签数据的分布情况及带标签数据类标可能受到攻击情况,在其原始特征空间和共享特征空间组成的扩展特征空间中训练分类模型.因为分类模型的训练既不仅仅是在原始特征空间也不仅仅是在共享特征空间中,从而既避免了原始类标受到噪声干扰又充分考虑了少

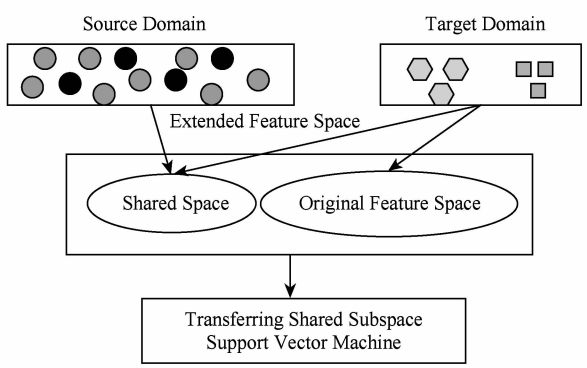


Fig. 1 The principal of TS^{3}VM.

图1 TS^{3}VM原理图

量带标签数据的数据分布. 迁移共享特征空间支持向量机算法(transferring shared subspace support vector machine, TS^{3}VM)原理如图 1 所示.

1 迁移共享空间支持向量机算法 TS^{3}VM

1.1 问题描述

本文提出的基于迁移共享空间的分类新方法的研究场景为:源域数据 $D_s=\{\boldsymbol{x}_i^s,\boldsymbol{x}_i^s\in\boldsymbol{X}^s,i=1,2,\cdots,n_s\}$,目标域数据 $D_t=\{(\boldsymbol{x}_j^t,y_j^t)\in\boldsymbol{X}^t\times\boldsymbol{Y}^t:j=1,2,\cdots,n_t\}$,其中 $\boldsymbol{x}_i^s\in\mathbb{R}^d,\boldsymbol{x}_j^t\in\mathbb{R}^d,y_j^t\in\{-1,+1\}$,$n_s\gg n_t$;学习任务是由源域的大量无标签数据来辅助目标域的少量带标签数据来为目标域学习到一个高效的分类器且线性分类器形式为 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$,其中 $\boldsymbol{w}$ 表示样本 $\boldsymbol{x}$ 在其所在特征空间的分类超平面参数.

1.2 TS^{3}VM 算法原理和目标函数

对于源域和目标域数据,根据 Parzen 窗法分别得到其概率分布的核密度估计函数如下:

$$P(D_s)=\hat{P}_{D_s}(\boldsymbol{x})=\frac{1}{n_s\sqrt{2\pi}\sigma_s}\sum_{i=1}^{n_s}\mathrm{e}^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i^s\|^2}{2\sigma_s^2}},\quad(1)$$

$$P(D_t)=\hat{P}_{D_t}(\boldsymbol{x})=\frac{1}{n_t\sqrt{2\pi}\sigma_t}\sum_{i=1}^{n_t}\mathrm{e}^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j^t\|^2}{2\sigma_t^2}},\quad(2)$$

其中,σ_s,σ_t 是窗宽参数且 $\sigma_s>0,\sigma_t>0$. 将源域和目标域数据 $\boldsymbol{x}_i^s,\boldsymbol{x}_j^t$ 由原来的 d 维特征扩展为 $(d+r)$ 维特征,原始数据 $\boldsymbol{x}_i^s,\boldsymbol{x}_j^t$ 在扩展的特征空间中对应的新数据为 $\tilde{\boldsymbol{x}}_i^s,\tilde{\boldsymbol{x}}_j^t$,其中 $\tilde{\boldsymbol{x}}_i^s\in\mathbb{R}^{d+r},\tilde{\boldsymbol{x}}_j^t\in\mathbb{R}^{d+r}$,且扩展的 r 维特征空间是源域和目标域数据通过共同的变换矩阵参数 $\boldsymbol{\Theta}$ 投影得到的. 假设在原始特征空间中目标域样本的分类超平面参数为 $\boldsymbol{w},\boldsymbol{w}\in\mathbb{R}^d$,$\boldsymbol{\Theta}$ 是对源域和目标域数据特征进行共同的正交变换矩阵参数,$\boldsymbol{\Theta}\in\mathbb{R}^{r\times d}$,那么 $\tilde{\boldsymbol{x}}_i^s,\tilde{\boldsymbol{x}}_j^t$ 可表示为 $\boldsymbol{\Theta}\boldsymbol{x}_i^s,\boldsymbol{\Theta}\boldsymbol{x}_j^t$,假设 $\boldsymbol{x}_i^s,\boldsymbol{x}_j^t$ 在 r 维的共享征空间中分类超平面参数为 $\boldsymbol{v}$,$\boldsymbol{v}\in\mathbb{R}^r$. 那么在扩展的 r 维特征空间中源域和目标域数据根据 Parzen 窗原理得到的概率分布核密度估计函数分别为

$$P(D_s)=\hat{P}_{D_s}(\boldsymbol{\Theta}\boldsymbol{x}_i^s)=\frac{1}{n_s\sqrt{2\pi}\sigma_s}\sum_{i=1}^{n_s}\mathrm{e}^{-\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{x}_i^s)(\boldsymbol{x}-\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}}{2\sigma_s^2}},\quad(3)$$

$$P(D_t)=\hat{P}_{D_t}(\boldsymbol{\Theta}\boldsymbol{x}_j^t)=\frac{1}{n_t\sqrt{2\pi}\sigma_t}\sum_{j=1}^{n_t}\mathrm{e}^{-\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{x}_j^t)(\boldsymbol{x}-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}}{2\sigma_t^2}}.\quad(4)$$

假设源域和目标域数据相互独立,那么在通过正交变换参数 $\boldsymbol{\Theta}$ 投影后在 r 维的特征空间中亦最大可能满足联合分布独立性假设,所以由式(3)(4)得到在 r 维的特征空间中源域和目标域数据的联合概率分布为

$$\hat{P}_{D_{s,t}}(\boldsymbol{\Theta}\boldsymbol{x})=\hat{P}_{D_s}(\boldsymbol{\Theta}\boldsymbol{x}^s)\hat{P}_{D_t}(\boldsymbol{\Theta}\boldsymbol{x}^t).\quad(5)$$

令:

$$G(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}},\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i,\sigma^2)=G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_i,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{x}_i)(\boldsymbol{x}-\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}}{2\sigma^2}},$$

那么 $\hat{P}_{D_s}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^s)$ 和 $\hat{P}_{D_t}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^t)$ 可表达为如下形式:

$$\begin{aligned}&\hat{P}_{D_s}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^s)=\hat{P}_{D_s}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}^s)=\\&\frac{1}{n_s}\sum_{i=1}^{n_s}G(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^s,\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i^s,\sigma^2)=\\&\frac{1}{n_s}\sum_{i=1}^{n_s}G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}^s,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_i^s,\sigma^2),\end{aligned}\quad(6)$$

$$\begin{aligned}&\hat{P}_{D_t}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^t)=\hat{P}_{D_t}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}^t)=\\&\frac{1}{n_t}\sum_{j=1}^{n_t}G(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}^t,\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}}_j^t,\sigma^2)=\\&\frac{1}{n_t}\sum_{j=1}^{n_t}G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}^t,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_j^t,\sigma^2).\end{aligned}\quad(7)$$

因为

$$\int G(\boldsymbol{x},\boldsymbol{x}_i^s,\sigma_s^2)G(\boldsymbol{x},\boldsymbol{x}_j^t,\sigma_t^2)\mathrm{d}x=G(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t,\sigma_s^2+\sigma_t^2),$$

所以有:

$$\begin{aligned}&\int\hat{P}_{D_s,D_t}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}})\mathrm{d}x=\int\hat{P}_{D_s,D_t}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x})\mathrm{d}x=\\&\int\frac{1}{n_s}\sum_{i=1}^{n_s}G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_i^s,\sigma^2)\times\\&\frac{1}{n_t}\sum_{j=1}^{n_t}G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_j^t,\sigma^2)\mathrm{d}x=\\&\frac{1}{n_s\times n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\int G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_i^s,\sigma^2)\times\\&G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x},\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_j^t,\sigma^2)\mathrm{d}x=\\&\frac{1}{n_s\times n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_i^s,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\boldsymbol{x}_j^t,2\sigma^2)=\\&\frac{1}{n_s\times n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\frac{1}{2\sqrt{\pi}\sigma}\mathrm{e}^{-\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}}{4\sigma^2}}.\end{aligned}$$

在 r 维的特征空间中要求源域和目标域数据的联合概率分布取得最大值,即要求最大化 $\int\hat{P}_{D_s,D_t}(\boldsymbol{v}^{\mathrm{T}}\tilde{\boldsymbol{x}})\mathrm{d}x$. 因此得到目标函数:

$$\arg\max_{\boldsymbol{v},\boldsymbol{\Theta}} J_1 = \arg\max_{\boldsymbol{v},\boldsymbol{\Theta}} \frac{1}{n_s \times n_t} \times \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \frac{1}{2\sqrt{\pi}\sigma} e^{-\frac{\boldsymbol{v}^T\boldsymbol{\Theta}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v}}{4\sigma^2}}$$

$$\text{s.t. } \boldsymbol{\Theta}\boldsymbol{\Theta}^T = \boldsymbol{I}_{r\times r}. \quad (8)$$

对于式(8)直接求解比较困难,利用泰勒展开定理进行近似变形得到:

$$\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \frac{1}{2\sqrt{\pi}\sigma} e^{-\frac{\boldsymbol{v}^T\boldsymbol{\Theta}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v}}{4\sigma^2}} \approx \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \frac{1}{2\sqrt{\pi}\sigma} \times \left(1-\frac{1}{4\sigma^2}\boldsymbol{v}^T\boldsymbol{\Theta}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v}\right), \quad (9)$$

结合式(8)(9)得到目标函数:

$$\arg\min_{\boldsymbol{v},\boldsymbol{\Theta}} J_2 = \arg\min_{\boldsymbol{v},\boldsymbol{\Theta}} \boldsymbol{v}^T\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v}$$

$$\text{s.t. } \boldsymbol{\Theta}\boldsymbol{\Theta}^T = \boldsymbol{I}_{r\times r}, \quad (10)$$

假设目标域数据在扩展后的总特征空间中分类超平面形式为

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}^t + \boldsymbol{v}^T\bar{\boldsymbol{x}}^t = \boldsymbol{w}^T\boldsymbol{x}^t + \boldsymbol{v}^T\boldsymbol{\Theta}\boldsymbol{x}^t = (\boldsymbol{w}^T + \boldsymbol{v}^T\boldsymbol{\Theta})\boldsymbol{x}^t, \quad (11)$$

其中,$\boldsymbol{x}^t$,$\bar{\boldsymbol{x}}^t$ 分别表示原始特征空间和 r 维共享空间中的目标域数据,$\boldsymbol{w}$,$\boldsymbol{v}$ 分别表示目标域数据 $\boldsymbol{x}^t$,$\bar{\boldsymbol{x}}^t$ 在其对应特征空间中的分类超平面参数.

根据支持向量机最大间隔分类原则,同时要求源域和目标域数据在扩展的 r 维空间中联合概率分布最大,且对目标域数据在扩展后的特征空间中施加分类误差约束,得到 TS^{3}VM 目标函数:

$$\arg\min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{\Theta}} J_3 = \arg\min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda_1}{2}\|\boldsymbol{w}+\boldsymbol{\Theta}^T\boldsymbol{v}\|^2 + \frac{C}{2}\sum_{j=1}^{n_t}\xi_j^2 + \frac{\lambda_2}{2}\boldsymbol{v}^T\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v},$$

$$\text{s.t. } y_j^t(\boldsymbol{w}^T + \boldsymbol{v}^T\boldsymbol{\Theta})\boldsymbol{x}_j^t = 1-\xi_j, j = 1,2,\cdots,n_t,$$

$$\boldsymbol{\Theta}\boldsymbol{\Theta}^T = \boldsymbol{I}_{r\times r}. \quad (12)$$

针对式(12),作如下说明:

1) 式(12)中第 1 项和第 2 项是结构风险项;第 3 项是经验风险项;第 4 项表示源域和目标域数据在扩展的 r 维的特征空间中的联合概率分布;通过对这 4 项的平衡期望达到最小的结构化风险.

2) 由式(12)可以看出,目标域数据的类标签仅仅用来维数扩展,从而避免了误标对最终分类模型的影响,使得最终分类模型有较强的抗噪性.

3) 由式(12)可以看出,在扩展后的特征空间中对目标域数据施加分类误差约束,既充分利用源域和目标域间的共性"知识"来获得正迁移效果,又充分考虑了目标域数据的数据分布,使得最终分类模型在目标域数据上性能最优,这也是本文采用扩维思想而不是采用传统降维思想来进行迁移分类的动机之一.

4) 从目标函数的构造过程可以看出,TS^{3}VM 方法在目标域数据数量远远小于源域数据数量且目标域数据类标受到攻击而产生误标情况下,具有很好的鲁棒性.

1.3 TS^{3}VM 算法目标函数参数学习规则

在不引起混淆的情况下,把约束项中的 y_j^t 记为 y_j,$\boldsymbol{x}_j^t$ 记为 $\boldsymbol{x}_j$,且为便于目标函数式(12)求解,引入辅助变量 $\boldsymbol{h}$,令 $\boldsymbol{h}=\boldsymbol{w}+\boldsymbol{\Theta}^T\boldsymbol{v}$,则式(12)变为

$$\arg\min_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{\Theta}} J_4 = \arg\min_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^T\boldsymbol{v}\|^2 + \frac{\lambda_1}{2}\|\boldsymbol{h}\|^2 + \frac{C}{2}\sum_{j=1}^{n_t}\xi_j^2 + \frac{\lambda_2}{2}\boldsymbol{v}^T\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^T\boldsymbol{\Theta}^T\boldsymbol{v}$$

$$\text{s.t. } y_j\boldsymbol{h}^T\boldsymbol{x}_j = 1-\xi_j, j = 1,2,\cdots,n_t,$$

$$\boldsymbol{\Theta}\boldsymbol{\Theta}^T = \boldsymbol{I}_{r\times r}. \quad (13)$$

其中有 3 个变量 $\boldsymbol{h}$,$\boldsymbol{v}$,$\boldsymbol{\Theta}$ 需要同时优化.直接对这些参数优化是比较困难的.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[12-15]对式(13)进行参数优化.在该迭代过程中包含 3 个主要步骤:

1) 固定($\{\boldsymbol{\Theta},\boldsymbol{v}\}$),优化式(13)学习得到$\{\boldsymbol{h}\}$;

2) 固定($\{\boldsymbol{\Theta},\boldsymbol{h}\}$),优化式(13)学习得到$\{\boldsymbol{v}\}$;

3) 固定($\{\boldsymbol{v},\boldsymbol{h}\}$),优化式(13)学习得到$\{\boldsymbol{\Theta}\}$.

下面通过 3 个定理给出目标函数中的 3 个变量的学习规则:

定理 1. 假设变量 $\boldsymbol{v}$,$\boldsymbol{\Theta}$ 固定,则式(13)中变量 $\boldsymbol{h}$ 的最优值可以通过对一个线性方程组求解得到.

证明. 假设变量 $\boldsymbol{v}$,$\boldsymbol{\Theta}$ 固定,当优化变量 $\boldsymbol{h}$ 时,式(13)可写为

$$\arg\min_{\boldsymbol{h}} J_5 = \arg\min_{\boldsymbol{h}} \frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^T\boldsymbol{v}\|^2 + \frac{\lambda_1}{2}\|\boldsymbol{h}\|^2 + \frac{C}{2}\sum_{j=1}^{n_t}\xi_j^2$$

$$\text{s.t. } y_j\boldsymbol{h}^T\boldsymbol{x}_j = 1-\xi_j, j = 1,2,\cdots,n_t, \quad (14)$$

构造目标函数 J_5 的拉格朗日函数:

$$L = \frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^T\boldsymbol{v}\|^2 + \frac{\lambda_1}{2}\|\boldsymbol{h}\|^2 + \frac{C}{2}\sum_{j=1}^{n_t}\xi_j^2 - \sum_{j=1}^{n_t}\alpha_j[y_j\boldsymbol{h}^T\boldsymbol{x}_j - 1 + \xi_j], \quad (15)$$

其中,α_j 为拉格朗日乘子且 $\alpha_j>0, j=1,2,\cdots,n_t$,式(14)取得极值的必要条件为式(15)对 $\boldsymbol{h}$ 和 $\xi_j(j=1,2,\cdots,n_t)$ 的偏导数为 0,于是得到:

$$h=\frac{\sum_{j=1}^{n_t}\alpha_j y_j \boldsymbol{x}_j+\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}}{1+\lambda_1},$$

$$\alpha_j=C\xi_j,$$

$$y_j\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}_j-1+\xi_j=0, j=1,2,\cdots,n_t, \quad (16)$$

式(16)写成线性方程组的形式为

$$\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} & -\boldsymbol{Z}^{\mathrm{T}}/(1+\lambda_1)\\ \boldsymbol{0} & C\boldsymbol{I} & -\boldsymbol{I}\\ \boldsymbol{Z} & \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{\xi}\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}/(1+\lambda_1)\\ \boldsymbol{0}\\ \boldsymbol{1}\end{bmatrix}, \quad (17)$$

其中,$\boldsymbol{Z}=[\boldsymbol{x}_1^{\mathrm{T}}y_1,\boldsymbol{x}_2^{\mathrm{T}}y_2,\cdots,\boldsymbol{x}_n^{\mathrm{T}}y_n]$. 从而,变量 $\boldsymbol{h}$ 的最优值可以通过对一个线性方程组求解得到. 证毕.

定理 2. 假设变量 $\boldsymbol{h},\boldsymbol{\Theta}$ 固定,则式(13)中变量 $\boldsymbol{v}$ 是一个解析解,具体形式为

$$\boldsymbol{v}=\boldsymbol{\Theta}(\boldsymbol{I}_{d\times d}-\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}})^{-1}\boldsymbol{h}. \quad (18)$$

证明. 假设变量 $\boldsymbol{h},\boldsymbol{\Theta}$ 固定,式(13)变为

$$\arg\min_{\boldsymbol{v}} J_7=\arg\min_{\boldsymbol{v}}\frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\|^2+\frac{\lambda_2}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}, \quad (19)$$

目标函数 J_7 取最小值的必要条件是其对 $\boldsymbol{v}$ 的偏导数为 $\boldsymbol{0}_{r\times 1}$,即:

$$\frac{\partial L}{\partial v}=\boldsymbol{\Theta}(\boldsymbol{h}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v})+\lambda_2\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}=\boldsymbol{0}_{r\times 1}.$$

从而有:

$$\boldsymbol{v}=\boldsymbol{\Theta}(\boldsymbol{I}_{d\times d}-\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}})^{-1}\boldsymbol{h},$$

此时 $\boldsymbol{v}$ 是一个解析解. 证毕.

定理 3. 假设变量 $\boldsymbol{h},\boldsymbol{v}$ 固定,则式(13)中变量 $\boldsymbol{\Theta}$ 的求解可以用梯度下降法求解得到,并且梯度下降法中的步长具有解析解的形式.

证明. 假设变量 $\boldsymbol{h},\boldsymbol{v}$ 固定,式(13)变为

$$\arg\min_{\boldsymbol{\Theta}} J_8=\arg\min_{\boldsymbol{\Theta}}\frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\|^2+\frac{\lambda_2}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}$$

$$\text{s.t. } \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\boldsymbol{I}_{r\times r}, \quad (20)$$

假设隐变量的域记为 $\Omega^{\perp}$,则 $\Omega^{\perp}=\{\boldsymbol{\Theta}\subset\mathbb{R}^{r\times d}\}\cap\{\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\boldsymbol{I}_{r\times r}\}$,因为 $\Omega^{\perp}$ 中的约束条件 $\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\boldsymbol{I}_{r\times r}$ 和正则化项具有相似的作用,所以可以把式(20)中的约束条件去掉而得到一个形式简单的优化问题. 目标函数式 J_8 对变量 $\boldsymbol{\Theta}^{\mathrm{T}}$ 求偏导数为

$$\frac{\partial J_8}{\partial\boldsymbol{\Theta}^{\mathrm{T}}}=\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}-\boldsymbol{h}\boldsymbol{v}^{\mathrm{T}}+\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}, \quad (21)$$

则变量 $\boldsymbol{\Theta}^{\mathrm{T}}$ 可以用梯度下降法求解得到,梯度下降规则为

$$\boldsymbol{\Theta}^{\mathrm{T}}\leftarrow\boldsymbol{\Theta}^{\mathrm{T}}-\gamma(\boldsymbol{I}_{d\times d}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta})\frac{\partial J_8}{\partial\boldsymbol{\Theta}^{\mathrm{T}}}=\boldsymbol{\Theta}^{\mathrm{T}}-\gamma\nabla\boldsymbol{\Theta}^{\mathrm{T}}, \quad (22)$$

为得到变量 $\boldsymbol{\Theta}^{\mathrm{T}}$ 的步长 γ 的解析解,把式(21)代入到式(22)中,得到:

$$\boldsymbol{\Theta}^{\mathrm{T}}\leftarrow\boldsymbol{\Theta}^{\mathrm{T}}-\gamma(\boldsymbol{I}_{d\times d}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta})\times(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}-\boldsymbol{h}\boldsymbol{v}^{\mathrm{T}}+\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}). \quad (23)$$

从而有:

$$\boldsymbol{\Theta}\leftarrow\boldsymbol{\Theta}-\gamma(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}-\boldsymbol{v}\boldsymbol{h}^{\mathrm{T}}+\lambda_2\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\times(\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}})^{\mathrm{T}})\times(\boldsymbol{I}_{d\times d}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}). \quad (24)$$

进一步,令:

$$\boldsymbol{M}=(\boldsymbol{I}_{d\times d}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta})(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}-\boldsymbol{h}\boldsymbol{v}^{\mathrm{T}}+\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}), \quad (25)$$

则有:

$$\boldsymbol{\Theta}^{\mathrm{T}}\leftarrow\boldsymbol{\Theta}^{\mathrm{T}}-\gamma\boldsymbol{M}, \quad (26)$$

$$\boldsymbol{\Theta}\leftarrow\boldsymbol{\Theta}-\gamma\boldsymbol{M}^{\mathrm{T}}, \quad (27)$$

把式(26)(27)代回式(20),得到关于 γ 的函数:

$$g(\gamma)=\frac{1}{2}\|\boldsymbol{h}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}+\gamma\boldsymbol{M}\boldsymbol{v}\|^2+\frac{\lambda_2}{2}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}-\gamma\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v}-\gamma\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}+\gamma^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})(\boldsymbol{x}_i^{\mathrm{s}}-\boldsymbol{x}_j^{\mathrm{t}})^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v}), \quad (28)$$

记:

$$\boldsymbol{t}_1=\boldsymbol{h}-\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v},$$

$$\boldsymbol{t}_2=\boldsymbol{M}\boldsymbol{v},$$

$$t_3 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v},$$

$$t_4 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v},$$

$$t_5 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v}+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v},$$

则式(28)变为

$$g(\gamma)=\frac{1}{2}\|\boldsymbol{t}_1+\gamma\boldsymbol{t}_2\|^2+\lambda_2\gamma^2 t_4-\lambda_2\gamma t_5+\lambda_2 t_3, \tag{29}$$

式(29)对变量 γ 求偏导令其结果为 0,得到关于步长 γ 的解析解形式为 $\gamma=\dfrac{\lambda_2 t_5-tr(\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{t}_2)}{2\lambda_2 t_4+tr(\boldsymbol{t}_2^{\mathrm{T}}\boldsymbol{t}_2)}$. 证毕.

1.4 TS³VM 算法描述

根据 1.1~1.3 节分析和推导,给出 TS³VM 算法描述.

算法 1. TS³VM 算法.

参数:源域和目标域数据特征扩展的维数 r、参数 C,λ_1,λ_2;

输入:目标域样本$\{(\boldsymbol{x}_j^t,y_j^t)\mid j=1,2,\cdots,n_t\}$、源域样本$\{(\boldsymbol{x}_i^s)\mid i=1,2,\cdots,n_s\}$,其中,$n_s\geqslant n_t$;

输出:变换参数矩阵 $\boldsymbol{\Theta}\in\mathbb{R}^{r\times d}$、目标域原始特征空间分类超平面参数 $\boldsymbol{w}$、扩展的 r 维特征空间分类超平面参数 $\boldsymbol{v}$.

初始化:参数 $\boldsymbol{h}_0$,$\boldsymbol{\Theta}_0$,其中 $\boldsymbol{h}_0\in\mathbb{R}^d$,$\boldsymbol{\Theta}_0\in\mathbb{R}^{r\times d}$(满足$\boldsymbol{\Theta}_0\boldsymbol{\Theta}_0^{\mathrm{T}}=\boldsymbol{I}_{r\times r}$),当前迭代值 $iter=0$,设置最大迭代次数 $itermax$ 和误差阈值 ε_1.

Repeat

Step1. 根据定理 2 计算 $\boldsymbol{v}_{\mathrm{iter}}$;

Step2. $iter=iter+1$;

Step3. 根据定理 3 采用梯度下降法计算 $\boldsymbol{\Theta}_{\mathrm{iter}}\in\mathbb{R}^{r\times d}$:初始化 $\boldsymbol{\Theta}_{\mathrm{iter}}^0\in\mathbb{R}^{r\times d}$,满足 $\boldsymbol{\Theta}_{\mathrm{iter}}^0(\boldsymbol{\Theta}_{\mathrm{iter}}^0)^{\mathrm{T}}=\boldsymbol{I}_{r\times r}$,当前迭代值 $t=0$,最大迭代次数 $t_{\max}$ 和误差阈值 ε_2;

Repeat $t=t+1$;根据式(22)计算 $\boldsymbol{\Theta}_{\mathrm{iter}}^t$;

Until $\|\boldsymbol{\Theta}_{\mathrm{iter}}^t-\boldsymbol{\Theta}_{\mathrm{iter}}^{t-1}\|<\varepsilon_2$ or $(t>t_{\max})$令 $\boldsymbol{\Theta}_{\mathrm{iter}}=\boldsymbol{\Theta}_{\mathrm{iter}}^t$;

Step4. 根据定理 1 计算 $\boldsymbol{h}_{\mathrm{iter}}$;

Until $\|\boldsymbol{h}_{\mathrm{iter}}-\boldsymbol{h}_{\mathrm{iter}-1}\|<\varepsilon_1$ or $\|\boldsymbol{v}_{\mathrm{iter}}-\boldsymbol{v}_{\mathrm{iter}-1}\|<\varepsilon_1$ or $\|\boldsymbol{\Theta}_{\mathrm{iter}}-\boldsymbol{\Theta}_{\mathrm{iter}-1}\|<\varepsilon_1$ or $(iter>iter_{\max})$.

1.5 TS³VM 算法收敛问题说明

对于 TS³VM 算法,以步骤 $iter+1$ 的迭代学习为例对其收敛性作如下分析:

1) 在步骤 $iter+1$,固定 $\boldsymbol{v}_{\mathrm{iter}}$,$\boldsymbol{\Theta}_{\mathrm{iter}}$不变,由定理 1 可知 $\boldsymbol{h}_{\mathrm{iter}+1}$是线性方程组的解,从而可以确保变量 $\boldsymbol{h}_{\mathrm{iter}+1}$在步骤 $iter+1$ 取得全局最优解.

2) 步骤 $iter+1$ 固定 $\boldsymbol{h}_{\mathrm{iter}+1}$,$\boldsymbol{\Theta}_{\mathrm{iter}}$,则 $\boldsymbol{v}_{\mathrm{iter}+1}=\boldsymbol{\Theta}_{\mathrm{iter}}(\boldsymbol{I}_{d\times d}-\lambda_2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}})^{-1}\boldsymbol{h}_{\mathrm{iter}+1}$,由于 $\boldsymbol{h}_{\mathrm{iter}+1}$,$\boldsymbol{\Theta}_{\mathrm{iter}}$在其当前迭代步骤中取得全局(或局部)最优解,从而可以确保变量 $\boldsymbol{v}$ 在步骤 $iter+1$ 取得全局最优解.

3) 在步骤 $iter+1$,固定 $\boldsymbol{h}_{\mathrm{iter}}$,$\boldsymbol{v}_{\mathrm{iter}}$不变,对式(20)进行迭代优化,根据梯度下降算法的性质可知所得到的解 $\boldsymbol{\Theta}_{\mathrm{iter}}$是式(20)的某个局部最优解的近似解,即 $J_8(\boldsymbol{\Theta}_{\mathrm{iter}+1})\leqslant J_8(\boldsymbol{\Theta}_{\mathrm{iter}})$.

据上述分析可知,TS³VM 算法并不能保证严格的收敛.交替迭代优化技术在智能模型学习方法中[9-12]经常被使用,如经典的 FCM 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的或是有条件收敛的,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[9-12].

正如文献[12]指出的: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.例如,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.

对 TS³VM 算法的说明有 2 点:

1) 在变量初始化阶段,变量 $\boldsymbol{h}$,$\boldsymbol{\Theta}$ 随机初始化,因此在实验部分进行多次实验取平均值作为最终实验结果,尽量减少随机初始化带来的不确定性.

2) 对于迭代过程中的收敛条件,在算法设计中一方面通过阈值来控制,即当连续多次(比如 10 次)更新迭代相应的变量不再发生变化时,达到收敛;另一方面为保险起见,设置了最大迭代次数,对算法能够进行的最大迭代次数进行限制,几个收敛条件满足一个即停止迭代.

2 实验分析

为验证 TS³VM 算法的有效性,在人造数据集

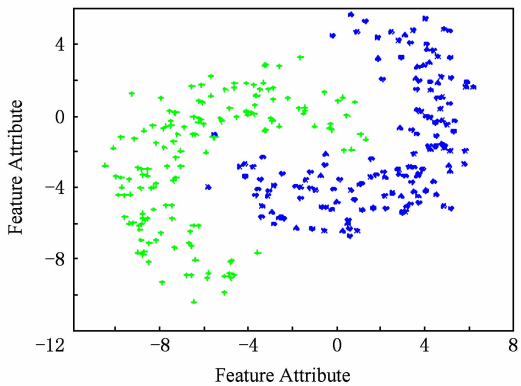
(2-moons)和3个文本数据集上(20NewsGroup[16], Reuters-21578[13], Email Spam[17])进行了实验. 采用比较算法有:SVM, TSVM[18], LMPROJ[4], TCA[6]. 采用SVM算法的目的是为了验证目标域中带标签数据数量非常少不足以训练一个高性能的分类器;算法TSVM是综合利用带标签数据和无标签数据的传统模式分类方法,采用其目的是为了验证在训练分类模型过程中引入迁移学习思想来综合利用带标签数据和无标签数据较传统模式分类方法优越;算法LMPROJ为迁移类算法,采用其目的是为了验证半监督的迁移学习方法和无监督的迁移学习方法的性能优劣;算法TCA为迁移类算法,采用其目的是为了验证在相同场景设置下,针对基于降维思想的迁移分类算法来说,本文所提的基于扩维的迁移分类算法表现出可比较的性能. 对于SVM,由libsvm[19]软件实现,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.

具体来说,对于算法TSVM,参数 C_1 和 C_2 的取值范围为 $\{10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},1,10,10^2,10^3,10^4,10^5\}$;对于算法LMPROJ,参数 λ 的取值范围为 $\{2^{10},2^{11},2^{12},2^{13},2^{14},2^{15},2^{16},2^{17},2^{18},2^{19},2^{20}\}$,参数 λ_2 的取值范围为 $\{2^{-5},2^{-4},2^{-3},2^{-2},2^{-1},1,2,2^2,2^3,2^4,2^5\}$,参数C的取值范围为 $\{2^{-6},2^{-5},2^{-4},2^{-3},2^{-2},2^{-1},1,2,2^2,2^3,2^4,2^5,2^6\}$;对于算法TCA,参数 μ 的取值范围为 $\{10^{-3},10^{-2},10^{-1},1,10,10^2,10^3\}$. 对于TS^{3}VM方法中的共享空间维数 r 需要人工确定,在当前主流的基于空间的分类方法中一般采用网格寻优或实验验证的方法确定,例如Ando[20]方法、Zheng[21]方法和Ji[22]方法. 对于文本数据集,本文借鉴Ando方法的思想,只关心共享特征空间维数 r 在10~100之间的范围{5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95,100};对于人造数据集,扩展维数固定为1. 参数 C 取值范围为 $\{10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},1,10,10^2,10^3,10^4,10^5\}$. 参数 λ_1,λ_2 的取值范围为 $\{2^{-6},2^{-5},2^{-4},2^{-3},2^{-2},2^{-1},1,2,2^2,2^3,2^4,2^5,2^6\}$. 所有参数 r,C,λ_1,λ_2 均通过网格搜索的方式确定最优值. 详细参数敏感性实验参见2.3节

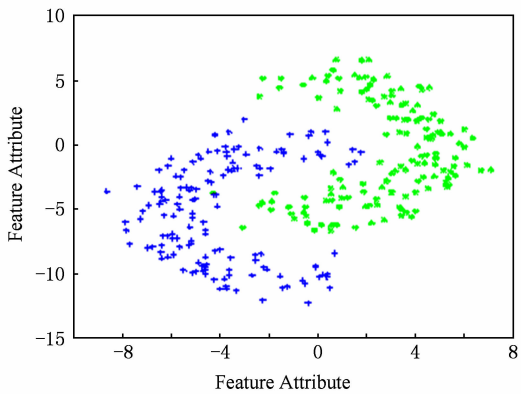
2.1 数据集描述与设置

1) 2-moons数据集:人工生成1个包含600个样本的双月形2维样本集作为源域数据集,正负类样本各为300个;将源域数据按逆时针方向分别旋转30°,45°,60°,得到目标域数据集. 另外,把目标域带标签数据的标签部分弄错,得到部分数据误标的目标域数据集,如图2所示. 为了适应TS^{3}VM算法之研究场景,对源域和目标域样本作如下处理:对于源域样本,去掉所有的标签信息;对于目标域样本,随机选择正负样本各30个作为训练样本. 另外为了验证TS^{3}VM算法对类标攻击的鲁棒性,把目标域中的部分训练样本类标弄错(误标样本比例设为20%).

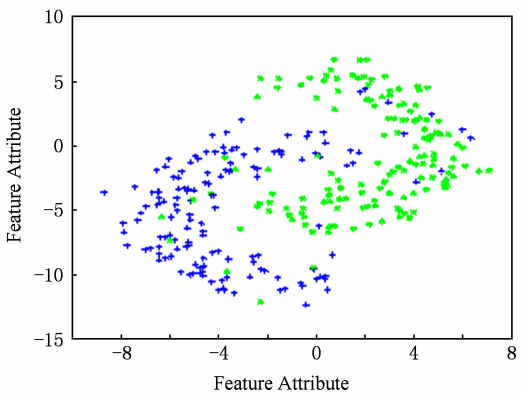
2) 文本数据集:20NewsGroup数据集包含近20 000新闻组的英文文档,大约分成20类,包含6个



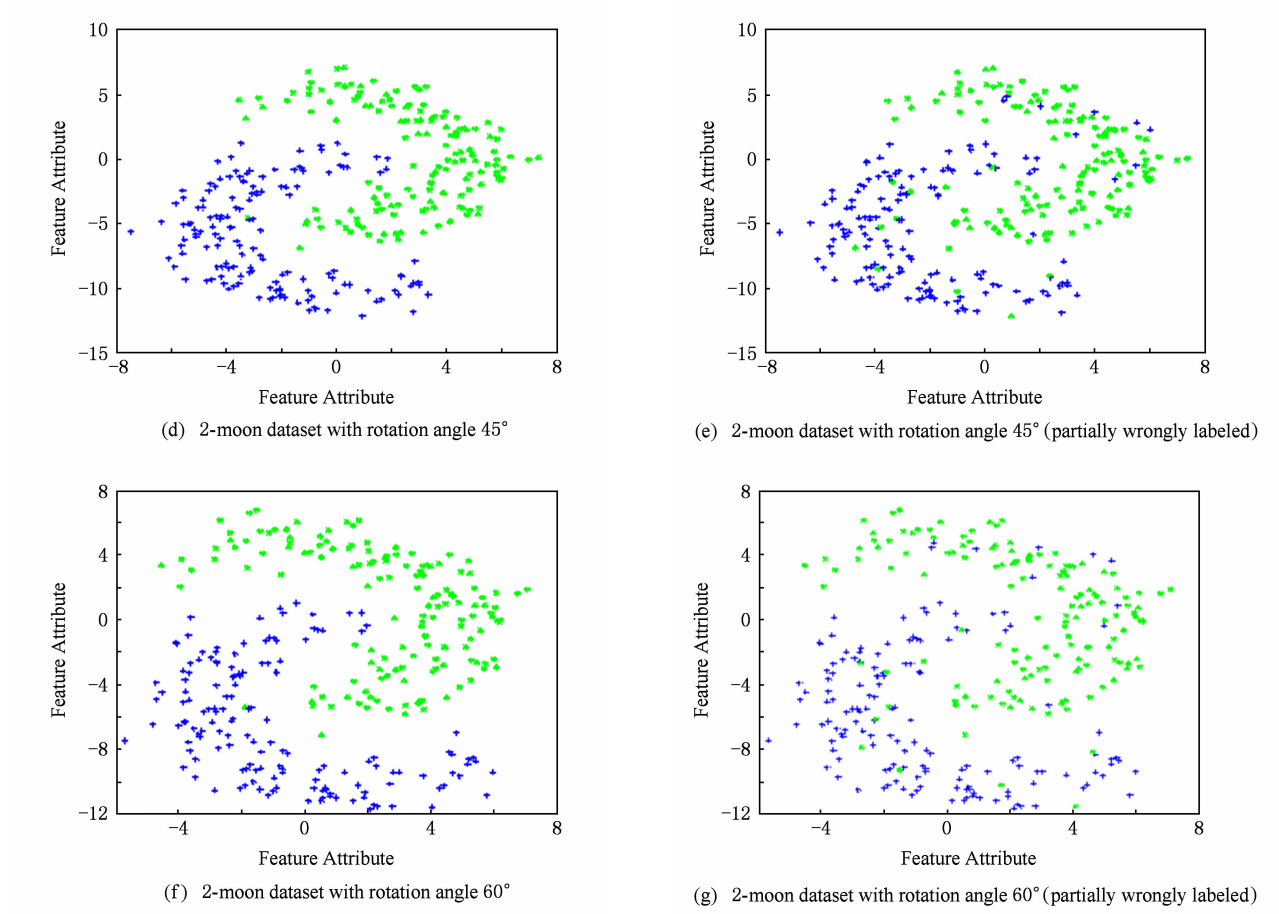
(a) Original 2-moon dataset



(b) 2-moon dataset with rotation angle 30°



(c) 2-moon dataset with rotation angle 30° (partially wrongly labeled)



(d) 2-moon dataset with rotation angle 45°

(e) 2-moon dataset with rotation angle 45° (partially wrongly labeled)

(f) 2-moon dataset with rotation angle 60°

(g) 2-moon dataset with rotation angle 60° (partially wrongly labeled)

Fig. 2 Four 2-moon datasets with different rotation degrees.

图 2 旋转不同角度的双月型数据集

顶层类别,每个顶层类别下分别包含若干个子类别. Reuter-21578 包含近 21 000 个文档集,分为 orgs, people,places 三个大类,每个类别下面包含了相应的子类别. 为了适应本文研究场景,对于 20News-Group 数据集:分别从顶层大类中抽取 4 个大类以构建学习数据集;对于 Reuter-21578 数据集:数据基于子类进行分割,分别从顶层大类中抽取 3 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不同的子类认为不同的领域. Email Spam 数据集包含 User1,User2,User3 这 3 个子集代表 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. 为适应本文研究场景设置,构造了 3 个迁移数据集. 详细信息见表 1 和表 2 所示.

Table 1 Transfer Learning Text Classification Datasets 20NewsGroup

表 1 迁移学习文本分类数据 20NewsGroup

ID	Dataset	Domain	Positive Data	Negative Data
NG1	comp vs. sci	Source	comp. graphics com. os. ms-windows. misc	sci. crypt sci. electronics
		Target	comp. sys. ibm. pc. hardware comp. sys. mac. hardware comp. windows. x	sci. med sci. space
NG2	rec vs. talk	Source	rec. autos rec. motorcycles	talk. politics. guns talk. politics. mideast
		Target	rec. sport. baseball rec. sport. hockey	talk. politics. misc talk. religion. misc

Continued (Table 1)

ID	Dataset	Domain	Positive Data	Negative Data
NG3	comp vs. talk	Source	comp. graphics comp. sys. mac. hardware comp. windows. x	talk. politics. mideast talk. religion. misc
		Target	comp. os. ma-windows. misc comp. sys. ibm. pa. hardware	talk. politics. guns talk. politics. misc
NG4	sci vs. talk	Source	sci. electronicis sci. med	talk. politics. misc talk. religion. misc
		Target	sci. crypt sci. space	talk. politics. guns talk. politics. mideast
NG5	comp vs. rec	Source	comp. graphics comp. sys. ibm. pc. hardware comp. sys. mac. hardware	rec. motorcycles rec. sport. hockey
		Target	comp. os. ms-windows. misc comp. windows. x	rec. autos rec. sport. baseball
NG6	rec vs. sci	Source	rec. autos rec. sport. baseball	sci. med sci. space
		Target	rec. motorcycles rec. sport. hockey	sci. crypt sci. electronics

Table 2 Transfer Learning Text Classification Datasets Reuter-21578 and Email Spam

表 2 迁移学习文本分类数据 Reuter-21578 和 Email Spam

Id	Dataset	Source Domain	Target Domain
Rt1	orgs vs. people	orgs. {*}, people. {*}	orgs. {*}, people. {*}
Rt2	orgs vs. places	orgs. {*}, places. {*}	orgs. {*}, places. {*}
Rt3	people vs. places	people. {*}, places. {*}	people. {*}, places. {*}
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

2.2 实验结果与讨论

人造数据集的实验结果如表 3 所示,3 个文本数据集的实验结果如表 4～5 所示. 需要说明的是,本文对所有方法对参数均使用交叉验证法进行学习,并取最优结果进行记录.

由表 3～5 所示结果可以看出:

1) 基线方法 SVM 的训练集是目标域中少量的带标签数据,不足以训练一个高性能的分类器,故在所有数据集上的分类性能均低于其他学习方法;

2) 另外一个基线方法 TSVM 虽然其训练集包括源域中大量不带标签数据,但是由于源域和目标域数据的数据分布不同,故在大部分数据集上分类性能低于其他迁移学习类方法,这也进一步说明了使用传统的半监督模式分类算法不能很好地解决数据分布不同之下的数据分类问题;

3) 迁移分类算法 LMPROJ 在大部分数据集上的分类性能高于传统半监督模式分类方法而低于其他 2 个迁移分类算法,因为 LMPROJ 算法中目标域数据不带有类标信息,属于无监督迁移分类;

4) 本文所提算法 TS^{3}VM 在 3 个迁移分类算法中表现出鲁棒的分类性能,这进一步说明在迁移分类学习中采用扩维的思想较 TCA 算法采用降维

Table 3 Classification Accuracy Comparison on 2-moon with Different Rotation Angle

表 3 不同旋转角度的 2-moon 数据集分类精度比较

Algorithm	Dataset with Correct Labels			Dataset with False Labels		
	30°	45°	60°	30°	45°	60°
SVM	0.745 8±0.002 7	0.695 8±0.002 7	0.520 8±0.007 9	0.533 3±0.005 4	0.504 2±0.001 4	0.479 2±0.005 5
TSVM	0.758 3±0.003 2	0.737 5±0.007 1	0.679 2±0.004 2	0.529 2±0.004 5	0.638 3±0.001 7	0.641 7±0.007 3
LMPROJ	0.875 0±0.009 8	0.812 5±0.006 7	0.770 8±0.000 3	0.687 5±0.009 8	0.758 3±0.005 2	0.762 5±0.008 1
TCA	0.837 5±0.006 5	**0.859 2±0.003 0**	0.754 2±0.005 1	0.816 7±0.007 0	0.825 8±0.002 9	0.670 8±0.006 7
TS^{3}VM	**0.891 7±0.010 0**	0.845 8±0.002 3	**0.808 3±0.009 9**	**0.887 5±0.003 0**	**0.837 5±0.004 8**	**0.804 2±0.001 8**

Table 4 Classification Accuracy Comparison on 20NewsGroup

表 4 20NewsGroup 数据集分类精度比较

Algorithm	20NewsGroup					
	NG1	NG2	NG3	NG4	NG5	NG6
SVM	0.7513±0.0007	0.8160±0.0079	0.8161±0.0045	0.7975±0.0094	0.8013±0.0027	0.8613±0.0078
TSVM	0.8310±0.0017	0.8160±0.0075	0.9011±0.0021	0.7964±0.0081	0.8620±0.0074	0.8697±0.0050
LMPROJ	0.8201±0.0047	0.7674±0.0030	**0.9402±0.0031**	0.7203±0.0020	0.8504±0.0062	0.8702±0.0074
TCA	0.8778±0.0032	0.8957±0.0028	0.9035±0.0121	0.8769±0.0054	0.9159±0.0061	**0.9352±0.0018**
TS^3VM	**0.8875±0.0087**	**0.9178±0.0010**	0.8954±0.0022	**0.8892±0.0030**	**0.9173±0.0058**	0.9276±0.0013

Table 5 Classification Accuracy Comparison on Reuter-21578 and Email Spam

表 5 Reuter-21578 和 Email Spam 文本分类精度比较

Algorithm	Reuters			Email Spam		
	Rt1	Rt2	Rt3	Spam1	Spam2	Spam3
SVM	0.7980±0.0030	0.7880±0.0018	0.7640±0.0028	0.7688±0.0028	0.8388±0.0018	0.8113±0.0010
TSVM	0.7230±0.0021	0.6721±0.0017	0.6932±0.0034	0.7864±0.0010	**0.8920±0.0021**	0.7011±0.0017
LMPROJ	0.8463±0.0010	0.8020±0.0036	0.7080±0.0023	0.6764±0.0020	0.7890±0.0013	0.6980±0.0037
TCA	0.8784±0.0021	0.8973±0.0029	0.8529±0.0021	0.8352±0.0026	0.8546±0.0038	0.8735±0.0025
TS^3VM	**0.9260±0.0024**	**0.9100±0.0020**	**0.8833±0.0026**	**0.8572±0.0030**	0.8917±0.0017	**0.9153±0.0110**

的思想,对最终模式分类器来说效果更好,因为扩维的思想既考虑了源域和目标域领域间的共性知识又充分考虑了目标域数据的特有数据分布情况;

5) 尽管在部分数据集上(比如 Spam2 数据集)本文所提算法没有表现出最佳的均值精度,但是在方差中表现出了优势.

6) 更重要的是,从表 3 可以看出,在目标域数据类标受到攻击产生误标的情况下,本文所提算法表现出较强的鲁棒性,对最终分类器的影响甚微.这也进一步表现出本算法的相对于传统半监督分类算法和迁移分类算法的优点:在一定程度上摆脱了对已知类标数据类标的依赖性.

2.3 参数敏感性实验

在评价某个参数的性能影响时,先固定其他 3 个参数的最优值.采用数据集 NG1 和 Spam2 作为实验数据,图 3~6 分别显示了上述 4 个参数对所提方法的性能影响.由此可得如下结论:

1) 由图 3 可以看出,所提方法是基于特征扩维的共享特征空间分类学习模型,因此对扩展的共享特征空间维数 r 具有较大程度的敏感性.即对于不同的数据集 r 的取值明显影响所提方法的最终性能,这也进一步说明了针对不同的数据集扩展的共享特征空间维数 r 协调的重要性,同时这也是本文作者下一步要深入研究的方向之一.

2) 由图 4 可以看出,在本文所考虑 C 的所有取值范围内,分类精度变化幅度较大.这进一步说明了,

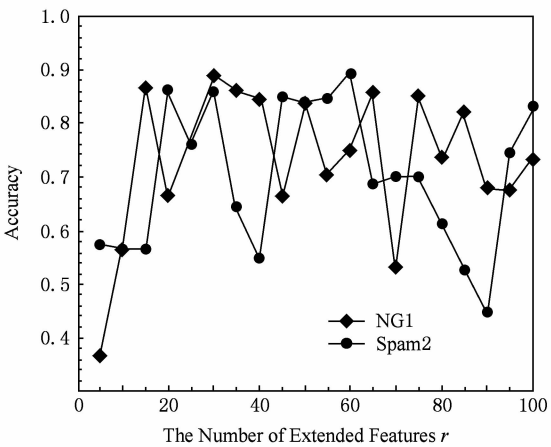


Fig. 3 Influence of parameter r on accuracy.

图 3 参数 r 对分类精度的影响

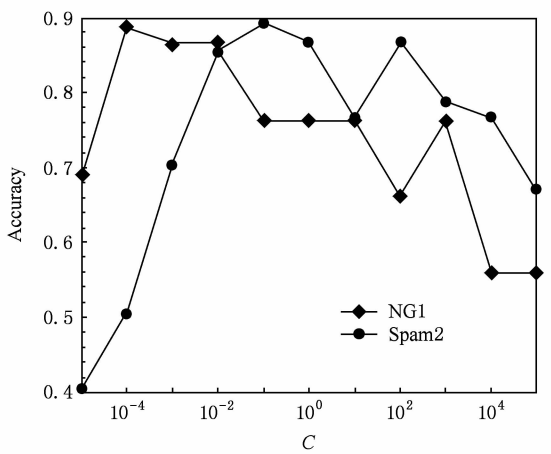


Fig. 4 Influence of parameter C on accuracy.

图 4 参数 C 对分类精度的影响

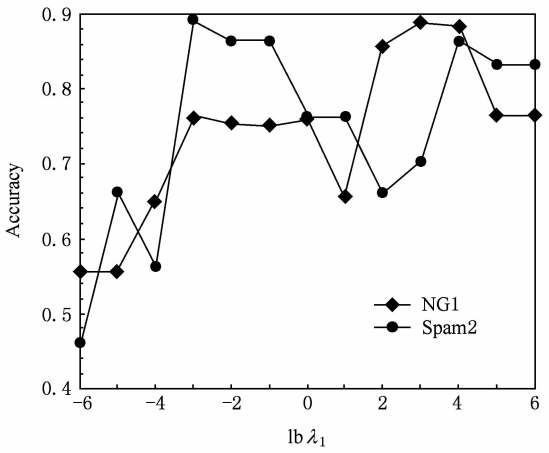


Fig. 5 Influence of parameter λ_1 on accuracy.

图 5 参数 λ_1 对分类精度的影响

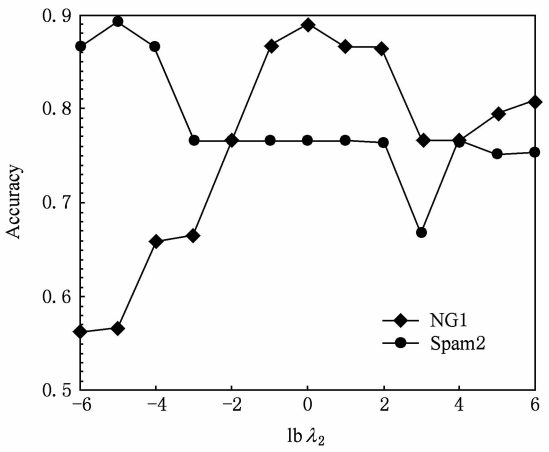


Fig. 6 Influence of parameter λ_2 on accuracy.

图 6 参数 λ_2 对分类精度的影响

基于结构风险最小化学习模型对参数 C 具有较大程度上的敏感性,即 C 在一定范围内的不同取值明显影响所提方法的泛化性能.

3) 由图 5 可以看出,由于本文所提方法基于最大间隔原则,故对平衡参数 λ_1 具有较大程度的敏感性,即 λ_1 在一定范围内的不同取值明显影响分类器最终性能.

4) 目标函数中第 4 项 $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}$ 表示在扩展的共享特征空间中目标域和源域数据的联合概率分布对分类器最终性能的影响,其联合概率分布越大,项 $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Theta}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)(\boldsymbol{x}_i^s-\boldsymbol{x}_j^t)^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{v}$ 的值越小,分类器性能越好;反之,性能越差. 由图 6 可以看出,λ_2 作为平衡分类器最终性能和训练数据在扩展的共享空间中联合概率分布的参数,对所提方法具有较大程度的影响.

3 结 论

综合利用相关领域的大量无标签数据来指导目标领域少量带标签数据来训练学习分类器是机器学习研究的热点之一. 本文在半监督学习中引入迁移思想,根据分类超平面最大间隔和联合概率分布最大原则来充分挖掘无标签和带标签数据间的"共同"知识,在"共同"知识基础之上又结合少量带标签数据的数据分布,提出了基于迁移共享特征空间的模式分类算法. 本文所提算法具有两大优点:1)从数据特征空间扩维的角度出发,既考虑带标签数据和无标签数据的"共同"知识,又考虑了少量带标签数据的数据分布;2)对目标域施加分类误差约束是在扩展后的特征空间而不是在原始特征空间,因此最终分类器在一定程度上摆脱了对带标签数据类标的依赖性. 在大量相关数据集上的实验也验证了本文所提方法的有效性.

目前来看,本文所提方法的一个不尽人意之处是扩展的特征空间维数的确定还没有一个理论性的结论,而是采用传统的实验验证及网格寻优的方法确定,因此如何理论性地给出共享特征空间维数的确定方法是一项具有挑战性的研究话题,这是我们今后研究的重点之一.

参 考 文 献

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

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

[3] Xie S, Fan W, Peng J, et al. Latent space domain Transfer between High Dimensional overleaping distributions [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 91-100

[4] 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

[5] Pan S J, Ni X S J, et al. Cross-domain sentiment classification via spectral feature alignment [C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 751-760

[6] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis [J]. IEEE Trans on Neural Network, 2011, 22(2): 199-210

[7] Hong Jiaming, Yin Jian, Huang Yun, et al. TrSVM: A transfer learning algorithm using domain similarity [J]. Journal of Computer Research and Development, 2011, 48(10): 1823-1830 (in Chinese)
(洪佳明, 印鉴, 黄云, 等. TrSVM: 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)

[8] Zhuang Fuzhen, Luo Ping, Shen Zhiyong, et al. Mining distinction and commonality across multiple domains using generative model for text classification [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(11): 2025-2039

[9] Shao M, Castillo C, Gu Z H, et al. Low-rank transfer subspace learning [C] //Proc of the 12th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2012: 1104-1109

[10] Gupta S K, Phung D, Adams B, et al. Regularized nonnegative shared subspace learning [J]. Data Mining and Knowledge Discovery, 2013, 26(1): 57-97

[11] Gu Xin, Wang Shitong. Fast cross-domain classification method for large multisources/small target domains [J]. Journal of Computer Research and Development, 2014, 51(3): 519-535 (in Chinese)
(顾鑫, 王士同. 大样本多源域与小目标域的跨领域快速分类学习[J]. 计算机研究与发展, 2014, 51(3): 519-535)

[12] Deng Z H, Choi K S, Chung F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information [J]. Pattern Recognition, 2010, 43(3): 767-781

[13] Yu J, Cheng Q S, Huang H K. Analysis of the weighting exponent in the FCM [J]. IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, 2004, 34(1): 164-176

[14] Yang S, Yan S, Zhang C, et al. Bilinear analysis for kernel selection and nonlinear feature extraction [J]. IEEE Trans on Neural Networks, 2007, 18(5): 1442-1452

[15] Jiang Yizhang, Deng Zhaohong, Wang Shitong. Mamdani-larsen type transfer learning fuzzy system [J]. Acta Automatica Sinica, 2012, 38(9): 1393-1409 (in Chinese)
(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)

[16] 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

[17] Bickel S. Ecml pkdd discovery challenge [EB/OL]. 2006 [2013-12-21]. http://www.ecmlpkdd2006.org/challenge.html

[18] Joachims T. Transductive inference for text classification using support vector machines [C] //Proc of ICML-99. San Francisco: Morgan Kaufmann, 1999: 200-209

[19] Chang C C, Lin C J. LIBSVM: A library for support vector machines [EB/OL]. 2001 [2013-11-15]. http://www.csie.ntu.edu.tw/~cjlin/libsvm

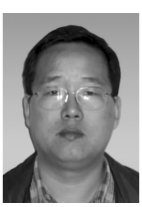
[20] Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data [J]. Journal of Machine Learning Research, 2005, 6: 1817-1853

[21] Zheng V W, Pan S J, Yang Q, et al. Transferring multi-device localization models using latent multi-task learning [C] //Proc of the 23rd Int Conf on Artificial intelligence. Menlo Park, CA: AAAI, 2008: 1427-1432

[22] Ji S, Tang L, Yu S, et al. A shared-subspace learning framework for multi-label classification [J]. IEEE Trans on Knowledge Discovery From Data, 2010, 4(2): 8:1-8:29

Dong Aimei, born in 1978. PhD candidate at the School of Digital Media, Jiangnan University, and lecturer at Qilu University of Technology. Her research interests include artificial intelligence, pattern recognition, and machine learning.

Bi Anqi, born in 1989. PhD candidate at the School of Digital Media, Jiangnan University. Her current research interests include pattern recognition, intelligent computation and their application.



Wang Shitong, born in 1964. Professor and PhD supervisor. Senior member of China Computer Federation. His current research interests include artificial intelligence, pattern recognition, fuzzy system, medical image processing, data mining, and intelligent computation and their applications.