计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
48(10): 1823-1830, 2011

TrSVM:一种基于领域相似性的迁移学习算法

洪佳明 印 鉴 黄 云 刘玉葆 王甲海
(中山大学信息科学与技术学院 广州 510006)
(hongjiaming8888@gmail.com)

TrSVM: A Transfer Learning Algorithm Using Domain Similarity

Hong Jiaming, Yin Jian, Huang Yun, Liu Yubao, and Wang Jiahai
(*School of Information Science and Technology, Sun Yat-sen University, Guangzhou* 510006)

Abstract Transfer learning algorithms focus on reusing related domain data to help solving learning tasks in the target domain. In this paper, we study the problem of inductive transfer learning. Most of the existing algorithms in inductive transfer learning might suffer from the problem of sample selection bias when the number of target domain data is too small. To address this problem, we propose to utilize domain similarity in a new approach. Through detailed discussion about the similarity of related domains, we define the concept of weak domain similarity. Using this concept to give additional constraints on the target classifiers, we develop a simple but effective approach to leverage the useful knowledge from the related domain, so that related domain data can be directly used in the training process. In this way, we are able to make the target classifier less sensitive to the small amount of target training data. Furthermore, we show that a modified SMO method can be applied to optimize the objective function in the algorithm effectively. The new algorithm is referred to as TrSVM, and can be seen as extension of support vector machines for transfer learning. Experiment results on extensive datasets show that TrSVM outperforms support vector machines and the state-of-the-art TrAdaBoost algorithm, and demonstrate the effectiveness of our algorithm.

Key words transfer learning; cross-domain learning; classification; SVM; domain similarity

摘 要 迁移学习是对传统监督学习的扩展,试图利用其他相关领域中的现存数据来帮助完成当前领域的学习任务. 对于归纳式迁移学习算法,当目标领域只有少量数据时,已有的算法容易受到选择性偏差的影响,不能充分发挥相关领域数据的作用. 为解决该问题,提出一种利用领域相似性的新途径:通过定义领域弱相似性的概念,将相似性的约束与目标分类器联系起来,能在训练过程中有效利用相关领域的大量数据,设计出一种基于支持向量机的迁移学习算法 TrSVM,并给出求解过程. 在大量数据集上的实验结果表明了新算法的有效性.

关键词 迁移学习;跨领域学习;分类;支持向量机;领域相似性

中图法分类号 TP391;TP181

近年来,随着信息化程度不断提高,数据挖掘和机器学习算法在实现信息自动化等方面发挥着越来越大的作用.同时,各种新的学习任务也对传统算法提出更高要求与挑战.迁移学习就是为了更好地利

收稿日期:2011-06-23;修回日期:2011-08-24
基金项目:国家自然科学基金项目(61033010,61070005);国家科技计划基金项目(2008ZX1005-013);广东科技计划基金项目(2009A080207005,2009B090300450,2010A040303004)

用各种数据而提出的新的研究方向[1].

通常,新的数据领域中只有少量的有标记数据,这给数据分析工作带来了困难.直接对该领域数据进行大量的人工标注代价昂贵,而且可能不存在大量的样本供用户进行标注.与此相反,人们在其他相关领域中已经积累了大量数据,这些数据中可能包含了可供借鉴的知识.例如,在对 blog 文档进行分类时,已经标注的新闻文档数据无疑可以对此提供帮助.尽管两者并不相同,但它们在很多方面的共同点(如相同的关键词)使得新闻文档的分类知识至少有一部分仍然适用于 blog 文档.如何将相关领域中的有用知识"迁移"到目标领域中,就是迁移学习所关注的问题.迁移学习不同于监督学习的重要特征是:训练数据与测试数据的分布不再是一致的,而这种一致性是传统机器学习的基本假设[2].

针对不同的应用,目前已提出归纳式迁移学习、直推式迁移学习和无监督迁移学习等算法[1,3].本文研究归纳式迁移学习算法(文献[1,3]的定义略有不同,本文研究文献[3]定义的情况),其特点是目标领域中仅有少量数据,需要利用其他领域数据来帮助训练.具体的问题如下:给定一个目标领域(target domain),只有少量的有标记样本以及另一个领域的大量有标记样本,这个领域与目标领域相关但不相同,称为源领域(source domain).学习目标是利用以上数据训练分类器 L,用 L 对目标领域中的未标记数据进行分类(测试数据在训练阶段未知,不能在训练过程中使用).

当前的归纳式迁移学习算法大都是基于实例进行知识迁移的方法,试图从源领域中选取部分相关的样本直接应用到目标领域中.但是,源领域的样本是否相关大都根据目标领域中的小数量样本来判断.由于小数量样本可能存在选择性偏差(sample selection bias),不能代表整个目标领域的分布情况,仅仅根据它们来选择源领域数据,可能会遗漏有用的数据,影响学习效果.

为了充分利用源领域的数据,本文提出新的算法设计思路:通过考察领域相关性,定义一种直观的领域相关性度量,称为领域弱相似性,并要求相关领域之间的领域弱相似性必须满足一定的约束.将这种约束与目标分类器联系起来,能自然地将源领域数据嵌入到支持向量机(SVM)[4]的训练过程中,得到新算法 TrSVM.

本文的主要贡献是:首先,讨论了领域相关性及其含义,给出领域弱相似性的定义;其次,利用以上概念扩展 SVM 算法,得到新算法 TrSVM,并给出其优化算法.

1 相关工作

当前大部分迁移学习算法都假设两个领域具有相同的特征空间与类标号集合[1].这些算法大都属于以下两大类:基于实例的迁移学习和基于特征表示的迁移学习.

基于实例的迁移学习方法从源领域中选取部分相关的样本,直接应用到目标领域的学习任务中.代表性的工作是 TrAdaboost 算法[5],通过改进 AdaBoost,根据目标领域数据逐步调整源领域样本的权重,利用不同权重的样本进行训练,达到有效迁移知识的目的.这种方法得到的分类器可能有较大的泛化误差(见文献[5]定理 4).

基于特征表示的迁移学习方法试图寻找新的特征表示,使得在新特征表示下,不同领域的分布差异减小,进而将知识顺利地"迁移"到目标领域.其中,Pan 等人[6]通过降维的方法使分布的距离减小.Xie 等人[7]提出一种结合回归与 SVD 降维的方法,能拉近两个领域分布的距离.这种方法要求目标领域存在大量的未标记数据,因此不适用于归纳式迁移学习问题.

2 问题定义

给定目标领域数据集 T 和源领域数据集 S,其样本的特征空间相同,x 为特征向量,y 为类标号,记 $T=\{(\boldsymbol{x}_i,y_i)\mid i=1,2,\cdots,m\}$,$S=\{(\boldsymbol{x}_j,y_j)\mid j=m+1,m+2,\cdots,m+n\}$,$m$ 和 n 为样本数量,$m\ll n$.本文考虑 2-分类问题,y_i 的值为 $+1$ 或 -1.假设源领域样本变量服从分布 P_S,简记为 $(X_S,Y_S)\sim P_S(X,Y)$,目标领域样本变量服从分布 P_T,记为 $(X_T,Y_T)\sim P_T(X,Y)$.对应的边缘分布与条件分布记为 $P_S(X)$,$P_S(Y\mid X)$,$P_T(X)$,$P_T(Y\mid X)$.在迁移学习中,P_S 与 P_T 相关但不相同.

本文研究的问题是给定样本集 S 和 T,其中 S 为由 $P_S(X,Y)$ 产生的独立同分布样本集,T 为由 $P_T(X,Y)$ 产生的独立同分布样本集,利用 S 与 T 训练适用于分布 $P_T(X,Y)$ 的分类器 L.

3 领域弱相似性

常见的估计分布距离的方法,如 KL 距离等[8],

要求在目标领域中存在大量样本,这在归纳式迁移学习中是无法满足的.另外,如果能将分类器用于衡量分布的相似性,可能给学习过程带来帮助.有鉴于此,本文提出领域弱相似性的概念.

3.1 领域弱相似性定义

首先,我们从直观角度讨论领域相似性的特点.对于任何目标领域样本 x_t,设其类标号为 y_t,则 $P_T(y_t|x_t)>0.5$,同时,假设某源领域样本 x_s 与 x_t 相同,那么,由于两个领域有较大相关性,x_s 的类标号很可能也是 y_t,即以很大的概率选择 x_s,$P_S(y_t|x_t)>0.5$.显然,与 $P_S(y_t|x_t)=P_T(y_t|x_t)$ 的要求相比,这是较弱的条件.如果以上假设不成立,那么,相同的样本在两个领域中的类标号相反的可能性非常大,不符合我们对相关性的直观理解.以下给出这种相关性的定义.

定义 1. 对分布 $P_S(X,Y)$ 和 $P_T(X,Y)$,X 为样本变量,Y 为类标号,取值为 -1 或 $+1$.如果能够以 $1-v$ 的概率推断,对 $x\sim P_S(X)$,有 $(0.5-P_T(1|x))(0.5-P_S(1|x))>0$ 成立,则称分布 P_S 以 $1-v$ 的概率弱相似于 P_T,称 v 为负相似度,记为 $v(P_S,P_T)$.

假设分类器 L 在 P_T 上的泛化误差小于一定上界,用 L 代替 P_T,得到定义 2:

定义 2. 对分布 $P_S(X,Y)$ 和 $P_T(X,Y)$,X 为样本变量,Y 为类标号,取值为 -1 或 $+1$,e 为小于 1 的正数.若存在分类器 $L:X\to\{-1,+1\}$,使 L 在 P_T 上的泛化误差小于 e,且能以 $1-v$ 的概率推断,对 $(x,y)\sim P_S(X,Y)$,有 $L(x)=y$ 成立,则称 P_S 以 $1-v$ 的概率 e-弱相似于 P_T,称 v 为负相似度,记为 $v(L,P_S,P_T)$.

当 e 值较小时,L 近似于目标领域的类标号生成函数(labeling function[9],是在目标领域泛化误差为零的分类器),可将 $v(L,P_S,P_T)$ 作为 $v(P_S,P_T)$ 的近似.定义 2 用分类器 L 代替了未知的 P_T,有利于应用到训练过程中.根据以上讨论,由于迁移学习中假设两个领域是相关的,负相似度 $v(L,P_S,P_T)$ 必须比较小.

3.2 SVM 中的负相似度

本节讨论如何将负相似度应用到 SVM 中.设 SVM 的判别函数为 $f(x)$,为简明起见,下面讨论中不引入核函数,设 $f(x)=w^{\mathrm{T}}x+b$.

在 SVM 中,对目标领域中的 $f(x)$,其分类决策函数为 $L(x)=sign(f(x))$.设 $L(x)$ 在目标领域的泛化误差 e 足够小,得到 SVM 中的负相似度 $v(sign(f),P_S,P_T)$(简记为 $v(f,P_S,P_T)$)如下:

$$v(f,P_S,P_T)=E_{(x,y)\sim P_S}(I(yf(x)<0)). \quad (1)$$

其中 $I(t)$ 为指示函数,条件 t 成立时为 1,否则为 0.利用 S 中的大量数据可得式(1)的经验估计如下:

$$v'(f,P_S,P_T)=\frac{1}{n}\sum_{l=m+1}^{m+n}I(y_if(x_i)<0). \quad (2)$$

4 TrSVM 算法

4.1 TrSVM 算法推导

当目标领域的训练样本数目 m 足够大时,可直接应用 SVM 算法求解[4]:

$$\begin{aligned}&\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i,\\&\text{s.t. } y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i,\\&\quad \varepsilon_i\geqslant 0,i=1,2,\cdots,m.\end{aligned} \quad (3)$$

当 m 较小时,SVM 得到的 $f(x)$ 有较大的泛化误差.为避免这一问题,我们利用弱相似度定义,给 $f(x)$ 赋予额外约束,要求 $f(x)$ 在源领域上的负相似度估计值 $v'(f,P_S,P_T)$ 不能过大,得到如下优化问题:

$$\begin{aligned}&\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i+nC_Sv'(f,P_S,P_T),\\&\text{s.t. } y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i,\\&\quad \varepsilon_i\geqslant 0,i=1,2,\cdots,m,\end{aligned} \quad (4)$$

其中 C_S 为平衡参数.

以上方法称为 TrSVM,其优化目标综合考虑 3 方面因素:1)控制模型复杂度;2)控制分类器在目标领域的经验风险;3)控制分类器在源领域的负相似度.

由于式(4)中 $v'(f,P_S,P_T)$ 不是凸函数,较难优化[10],利用代理损失函数方法,用易于优化的损失函数来代替式(2).注意到,对源领域样本 x,当 $yf(x)<0$ 时,x 被错分,$yf(x)$ 绝对值越大则其距离决策边界越远,经验风险越大,为此用以下 $V(f,P_S,P_T)$ 来替代 $v'(f,P_S,P_T)$:

$$V(f,P_S,P_T)=\frac{1}{n}\sum_{i=m+1}^{m+n}\max(0,-y_if(x_i)). \quad (5)$$

由此,式(4)转化为:

$$\begin{aligned}&\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i+\\&\quad C_S\sum_{j=m+1}^{m+n}\max(0,-y_i(w^{\mathrm{T}}x_i+b)),\\&\text{s.t. } y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i,\end{aligned}$$

$$\varepsilon_i \geqslant 0, i = 1,2,\cdots,m. \tag{6}$$

TrSVM 与 SVM 的不同在于：优化目标中增加了 $V(f,P_S,P_T)$，能够利用源领域数据来帮助训练. 对目标领域，TrSVM 要求样本点落在最大间隔超平面之外，对不满足条件(即 $y_i f(\boldsymbol{x}_i)<1$)的点进行惩罚；对源领域的约束较弱，仅要求其样本点落在决策边界的正确一侧(即 $y_i f(\boldsymbol{x}_i)>0$). 通过这种方式，TrSVM 使尽量多的源领域样本被 f 正确分类，保证领域弱相似性较大，而弱化其在分布上要与目标领域一致的过高限制(不要求其处于最大间隔平面之外)，体现源领域在训练过程的辅助作用.

4.2 TrSVM 的优化问题

引入松弛变量，将式(6)转化为：

$$\begin{aligned}&\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\varepsilon_i + C_S\sum_{j=m+1}^{m+n}\varepsilon_j,\\ &\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1-\varepsilon_i, i = 1,2,\cdots,m,\\ &\quad y_j(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j + b) \geqslant 0-\varepsilon_i,\\ &\quad j = m+1,m+2,\cdots,m+n,\\ &\quad \varepsilon_k \geqslant 0, k = 1,2,\cdots,m+n.\end{aligned} \tag{7}$$

以上问题的拉格朗日函数是：

$$\begin{aligned}L(\boldsymbol{w},b,\varepsilon,\alpha,r) = &\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\varepsilon_i + C_S\sum_{j=m+1}^{m+n}\varepsilon_j -\\ &\sum_{i=1}^{m+n} r_i\varepsilon_i - \sum_{i=1}^{m}\alpha_i(y_i f(\boldsymbol{x}_i) - 1 + \varepsilon_i) -\\ &\sum_{i=m+1}^{m+n}\alpha_i(y_i f(\boldsymbol{x}_i) + \varepsilon_i).\end{aligned} \tag{8}$$

将式(8)对 $\boldsymbol{w},b$ 求偏导并置零，得到：

$$\begin{aligned}\frac{\partial L}{\partial w} &= \boldsymbol{w} - \sum_{i=1}^{m+n}\alpha_i y_i \boldsymbol{x}_i = 0;\\ \frac{\partial L}{\partial b} &= \sum_{i=1}^{m+n}\alpha_i y_i = 0.\end{aligned} \tag{9}$$

将式(9)代入式(7)，由对偶定理[10]，得到：

$$\begin{aligned}&\max \sum_{i=m+1}^{m+n}\delta_i\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m+n}\alpha_i\alpha_j y_i y_j\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle,\\ &\text{s.t. } 0 \leqslant \alpha_i \leqslant C_i, i = 1,2,\cdots,m+n,\\ &\quad C_i = C,\delta_i = 1, \forall\, i = 1,2,\cdots,m,\\ &C_i = C_S,\delta_i = 0, \forall\, i = m+1,m+2,\cdots,m+n,\\ &\quad \sum_{i=1}^{m+n}\alpha_i y_i = 0.\end{aligned} \tag{10}$$

以上引入记号 C_i 及 δ_i，方便下文论述. 由 KKT 条件，最优解需满足如下条件：

$$\begin{aligned}&\alpha_i = 0 \Rightarrow y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant \delta_i,\\ &0 < \alpha_i < C_i \Rightarrow y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) = \delta_i,\\ &\alpha_i = C_i \Rightarrow y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \leqslant \delta_i.\end{aligned} \tag{11}$$

求得最优解后，利用式(9)，得到 $\boldsymbol{w}$ 如下：

$$\boldsymbol{w} = \sum_{i=1}^{m+n}\alpha_i y_i \boldsymbol{x}_i. \tag{12}$$

而 b 可由式(11)直接得到，从而求得 $f(x)$.

4.3 问题求解

式(10)是与 SVM 相似的优化问题. 通过适当修改，可用与 SMO[11] 相同的流程来求解. 以下略述 SMO 的求解过程，然后指出如何对其进行修改，用于求解 TrSVM 的问题.

SMO 利用 KKT 约束条件，通过迭代方法逐步求解. 主要步骤是 1)使用启发式规则选择两个变量 α_i 和 α_j；2)固定其余变量不变，求解一个基于盒约束的二次优化问题，将 α_i 和 α_j 更新为 α_i^{new} 和 α_j^{new}；3)重复以上步骤直至收敛.

SMO 的第 1 个启发式规则是：优先从处于(0, C)中的变量中选取不满足 KKT 条件的变量 α_i，目的是加速算法收敛.

选定变量 α_i 后，SMO 使用第 2 个启发式方法选择变量 α_j. 定义当前 $f(x)$的输出误差如下：

$$E_l = f(x_l) - y_l, l = 1,2,\cdots,m. \tag{13}$$

SMO 选择使 $|E_i - E_j|$ 较大的 α_j 以加速收敛：若 E_i 为正，则选择 E_j 最小的 α_j，否则选择 E_j 最大的 α_j.

选定 α_i 和 α_j 以后，步骤 2)对其值进行更新，进而对判别函数进行更新，如此迭代直至收敛. 详细过程请参阅文献[11]. 相同的求解流程可用于求解式(12)，仅在以下 3 处作了一些调整：

首先，在选择第 1 个变量时，优先选取不满足 KKT 条件的非边界变量，同时优先选择目标领域的变量，即优先从 $i>m$ 的 α_i 中选取；

其次，由于两个领域的损失函数不同，输出误差函数修改如下：

$$E_i = f(\boldsymbol{x}_i) - \delta_i y_i, \tag{14}$$

由 5.1 节的讨论，以上修改适应了两个领域的不同情况，使我们仍然能够根据 E 值来选择要优化的变量；

最后，由式(10)，变量的区间发生变化，其更新公式也应随之改变. 设 α_i 与 α_j 更新为 α_i^{new} 与 α_j^{new}，可求得 α_j^{new} 的取得范围$[L,H]$为：

当 $y_i y_j = -1$ 时，

$$\begin{aligned}L &= \max(0,\alpha_i - \alpha_j),\\ H &= \min(C_j, C_i + \alpha_j - \alpha_i);\end{aligned} \tag{15}$$

当 $y_i y_j = 1$ 时，

$$L = \max(0,\alpha_j + \alpha_i - C_i),$$

$$H=\min(C_j,\alpha_i+\alpha_j); \tag{16}$$

当不考虑区间约束时，式(10)的最大值点为

$$\alpha_j^{\max}=\alpha_j-\frac{y_j(E_i-E_j)}{(2X_{ij}-X_{ii}-X_{jj})}, \tag{17}$$

这与 SMO 中用到的计算公式形式相同.

当 $\alpha_j^{\max}$ 在以上区间时，α_j^{new} 直接更新为该值，否则极值在区间端点取得. 这在形式上与 SMO 的更新公式相同，此处不再赘述.

显然，由于采用了 SMO 方法，TrSVM 也继承了 SMO 的高效性.

5 实验结果与分析

5.1 数据集

我们选用以下文本数据集进行实验：SRAA 数据集、20 newsgroups 数据集以及 Reuters-21758 数据集. 这些数据集都采用层次化的组织方式，通过将顶层的类标号作为分类标号，将顶层类别的数据依据不同的子类别分割为两个子集，就能产生两个相关但又不相同的数据集，分别作为目标领域和源领域. 例如，在 20 newsgroups 数据集中，包含 comp 与 rec 两个顶层分类，而 comp 之下包含 comp. windows. x，comp. graphics 等子类别，将 comp. windows. x 中的文档视为源领域中类标号为 comp 的数据，而将 comp. graphics 中的文档视为目标领域中类标号为 comp 的数据，就能构造两个不同的数据领域. 我们将文档数据转化为单词向量的形式，具体做法参照文献[7]，数据集详情参见表1. 前两个数据集来自 SRAA，中间的 6 个数据集来自 20 newsgroups，最后 3 个数据集来自 Reuters-21578，表中 M，N 为源领域与目标领域样本数.

Table 1 Data Summary

表 1 实验数据

Dataset	Number of Features	Number of Instances (M:N)
auto. vs. aviation	5 991	2 005:1 980
real. vs. simulated	6 161	2 020:2 008
comp. vs. rec	5 681	2 431:1 951
comp. vs. sci	6 773	2 007:2 373
comp. vs. talk	6 418	2 218:1 837
rec. vs. sci	6 935	1 963:1 992
rec. vs. talk	6 203	1 885:1 761
sci. vs. talk	6 390	1 663:1 939
orgs. vs. places	4 415	1 016:1 046
orgs. vs. people	4 771	1 239:1 210
people. vs. places	4 562	1 079:1 080

5.2 对比算法

作为对比，我们同时测试了当前领先的迁移学习算法与 SVM 算法. 为使比较更全面，使用不同的训练集学习 SVM：使用源领域数据作为训练集(记为 SVM-S)，使用目标领域数据作为训练集(记为 SVM-T)，以及同时使用两个领域数据作为训练集(记为 SVM-ST). 在迁移学习算法方面，我们选择 TrAdaBoost 算法[4]进行比较，选择 SVM 作为其基分类器.

在算法实现方面，采用 Weka 与 LibSVM，算法参数依照程序包的默认设置，TrAdaBoost 依照文献[5]设置参数. 算法中的核函数使用线性核函数.

为衡量各种算法的整体表现，采用与文献[5]类似的方法：在每个数据集上，设源领域数据总量为 M，从目标领域中随机选取大约 $0.01\times M$ 的样本，加上所有的源领域样本作为训练集，目标领域中其余的样本则作为测试集. 以上过程重复 10 次，计算其平均错误率.

5.3 参数敏感性测试

首先测试参数选择对 TrSVM 的影响. 在 sci. vs. talk 数据集上，分别随机抽取数据进行实验. 首先，固定 C 为 1，比较参数 C_S 分别取值为 0.5，1.0，1.5，2.0 对分类性能的影响. 实验结果如图 1 所示. 其次，固定 C_S 为 1，参数 C 分别取值为 0.5，1.0，1.5，2.0，实验结果如图 2 所示.

可以看出，随着参数的变化，算法的分类性能并没有受到很大影响，其分类错误率比较接近，说明 TrSVM 不会因参数的微小变动影响分类性能，在其他数据集上也有相似的现象. 在实验中，选取 C，C_S 为 0.5～2 之间的值即可.

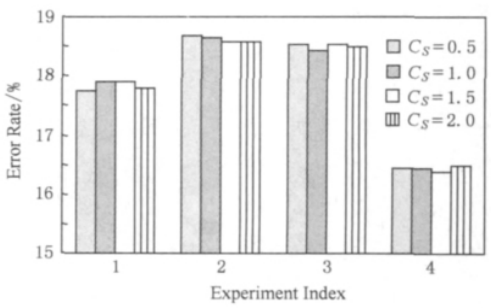

Fig. 1 Parameter sensitivity test on sci. vs. talk dataset with respect to C_S.

图 1 在 sci. vs. talk 数据集上对 C_S 的参数敏感性测试

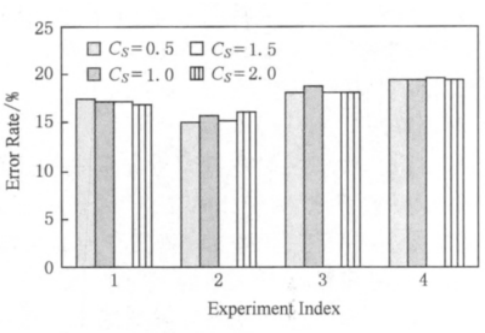

Fig. 2 Parameter sensitivity test on sci. vs. talk dataset with respect to C.

图 2 在 sci. vs. talk 数据集上对 C 的参数敏感性测试

5.4 分类性能比较

在本节中,我们在以上 11 个基准数据集上进行实验,比较 TrSVM 算法与其他算法的分类效果. 对 TrSVM 的参数,固定 C 为 1,对 C_S 分别取值为 1 和 2,对应的分类器称为 TrSVM-1,TrSVM-2.

表 2 列出各种算法在 11 个数据集上的平均错误率. 可以看出,TrSVM 的分类效果比较突出:在其中 10 个数据集上,TrSVM 的错误率是最低的,而在 SVM-ST 占优的 comp. vs. talk 数据集上,TrSVM 的错误率也与最优值非常接近(仅有约 0.44%的差距).

Table 2 Comparision of Perfomance

表 2 各算法实验结果比较

Dataset	SVM-S	SVM-T	SVM-ST	TrAdaBoost	TrSVM-1	TrSVM-2
comp. vs. rec	16.51	32.19	14.54	14.42	**13.73**	13.75
comp. vs. sci	33.71	34.13	29.52	26.58	**22.50**	**22.50**
comp. vs. talk	8.33	26.39	**7.57**	8.24	8.01	8.02
rec. vs. sci	28.65	35.69	23.83	21.02	17.43	**17.42**
rec. vs. talk	31.10	30.83	22.09	20.24	11.49	**11.46**
sci. vs. talk	29.85	38.40	24.84	22.74	18.23	**18.22**
real. vs. simulated	38.11	36.39	32.03	26.54	21.49	**21.45**
auto. vs. auitation	27.24	33.86	24.46	23.22	**20.61**	**20.61**
orgs. vs. places	33.62	36.73	32.54	30.49	29.93	**29.43**
orgs. vs. people	30.58	42.07	29.23	29.39	**27.79**	27.97
people. vs. places	47.46	42.36	46.22	44.21	**39.33**	41.06

观察表 2 中 SVM-T 的分类错误率,可以发现训练数据不足对分类器性能的影响:除了 comp. vs. talk 数据集,SVM-T 的错误率都在 30%以上,说明仅仅依靠本领域的少量数据进行训练并不足够. 比较表 2 前 3 列的分类结果,SVM-S 与 SVM-ST 的分类效果相对 SVM-T 有一定提高,说明源领域数据确实能为目标领域的分类任务提供帮助,而分类效果的改进程度说明两个领域相关性的大小:在 comp. vs. talk 和 comp. vs. rec 数据集上,SVM-S 与 SVM-ST 分类效果有成倍的提高,而在 people. vs. places 数据集上,错误率反而从 42.36%上升到 47%左右,表明两个领域的相关程度不高.

由于 SVM-S 与 SVM-ST 直接利用监督学习算法进行训练,并不考虑领域数据的分布差异,故而其分类效果受到影响. 将其与最后 3 列几种迁移学习算法相比较,可以观察到,在训练过程中区分了领域差异的迁移学习算法,其分类效果有一定提高. 例如,在 real. vs. simulated 数据上,错误率从 SVM-ST 的 32.03%下降到 TrAdaBoost 的 26.54%,而 TrSVM 进一步降低到 20.61%,在其他数据上也大都有不同程度的改善. 比较两种迁移学习算法,TrSVM 在很多数据集上效果都有明显提高,在部分数据集上效果略优于 TrAdaBoost. 图 3 依次标出 SVM-ST、TrAdaBoost 和 TrSVM 3 种算法在各个数据集上的错误率,相对于 TrAdaBoost 对 SVM-ST 的改善程度,TrSVM 的改善效果更加明显.

以上实验数据及分析表明 TrSVM 算法确实是有效的.

下一组实验研究目标领域训练样本增加时,TrSVM 与监督学习分类效果的变化. 与以上实验相同,固定源领域样本个数 M 不变,令目标领域训练样本数量 $t\times M$ 逐渐变化,t 从 0.01 增加到 0.5,

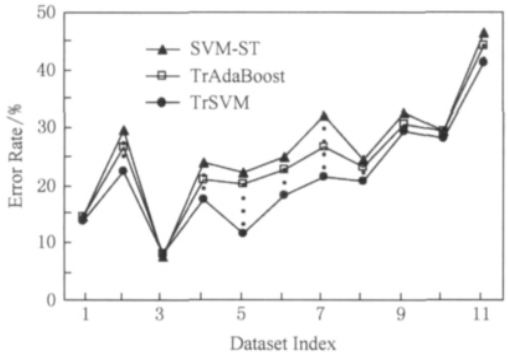

Fig. 3 Performance comparison on 11 datasets.

图3 在11个数据集上的分类效果比较

比较SVM-T，TrAdaBoost与TrSVM在rec. vs. talk上的分类效果.错误率曲线如图4所示.在rec. vs. talk数据集上，当比率$t<0.3$时，迁移学习的效果较明显.当比率逐渐增大，目标领域的训练样本已足够多时，TrSVM的分类效果与SVM-T没有太明显的差距，当比率足够大时(在此数据集上约为0.5)，SVM-T的效果超过TrSVM.相似的情况也可在其他数据集上观察到(但具体的比率不一定相同).

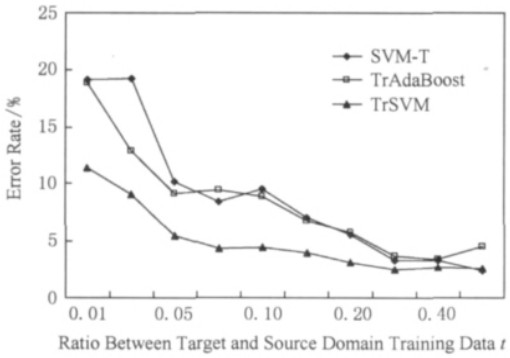

Fig. 4 Error rate curves on rec. vs. talk dataset.

图4 在rec. vs. talk数据集上的错误率曲线

6 结 论

迁移学习研究如何利用其他相关领域的已有知识，来帮助完成目标数据领域的学习任务.本文研究归纳式迁移学习问题，通过定义领域弱相似性的概念，给出领域相关性直观的衡量方法，能够充分利用相关领域的大量数据来帮助训练，并转化为易于求解的优化问题.大量数据集上的实验表明了新算法的有效性.下一步将研究如何利用弱相似性扩展其他算法，如贝叶斯算法等.其次，当存在多个相关领域时，能否利用弱相似性的概念设计算法也是以后要探讨的问题.

参 考 文 献

[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] Vapnik V. An overview of statistical learning theory [J]. IEEE Trans on Neural Networks, 1999, 10(5): 988-999

[3] Shi Y, Lan Z, Liu W, et al. Extending semi-supervised learning methods for inductive transfer learning [C] //Proc of the 9th IEEE Int Conf on Data Mining. Los Alamitos: IEEE Computer Society, 2009: 483-492

[4] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167

[5] 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, 2007: 193-200

[6] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2008: 677-682

[7] 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, 2009: 91-100

[8] Xu Zhen, Sha Chaofeng, Wang Xiaoling, et al. A semi-supervised learning algorithm from imbalanced data based on KL divergence [J]. Journal of Computer Research and Development, 2010, 47(1): 81-87
(许震，沙朝锋，王晓玲，等. 基于KL距离的非平衡数据半监督学习算法[J]. 计算机研究与发展，2010，47(1)：81-87)

[9] Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains [J]. Machine Learning, 2010, 79(1/2): 151-175

[10] Boyd S, Vandenberghe L. Convex Optimization [M]. Cambridge: Cambridge University Press, 2004

[11] Platt J C. Sequential minimal optimization: A fast algorithm for training support vector machines, MST-TR-98-14 [R]. Redmond, WA: Microsoft Research, 1998

Hong Jiaming, born in 1984. PhD candidate in the School of Information Science and Technology, Sun Yat-sen University. His major research interests include data mining and machine learning.

Yin Jian, born in 1968. Professor and PhD supervisor in Sun Yat-sen University. His major research interests include data mining and machine learning (issjyin@mail.sysu.edu.cn).

Liu Yubao, born in 1975. PhD, associate professor in Sun Yat-sen University. His major research interests include database system and data mining (liuyubao@mail.sysu.edu.cn).

Huang Yun, born in 1976. PhD candidate in Sun Yat-sen University. His major research interests include data mining (huangyun@ustc.edu).

Wang Jiahai, born in 1977. PhD, associate professor in Sun Yat-sen University. His major research interests include artificial intelligence and data mining (wangjiah@mail.sysu.edu.cn).

2012 年《计算机研究与发展》专题征文通知
——新型存储系统及其关键技术

近年来,随着国家和社会信息化发展的不断加速,信息存储的需求越来越广泛,数据存储量越来越大,目前存储系统的性能、功耗、容量、可靠性、安全性等各种问题严重阻碍了我国的信息化进程.因此,对云存储技术、SSD 技术、新兴存储芯片技术等新型存储系统及其关键技术的研究成了目前存储领域的研究热点.新型存储技术的发展将对现有存储系统结构产生巨大冲击和深远影响.近年来,国内外相关学者已在这些方面开展了大量的相关研究,也取得了一些重要研究成果.为此,《计算机研究与发展》将于 2012 年出版一个"新型存储系统及其关键技术"专题,报道该领域国内外科技工作者所取得的研究成果,提炼新型存储系统与应用中有待解决的关键问题,同时展望存储系统未来的研究方向,以推动我国存储系统与关键技术的科学研究和工程应用.

本期"新型存储系统及其关键技术"专题将面向国内外征集论文,欢迎广大学者、专家、工程技术人员积极投稿,现将专题论文征集的有关事项通知如下.

征文范围(但不限于):

①云存储技术与应用;②SSD (Solid State Disk)技术与应用;③新型的存储体系结构;④新兴存储技术及应用;⑤分布式及并行 I/O 技术;⑥存储安全问题;⑦存储系统可靠性、可用性及容灾问题;⑧归档存储系统;⑨存储虚拟化技术;⑩缓存技术及其一致性问题;⑪数据库存储技术;⑫移动存储技术;⑬存储能耗问题;⑭重复数据删除技术;⑮分级存储技术;⑯存储文件系统设计.

投稿要求

① 来稿应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表或宣读过.

② 论文一律用 word 格式排版,论文格式体例参考近期出版的《计算机研究与发展》的要求(http://crad.ict.ac.cn/).

③ 论文通过专辑投稿信箱(storage2012@126.com)发送电子稿,投稿时提供作者的联系方式(Excel 文档,按顺序包含:作者、论文标题、联系人、email、电话(手机)、通信地址、邮编).

重要时间

截稿日期:2012 年 4 月 10 日

结果通知日期:2012 年 8 月 15 日

特约编辑

舒继武 教授 清华大学计算机科学与技术系 shujw@tsinghua.edu.cn

方 粮 研究员 国防科学技术大学计算机学院 Lfang@nudt.edu.cn

专辑投稿信箱:storage2012@126.com

通信地址:北京市海淀区清华大学计算机科学与技术系 邮编:100084

联 系 人:舒继武

电 话:010-62783505-5