中国科学: 信息科学 2021 年 第 51 卷 第 10 期: 1609–1624
SCIENTIA SINICA Informationis
论文

基于联合分布核适配的迁移学习及其隐私保护

倪宣明[1], 沈鑫圆[1], 张海[2,3*]

1. 北京大学软件与微电子学院, 北京 100871
2. 西北大学数学学院, 西安 710127
3. 华东师范大学统计与数据科学前沿理论及应用教育部重点实验室, 上海 200062
* 通信作者. E-mail: zhanghai@nwu.edu.cn

收稿日期: 2020–02–08; 修回日期: 2020–04–25; 接受日期: 2020–06–15; 网络出版日期: 2021–10-12

国家自然科学基金委员会 – 广东省人民政府大数据科学研究中心项目 (批准号: U1811461) 资助

摘要 迁移学习利用不同但相关的源域标记数据来解决目标领域的学习问题, 大多数减小域间分布差异的方法依赖于最大均值差异距离, 但其仅仅能匹配域间数据分布的各阶矩. 此外, 隐私保护意识的增强限制了对数据源的访问, 对迁移学习的发展提出了新的挑战. 本文提出一种基于联合分布核适配的迁移学习及其隐私保护方法, 直接在再生核希尔伯特空间中同时减小域间边缘分布和条件分布的差异, 从而学习一个域不变核矩阵. 此外, 我们设置数据源双方首先访问一个相同的随机投影函数, 然后聚合器发布基于目标扰动的差分隐私核分类器, 在实现基于核的联合分布适配的同时, 避免了数据源与聚合器直接共享原始特征数据. 在多个文本和图像迁移学习基准数据集上进行了对比实验和参数分析, 结果显示本文方法具有良好的有效性.

关键词 迁移学习, 隐私保护, 分布适配, 谱学习, 差分隐私

1 引言

分类 (classification) 方法是机器学习领域使用最广泛的技术之一, 它根据带有标签的数据样本 (训练样本) 训练分类模型, 然后运用分类模型对新数据样本 (测试样本) 的类型进行预测, 目前已广泛应用于文本情感分析、人脸识别、语音识别等领域. 为了保证训练得到的分类模型具有准确性和高可靠性, 传统分类学习需要满足两个基本假设: (1) 用于学习的训练样本与新的测试样本满足独立同分布; (2) 需要有足够的训练样本才能学习得到一个好的分类模型 [1]. 但是, 训练样本与测试样本概率分布失配 (包括边缘分布失配和条件分布失配) 的问题在实际中广泛存在, 如在自然语言处理 [2]、计算机视觉 [3] 领域, 使得已有的训练样本不足以通过训练得到一个可靠的分类模型; 另外, 有标签样本往往很缺乏, 而且很难获得, 导致实际应用中缺乏足够多的有标签数据来训练模型.

引用格式: 倪宣明, 沈鑫圆, 张海. 基于联合分布核适配的迁移学习及其隐私保护. 中国科学: 信息科学, 2021, 51: 1609–1624, doi: 10.1360/SSI-2020-0020
Ni X M, Shen X Y, Zhang H. Transfer learning based on joint distribution kernel adaptation and its privacy protection (in Chinese). Sci Sin Inform, 2021, 51: 1609–1624, doi: 10.1360/SSI-2020-0020

上述问题可以通过引入迁移学习 (transfer learning)[4] 机制来解决, 它放宽了传统分类学习中的两个基本假设, 学习的目标是修正领域间的概率分布失配, 使得标准分类器可以在领域间有效迁移. 迁移学习任务包括两种不同类型的数据集, 分别来自源域和目标域. 源域包含大量有标签数据足以训练准确的分类器; 目标域包含大量无标数据, 服从与源域显著不同但又潜在相关的概率分布.

由于源域和目标域的分布不同, 迁移学习面临的首要问题是如何减小分布间的差异, 为此大量研究成果集中于通过参数化或非参数化距离的最小化实现领域间的概率分布适配. 这些方法的主要思想是学习隐含特征表示或实例权重, 使得源域和目标域间出现共享特性, 并显式地修正分布失配 [5]. 常用的概率分布距离度量函数包括 Kullback-Leibler (KL) 散度 [6]、Bregman 散度 [7] 和最大均值差异 (maximum mean discrepancy, MMD) [8,9]. 例如迁移子空间学习 (transfer subspace learning, TSL) [7] 采用 Bregman 散度作为比较边缘分布的距离, 但是使用参数化距离通常需要先进行非平凡的分布密度估计, 极大地增加了机器学习模型设计的复杂性. 因此, 相当一部分研究聚焦于如何最小化 MMD 这一非参数化统计量, 它通过核空间中源域和目标域间的概率分布均值距离来度量概率分布差异. 如, 迁移成分分析 (transfer component analysis, TCA) [9] 将 MMD 作为比较边缘分布的距离, 联合分布自适应 (joint distribution adaptation, JDA) [10] 通过同时减少边缘 MMD 和条件 MMD 的距离来减小域间差异, 其他工作通过添加正则化 [11]、结构一致性 [12]、标签传播 [13] 等来扩展 JDA. 随着深度学习方法的不断发展, MMD 也被广泛应用于深度迁移学习中, 如: 深度适配网络 (deep adaptation networks, DAN) [14] 引入多核变量 MMD, 通过多核优化方法使不同域之间的距离最小化; 加权域适应网络 (weighted domain adaptation network) [15] 利用源和目标域上的类别先验改进 MMD; 联合适配网络 (joint adaptation networks, JAN) [16] 利用多个域特定层的联合 MMD 来学习迁移网络.

谱学习 (spectral learning) [17] 算法是近年来机器学习领域的一个研究热点, 其建立在谱理论的基础上. 谱学习将连续空间上的复杂问题表示为一系列离散空间上的简单问题的组合, 通过求解这些简单问题获得其近似数值解, 而且能够保证多项式的计算复杂度 [18]. 相比 MMD, 谱学习对于分布差异的表征具有理论上的优势, 虽然同样可以将欧氏空间中的相关对齐问题推广到无限维特征空间, 但 MMD 通过核空间中两个域间的概率分布均值距离来度量概率分布差异, 仅仅能匹配不同概率分布的各阶矩, 而利用谱学习可以获得高维数据的低维谱表示. 一方面, 它能保持数据内部潜在结构不变, 另一方面, 由于是从全局结构出发考虑问题, 因此不同于传统基于局部结构和数值平均的学习算法, 它能够获得问题的全局最优解 [19]. Long 等 [5] 提出的迁移核学习 (transfer kernel learning, TKL) 是一种数据依赖的谱学习方法. TKL 将源域真实核矩阵和目标域插值核矩阵的近似误差作为域间概率分布差异的度量, 不需要参数化距离度量的分布密度估计过程, 且较非参数化距离 MMD 能更精细地度量概率分布差异. 但是 TKL 只适配了边缘分布, 忽略了条件概率分布的对齐. 因此, 有必要对基于谱学习的联合分布适配进行更深入的研究.

另一方面, 迁移学习的顺利实施通常需要源域和目标域直接共享原始数据, 这在一些情况下无法满足, 特别当涉及到一些有保密性要求或者敏感性的数据时. 世界各国正在加强数据安全和隐私的法律保护, 欧盟《通用数据保护条例》等隐私法规的出台严重制约了企业间多方合作训练模型的模式, 也使得直接的迁移学习应用不再有效. 已有的机器学习隐私保护算法主要考虑标准分类学习中的隐私经验风险最小化 (empirical risk minimization, ERM), 即试图在分类器的准确性和与训练集相关的差分隐私 (differentially private) [20] 保障之间优化平衡 [21]. 比如, Chaudhuri 等 [22] 通过输出扰动和目标扰动实现带 l_2 正则化项的 Logistic 回归和支持向量机 (support vector machine, SVM) 的 ERM 差分隐私, 其结果要求正则化器具有强凸性和可微性; Wang 和 Zhang [23] 进一步针对凸和非凸稀疏分类方法, 基于 ADMM (alternating direction method of multiplier) 算法将稀疏问题的求解转化为多步迭代

过程, 实现了带 l_1 和 $l_{1/2}$ 正则化项的 Logistic 回归的 ERM 差分隐私. 最近的研究工作集中在分布式机器学习算法下的 ERM 差分隐私方法 [24,25], 但目前这些工作基本都没有考虑数据源之间的分布差异, 因而与迁移学习的目标并不一致.

迁移学习过程中的隐私保护研究非常稀少, 最近的相关工作是 Wang 等 [26] 提出的用于 Logistic 回归的差分隐私假设迁移学习 (hypothesis transfer learning) 方法, 他们借助公共无标签源数据集度量源和目标之间的关系, 利用在源域上训练的假设 (应用输出扰动方法 [22] 来扰动源假设) 改进目标假设的学习. 其他相关研究关注迁移学习的变体, 如迭代式的差分隐私多任务学习 [27]、考虑协方差偏移 (covariate shift) 的分布式训练数据汇总 [28], 这方面的工作或者没有考虑数据源间的分布差异, 或者没有考虑数据的条件分布和模型的判别能力, 都与隐私保护迁移学习任务有一定的区别.

本文假设数据持有方和查询方分别提供源域数据和目标域数据, 目标是利用数据持有方的特征数据和样本标签预测查询方数据的标签, 由一个聚合器将双方数据集中起来进行模型训练. 本文的主要工作和贡献如下:

(1) 提出一种基于联合分布核适配的无监督迁移学习算法, 拓展了 TKL [5], 在再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS) 中同时对齐边缘分布和条件分布, 以学习一个域不变核矩阵. 具体来说, 本文利用源域弱分类器为目标数据赋伪标记, 通过引入同类标记矩阵度量类条件分布距离, 使得领域间的判别结构能更好的对应上. 然后引入自适应权衡参数 [29], 构建联合分布核适配的优化目标. (2) 为上述无监督迁移核学习算法提供一定的隐私保护. 本文引入 Rahimi-Recht 随机函数 [30] 得到源域和目标域数据的随机投影, 使得数据持有方和查询方仅需将数据点的随机投影值与聚合器共享. 随后, 聚合器通过优化一个基于目标扰动的 SVM 分类器, 在为无标记目标数据构建弱分类器的同时实现了对源域数据的差分隐私保护. (3) 在公开的文本和图像数据集上进行了充分的实验分析, 通过与多个基准和先进算法的分类精度对比以及系统的参数敏感性分析, 验证了本文所提方法的有效性.

2 支持隐私保护的迁移核学习算法

假设数据持有方和查询方分别提供源域数据 $\mathcal{S} = \{(\boldsymbol{x}_{s1}, y_1), \ldots, (\boldsymbol{x}_{sm}, y_m)\}$ 和目标域数据 $\mathcal{T} = \{\boldsymbol{x}_{t1}, \ldots, \boldsymbol{x}_{tn}\}$, 即目标域仅有无标记数据, 由一个聚合器将双方数据集中起来进行建模和预测. 本节首先提出一种基于联合分布核适配的迁移学习方法, 并给出求解方法. 然后, 考虑到数据持有方和查询方不希望在数据聚合过程中暴露各自的敏感信息, 提出了迁移算法的隐私保护程序.

2.1 迁移核学习算法

假设 $\mathcal{S}$ 和 $\mathcal{T}$ 的特征空间 $\mathcal{F}_S = \mathcal{F}_T$, 同时标签空间 $\mathcal{Y}_S = \mathcal{Y}_T$, 但是边缘概率分布 $P(\boldsymbol{x}_s) \neq P(\boldsymbol{x}_t)$, 并且条件概率分布 $P(y|\boldsymbol{x}_s) \neq P(y|\boldsymbol{x}_t)$. 为了在 RKHS 中对齐源域和目标域的概率分布, 即 $P(\phi(\boldsymbol{x}_s)) \simeq P(\phi(\boldsymbol{x}_t))$, 迁移核学习算法直接在 RKHS 中匹配源域和目标域的核矩阵.

2.1.1 目标函数构建

首先对目标核 $\boldsymbol{K}_T$ 进行标准特征分解得到特征系统 $\{\boldsymbol{\Lambda}_T, \boldsymbol{\Phi}_T\}$:

$$\boldsymbol{K}_T \boldsymbol{\Phi}_T \simeq \boldsymbol{\Phi}_T \boldsymbol{\Lambda}_T. \tag{1}$$

根据 Mercer 定理计算 $\{\boldsymbol{\Lambda}_T, \boldsymbol{\Phi}_T\}$ 在源域数据集 $\mathcal{S}$ 的取值, 得到源核 $\boldsymbol{K}_S$ 特征向量的插值近似:

$$\overline{\boldsymbol{\Phi}}_S \simeq \boldsymbol{K}_{ST}\boldsymbol{\Phi}_T\boldsymbol{\Lambda}_T^{-1}, \tag{2}$$

其中 $\boldsymbol{K}_{ST}$ 是由核表示的 $\mathcal{S}$ 和 $\mathcal{T}$ 之间的跨域相似矩阵. 特征向量 $\overline{\boldsymbol{\Phi}}_S \in \mathbb{R}^{m\times n}$ 保留了目标域的关键结构信息, 将作为知识迁移的桥梁.

Nyström 近似 [31] 基于 Mercer 定理用已知核的特征系统对目标核进行低秩近似. 在标准 Nyström 法中 $\boldsymbol{K}_S \simeq \overline{\boldsymbol{\Phi}}_S\boldsymbol{\Lambda}_T\overline{\boldsymbol{\Phi}}'_S$, 其中 "$'$" 表示转置. Long 等 [5] 提出的 TKL 方法通过将特征谱 $\boldsymbol{\Lambda}_T$ 松弛为待学习参数 $\boldsymbol{\Lambda}$, 经过谱核设计 [32] 得到一族由 $\{\boldsymbol{\Lambda}_T, \boldsymbol{\Phi}_T\}$ 外插值到源域数据集上生成的核矩阵:

$$\overline{\boldsymbol{K}}_S = \overline{\boldsymbol{\Phi}}_S\boldsymbol{\Lambda}\overline{\boldsymbol{\Phi}}'_S. \tag{3}$$

生成的核矩阵 $\overline{\boldsymbol{K}}_S \in \mathbb{R}^{m\times m}$ 通过 $\overline{\boldsymbol{\Phi}}_S$ 保留了目标核 $\boldsymbol{K}_T$ 的关键结构, 同时其维度被灵活地改变至与源核维度相同. 进一步, 通过最小化 Nyström 近似误差确定了最小化分布差异的最优特征值, 从而实现领域间边缘分布的对齐:

$$\min_{\boldsymbol{\Lambda}} \left\| \overline{\boldsymbol{\Phi}}_S\boldsymbol{\Lambda}\overline{\boldsymbol{\Phi}}'_S - \boldsymbol{K}_S \right\|_F^2, \tag{4}$$

其中 $\|\cdot\|_F$ 表示矩阵的 Frobenius 范数 (F – 范数).

最小化 Nyström 近似误差等价于要求源域和目标域数据充分重叠, 因此 TKL 实现了源域和目标域聚类结构重叠程度最大化 [5]. 然而由于源域和目标域之间的判别超平面也极有可能并不相同, 因此仅最小化边缘分布之间的差异并不能获得满足应用需求的迁移学习性能, 条件分布间的距离也应显式最小化.

由于目标域仅有无标记数据, 无法直接评价目标域的条件概率分布 $P(y_t|\phi(\boldsymbol{x}_t))$. 因此, 我们用类条件概率 $P(\phi(\boldsymbol{x}_t)|(y_t = c))$ 近似 $P(y_t|\phi(\boldsymbol{x}_t))$, 当样本个数足够大时, 二者有着很好的相似性 [29]. 为了得到 $P(\phi(\boldsymbol{x}_t)|(y_t = c))$, 首先在源域 $\mathcal{S}$ 上训练一个弱分类器, 然后将此弱分类器应用到目标域 $\mathcal{T}$ 上进行预测, 得到目标数据的伪标记. 由于分布差异的存在, 这些伪标记的置信度可能不高, 但它是当前学习条件下所能获得的最佳分类结果. 通过迭代式地修正预测结果, 一般能进一步提高分布适配程度和目标域分类正确率.

为了在核矩阵适配过程中体现类别信息, 定义一个同类标记矩阵 $\boldsymbol{U}(a,b) \in \mathbb{R}^{m\times n}$, 如果 y_{s_i} 和 y_{t_j} 的伪标记属于同一类, 那么 $\boldsymbol{U}_{ij} = a \geqslant 1$, 否则 $0 \leqslant \boldsymbol{U}_{ij} = b < 1$. (a,b) 的取值可以根据具体分类任务的特点确定, 但其值应接近 1.

将该同类标记矩阵与跨域核矩阵 $\boldsymbol{K}_{ST}$ 相结合, 得到携带了源域和目标域类别标记信息以及样本相似度的矩阵. 紧接着在源域样本上计算目标域特征系统的外插值:

$$\overline{\boldsymbol{\Phi}}_{SU} \simeq \boldsymbol{U} \odot \boldsymbol{K}_{ST}\boldsymbol{\Phi}_T\boldsymbol{\Lambda}_T^{-1}, \tag{5}$$

其中 $\odot$ 是矩阵的 Hadamard 积算子. $\overline{\boldsymbol{\Phi}}_{SU} \in \mathbb{R}^{m\times n}$ 不仅包含了目标域的关键性结构信息, 同时也包含了源域和目标域的类别信息. 注意到若同类标记矩阵中的参数 $b > 0$, 会突出两个域之间同类数据结构的重要性但仍然考虑不同类别数据的数据结构; 若 $b = 0$, 则表示不考虑不同类别数据的数据结构. 进一步通过源域样本所属类别样本划分 $\overline{\boldsymbol{\Phi}}_{SU}$, 得到类别 $g \in \{1, \ldots, G\}$ 对应的 $\overline{\boldsymbol{\Phi}}_{SU}^{(g)}$, 将其用于度量类条件分布的距离:

$$\sum_{g=1}^{G} \frac{m^{(g)} + n^{(g)}}{m+n} \left\| \overline{\boldsymbol{\Phi}}_{SU}^{(g)}\boldsymbol{\Lambda}\overline{\boldsymbol{\Phi}}_{SU}^{(g)'} - \boldsymbol{K}_S^{(g)} \right\|_F^2, \tag{6}$$

其中 $m^{(g)}$ 和 $n^{(g)}$ 分别表示源域和目标域中 (伪) 标记属于类别 g 的样本数.

结合式 (4) 和 (6), 并以边缘分布和条件分布的差异程度作为权衡参数, 得到联合分布核适配目标函数:

$$\begin{aligned}&\min_{\boldsymbol{\Lambda}} \sum_{g=0}^{G} \frac{\mu^{(g)}}{\sum_{g=0}^{G}\mu^{(g)}} \left\| \overline{\boldsymbol{\Phi}}_{SU}^{(g)} \boldsymbol{\Lambda} \overline{\boldsymbol{\Phi}}_{SU}^{(g)'} - \boldsymbol{K}_S^{(g)} \right\|_F^2,\\ &\text{s.t. } \lambda_i \geqslant \zeta \lambda_{i+1},\ i=1,\ldots,n-1,\\ &\qquad \lambda_i \geqslant 0,\ i=1,\ldots,n.\end{aligned} \tag{7}$$

目标函数中, $\mu^{(g)} = \frac{m^{(g)}+n^{(g)}}{m+n} d_A(\mathcal{S}^{(g)}, \mathcal{T}^{(g)})$ 表示自适应权衡参数, $g=0$ 对应边缘分布情况, 即 $\overline{\boldsymbol{\Phi}}_{SU}^{(0)} = \overline{\boldsymbol{\Phi}}_S$; $x^{(0)} = x$, $x \in \{m, n, \mathcal{S}, \mathcal{T}\}$. 由于目标域中无数据标签, 难以通过交叉验证 (cross validation) 策略优化其取值, 因此本文利用 Wang 等 [29] 提出的动态特征对齐方法: 利用 $\mathcal{A}$ – 距离作为分布差异的度量方式, $d_A(\mathcal{S}, \mathcal{T}) = 2(1-2\epsilon(h))$, 其中 $\epsilon(h)$ 定义为一个线性分类器判别两个域 $\mathcal{S}, \mathcal{T}$ 的错误率 (源域和目标域数据分别为正、负类). 由于正定核的特征谱遵循幂律分布衰减 [33], 因此在式 (7) 中对要求解的特征谱施加了特征值阻尼约束, 其中 $\zeta \geqslant 1$ 为特征谱阻尼系数, 编码了 $\overline{\boldsymbol{K}}_S$ 的先验衰减趋势, 同时鼓励较大的特征值对应的特征向量对知识迁移做出更大的贡献 [5].

2.1.2 域不变核矩阵

记 $\overline{\boldsymbol{\Phi}}_A \triangleq [\overline{\boldsymbol{\Phi}}_S; \boldsymbol{\Phi}_T]$ 为源域和目标域全集 $\mathcal{A} = \mathcal{S} \cup \mathcal{T}$ 上的特征向量. 基于谱核设计, 将学习到的最优特征谱 $\boldsymbol{\Lambda}$ 与 $\overline{\boldsymbol{\Phi}}_A$ 结合为域不变特征系统, 构建域不变核矩阵 $\overline{\boldsymbol{K}}_A$:

$$\overline{\boldsymbol{K}}_A = \begin{bmatrix} \overline{\boldsymbol{K}}_S & \overline{\boldsymbol{K}}_{ST} \\ \overline{\boldsymbol{K}}_{TS} & \overline{\boldsymbol{K}}_T \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{\Phi}}_S \boldsymbol{\Lambda} \overline{\boldsymbol{\Phi}}_S' & \overline{\boldsymbol{\Phi}}_S \boldsymbol{\Lambda} \boldsymbol{\Phi}_T' \\ \boldsymbol{\Phi}_T \boldsymbol{\Lambda} \overline{\boldsymbol{\Phi}}_S' & \boldsymbol{\Phi}_T \boldsymbol{\Lambda} \boldsymbol{\Phi}_T' \end{bmatrix} = \overline{\boldsymbol{\Phi}}_A \boldsymbol{\Lambda} \overline{\boldsymbol{\Phi}}_A'. \tag{8}$$

$\overline{\boldsymbol{K}}_S$ 上构建的标准核机器 (如 SVM) 可以直接应用于跨域核矩阵 $\overline{\boldsymbol{K}}_{TS} = \boldsymbol{\Phi}_T \boldsymbol{\Lambda} \overline{\boldsymbol{\Phi}}_S'$ 上, 相当于推广到了重构的目标域上, 因此有

$$\hat{\boldsymbol{y}}_t = \overline{\boldsymbol{K}}_{TS} (\widetilde{\boldsymbol{\alpha}} \odot \boldsymbol{y}_s) + \widetilde{\boldsymbol{b}}, \tag{9}$$

其中 $\widetilde{\boldsymbol{\alpha}}$, $\widetilde{\boldsymbol{b}}$ 分别表示新生成的核 SVM 的拉格朗日乘子向量和分类面截距.

2.1.3 优化目标求解

类似 TKL 的求解方法, 将目标函数(7)化为带不等式约束的二次规划形式求解, 记 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ 为 n 个待学习的特征谱参数, 即 $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. 首先将约束条件 $\lambda_i \geqslant \zeta \lambda_{i+1}$, $i = 1, \ldots, n-1$ 记为 $(\boldsymbol{I} - \zeta \overline{\boldsymbol{I}})\boldsymbol{\lambda} \geqslant 0$, 其中 $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ 是单位矩阵, $\overline{\boldsymbol{I}} \in \mathbb{R}^{n \times n}$ 的全部非零元素为 $\overline{I}_{i,i+1} = 1$, $i = 1, \ldots, n-1$. 进一步令

$$\begin{aligned} \boldsymbol{Q} &= \sum_{g=0}^{G} \frac{\mu^{(g)}}{\sum_{g=0}^{G}\mu^{(g)}} \left(\overline{\boldsymbol{\Phi}}_{SU}^{(g)'} \overline{\boldsymbol{\Phi}}_{SU}\right) \odot \left(\overline{\boldsymbol{\Phi}}_{SU}^{(g)'} \overline{\boldsymbol{\Phi}}_{SU}^{(g)}\right), \\ \boldsymbol{v} &= \sum_{g=0}^{G} \frac{\mu^{(g)}}{\sum_{g=0}^{G}\mu^{(g)}} \mathrm{diag}\left(\overline{\boldsymbol{\Phi}}_{SU}^{(g)'} \boldsymbol{K}_S \overline{\boldsymbol{\Phi}}_{SU}^{(g)}\right), \\ \boldsymbol{C} &= \begin{bmatrix} \boldsymbol{I} - \zeta \overline{\boldsymbol{I}} \\ \boldsymbol{I} \end{bmatrix}. \end{aligned} \tag{10}$$

由此得到了带不等式约束的标准二次规划问题:

$$\begin{aligned}&\min_{\boldsymbol{\lambda}} \boldsymbol{\lambda}'\boldsymbol{Q}\boldsymbol{\lambda} - 2\boldsymbol{v}'\boldsymbol{\lambda}\\&\text{s.t. } \boldsymbol{C}\boldsymbol{\lambda} \geqslant \boldsymbol{0}.\end{aligned} \tag{11}$$

实际问题中核矩阵通常显示出 "谱隙" (eigengap) 的特点, 即其最大 r 个特征值远大于余下的特征值 [33]. 因此我们保留最大的 r 个特征系统, 相关变量维度降为 $\overline{\boldsymbol{\Phi}}_S \in \mathbb{R}^{m\times r}$, $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$, $\boldsymbol{\lambda} \in \mathbb{R}^{r\times 1}$, 从而显著降低计算开销, 同时减轻噪声的影响.

对式 (11) 使用拉格朗日乘子法可得到其对偶问题 (dual problem) 形式. 首先将该问题的拉格朗日函数写为

$$L(\boldsymbol{\lambda}, \boldsymbol{\alpha}) = \boldsymbol{\lambda}'\boldsymbol{Q}\boldsymbol{\lambda} - 2\boldsymbol{v}'\boldsymbol{\lambda} + \boldsymbol{\alpha}'\boldsymbol{C}\boldsymbol{\lambda}, \tag{12}$$

其中 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)'$, $\alpha_i \geqslant 0$ 是拉格朗日乘子. 原问题就变成了 $\min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}, \boldsymbol{\alpha})$, 定义 $\boldsymbol{\lambda}^*$ 为原问题的最优解, $\boldsymbol{\alpha}^*$ 为对偶问题的最优解. 根据 Hadamard 乘积矩阵的性质, 易证 $\boldsymbol{Q}$ 为正定矩阵, 则式 (11) 为严格的凸二次规划问题, 具有唯一的全局最优解 $\boldsymbol{\lambda}^*$, 则 $\boldsymbol{\lambda}^*, \boldsymbol{\alpha}^*$ 分别是原始问题和对偶问题的最优解的充要条件是 $\boldsymbol{\lambda}^*$, $\boldsymbol{\alpha}^*$ 满足 KKT (Karush-Kuhn-Tucker) 条件:

$$\begin{cases}\boldsymbol{\lambda}^* = \boldsymbol{Q}^{-1}\boldsymbol{v} - \dfrac{1}{2}\boldsymbol{Q}^{-1}\boldsymbol{C}'\boldsymbol{\alpha},\\ \boldsymbol{\alpha}^{*\prime}\boldsymbol{C}\boldsymbol{\lambda}^* = \boldsymbol{0},\\ \boldsymbol{C}\boldsymbol{\lambda}^* \geqslant \boldsymbol{0},\\ \alpha_i^* \geqslant 0.\end{cases} \tag{13}$$

2.2 差分隐私迁移核学习算法

在本文的迁移算法中, 由第三方聚合器计算源域和目标域数据的交叉核矩阵, 这需要数据持有方和查询方向聚合器提供自己的原始特征数据. 此外, 聚合器还将利用源域数据训练一个初始分类器, 进而实现条件分布适配, 但这会不可避免地导致聚合器可以接触到源域的特征数据和对应的样本标签, 带来隐私泄露风险. 为了解决这些问题, 本小节提出了本文迁移算法的隐私保护程序.

2.2.1 数据源: 随机函数投影原数据

为了避免聚合器需要得到数据持有方和查询方的原始特征数据以计算交叉核矩阵和自适应权衡参数 μ, 受 Sarpatwar 等 [28] 的启发, 设计数据持有方和查询方都访问一个随机函数 $h(\cdot)$, 并共用相同的随机种子. $h(\cdot)$ 根据核函数显式构造特征映射, 保证任意两点随机特征映射的内积近似对应的核函数值. 利用 Rahimi 和 Recht [30] 所提随机傅里叶特征映射近似核函数的方法, 设计随机函数 $h(\cdot)$ 使得下式成立:

$$h(\boldsymbol{x})'h(\boldsymbol{y}) = \mathrm{E}_\omega[\zeta_\omega(\boldsymbol{x})\zeta_\omega(\boldsymbol{y})^*] = \int_{\mathbb{R}^d} p(\omega)\mathrm{e}^{-\mathrm{i}\omega'(\boldsymbol{x}-\boldsymbol{y})}\mathrm{d}\omega = k(\boldsymbol{x}, \boldsymbol{y}), \tag{14}$$

其中 $\zeta_\omega(\boldsymbol{x}) = \mathrm{e}^{-\mathrm{i}\omega'\boldsymbol{x}}$, $k(\cdot)$ 表示某一平移不变核函数, $p(\omega)$ 为其诱导的概率密度函数. 高斯核作为常用的核函数, 具有优良的函数表达能力, 当给出的数据缺少先验知识时, 高斯核往往是一种较好的选择, 能够给出平滑的估计, 因此本文选用高斯核构造核矩阵用于表征输入数据. 对于高斯核函数 $k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}$, 有 $\omega \sim \mathcal{N}(\boldsymbol{0}, 2\gamma\boldsymbol{I})$, 其中 $\boldsymbol{I}$ 表示单位矩阵.

Rahimi 和 Recht [30] 给出随机傅里叶特征的一种构造方法为 $z_\omega(\boldsymbol{x})=\sqrt{2}\cos(\omega'\boldsymbol{x}+b)$, b 服从 $[0,2\pi]$ 上的均匀分布, 如此可满足 $\mathrm{E}_\omega[z_\omega(\boldsymbol{x})'z_\omega(\boldsymbol{y})]=k(\boldsymbol{x},\boldsymbol{y})$. 为了进一步减小 $z_\omega(\boldsymbol{x})'z_\omega(\boldsymbol{y})$ 的方差, 将 D 个独立同分布的 z_ω 串联成一个列向量并进行标准化, 得到了每个数据点的随机投影

$$h(\boldsymbol{x})=\sqrt{\frac{2}{D}}\big[\cos(\omega_1'\boldsymbol{x}+b_1),\cos(\omega_2'\boldsymbol{x}+b_2),\ldots,\cos(\omega_D'\boldsymbol{x}+b_D)\big]', \tag{15}$$

其中 D 为随机特征空间的维度. 数据持有方和查询方随后将随机投影 $\boldsymbol{H}_S=[h(\boldsymbol{x}_{s1}),\ldots,h(\boldsymbol{x}_{sm})]$, $\boldsymbol{H}_T=[h(\boldsymbol{x}_{t1}),\ldots,h(\boldsymbol{x}_{tn})]$ 传输给聚合器. 如此, 保证了聚合器无需获取用户的原始特征数据以及随机函数 $h(\cdot)$ 的种子, 即可进行基于非线性核的分类器构建和联合分布核适配操作.

2.2.2 聚合器: 源域差分隐私分类器发布

聚合器需要利用源域随机特征数据构建基于平移不变核的分类器 (下文以 SVM 为例), 它将有权限访问源域样本标签 $\boldsymbol{y}$. 若聚合器可以直接得到原始分类器, 则可能暴力求解出原始数据. 因此设计让聚合器执行一个隐私 SVM 分类器, 使得聚合器仅能得到添加噪声的分类器[1)].

采用 Chaudhuri 等 [22] 提出的针对非线性核的目标扰动 ERM 差分隐私算法, 聚合器使用源域随机特征和标签数据 $\mathcal{D}=\{(h(\boldsymbol{x}_{si}),y_i)|i=1,2,\ldots,m\}$, 发布差分隐私 SVM 分类器. 添加噪音扰动的目标函数 $\Omega_{\mathrm{priv}}(\boldsymbol{w},\mathcal{D})$ 表示如下:

$$\begin{aligned}\min_{\boldsymbol{w}}\Omega_{\mathrm{priv}}(\boldsymbol{w},\mathcal{D})&=\Omega(\boldsymbol{w},\mathcal{D})+\frac{1}{m}\boldsymbol{\varphi}'\boldsymbol{w},\\ \Omega(\boldsymbol{w},\mathcal{D})&=\frac{1}{2}\|\boldsymbol{w}\|_2^2+C\sum_{i=1}^{m}\ell\big(\boldsymbol{w}'h(\boldsymbol{x}_{si}),\boldsymbol{y}_i\big),\end{aligned} \tag{16}$$

其中 $\ell(\cdot)$ 表示损失函数, C 为正则化常数, 是用来控制损失函数的惩罚系数. $\boldsymbol{\varphi}$ 服从概率密度为 $v(\boldsymbol{\varphi})=\frac{1}{\alpha}\mathrm{e}^{-\beta\|\boldsymbol{\varphi}\|_2}$ 的分布, 其中 α 为标准化常数, β 是隐私预算 ε 的函数. ε 量化了隐私要求, 更小的 ε 意味着更高的隐私保护需求. 由于 SVM 的损失函数 $\ell(\cdot)$ 连续但不可微, 利用 Huber 损失函数代替[2)].

将求解式 (16) 得到的近似最优预测器记为 $\boldsymbol{w}_{\mathrm{priv}}$, 聚合器利用 $\boldsymbol{w}_{\mathrm{priv}}$ 和目标数据随机投影 $\boldsymbol{H}_T$ 便可得到目标数据的伪标记. 值得注意的是, 利用本文迁移算法得到域不变核矩阵后, 还将在其上构建一次标准分类器, 但由于输入的是重构的核矩阵, 聚合器无法获取原始特征数据.

将本文所提基于差分隐私保护的核迁移算法总结为算法 1.

2.3 算法时间复杂度分析

记 m,n 分别为源域和目标域的样本规模, d,D 分别为样本维度和随机特征空间维度, r 为保留最大特征向量的数目. 则算法 1 步骤 2 中随机特征映射的时间复杂度为 $O((m+n)Dd)$. 步骤 3 中, 构建目标扰动的差分隐私 SVM 分类器, 主要的计算负担在于扰动优化目标的求解. 由于仅在优化迭代处添加噪声扰动, 因此计算复杂度应与不添加差分隐私的 SVM 相同, 随机特征的运用使得该部分 SVM 的计算过程类似线性核, 因此求得数值解的时间复杂度可以认为是 $O(mD)$ [34]. 步骤 4 中的 3 个核矩阵计算共需要 $O(D(m+n)^2)$ 时间. 式 (1) 目标核的特征值分解理论上需要 $O(rn^2)$ 时间. 式 (2) 和 (5) 中目标域特征系统的外插值需要 $O(mnr)$ 时间. 二次规划问题求解需要 $O(rn^2+r^3)$ 时间 [5]. 步骤 5

1) 除本文选择的 SVM 分类器以外, 其他的核机器, 如核脊回归 (kernel ridge regression), 或者这些方法的高效变种算法, 可以通过施加差分隐私保护整合到本文提出的算法框架中.

2) 具体算法见 Chaudhuri 等 [22] 中的 Algorithm 3.

Algorithm 1 Differentially private transfer kernel learning

Input: Source data $\mathcal{S}=\{(\boldsymbol{x}_{s1},y_1),\ldots,(\boldsymbol{x}_{sm},y_m)\}$ provided by data holder, target data $\mathcal{T}=\{\boldsymbol{x}_{t1},\ldots,\boldsymbol{x}_{tn}\}$ by inquirer, eigen-damping factor ζ, Gaussian kernel parameter γ, random space dimension D, privacy parameters ε,Λ,c.

Output: Predicted labels $\hat{\boldsymbol{y}}_t$ for target data $\mathcal{T}$.

1: Draw D iid $\{\omega_i\}_{i=1}^{D}$ and $\{b_i\}_{i=1}^{D}$ from $p(\omega)=\mathcal{N}(\boldsymbol{0},2\gamma\boldsymbol{I}_d)$ and $[0,2\pi]$, respectively. Parameters $\{\omega_i,b_i\}$ are shared between the data holder and the inquirer;

2: **Data holder and inquirer** respectively calculate random projections of $\mathcal{S}$ and $\mathcal{T}$ by (15), obtain $\boldsymbol{H}_S=[h(\boldsymbol{x}_{s1},\ldots,h(\boldsymbol{x}_{sm})]$, $\boldsymbol{H}_T=[h(\boldsymbol{x}_{t1},\ldots,h(\boldsymbol{x}_{tm})]$, send to the aggregator;

3: **Aggregator** establishes ε-differentially private approximate SVM classifier as (16) with $\{\boldsymbol{H}_S,\boldsymbol{y}\}$ and privacy parameters ε,Λ,c, releases approximate minimizer $\boldsymbol{w}_{\text{priv}}$ and then obtains the pseudo-label of the target data using $\boldsymbol{H}_T$;

4: **Aggregator** computes approximate kernel matrices $\boldsymbol{K}_S=\boldsymbol{H}_S'\boldsymbol{H}_S$, $\boldsymbol{K}_T=\boldsymbol{H}_T'\boldsymbol{H}_T$, $\boldsymbol{K}_{ST}=\boldsymbol{H}_S'\boldsymbol{H}_T$ and adaptive tradeoff parameter μ, solves quadratic progamming problem (11) for optimal eigenspectrum $\boldsymbol{\Lambda}$;

5: **Aggregator** computes domain-invariant kernel matrix $\overline{\boldsymbol{K}}_A$ as (8), returns the predicted label $\hat{\boldsymbol{y}}_t$ by (9) to inquirer.

中域不变核矩阵的构建需要 $O(r(m+n)^2)$ 时间, 目标数据的预测中主要的计算负担在于利用 $\overline{\boldsymbol{K}}_S$ 进行 SVM 的训练, 采用 SMO 求解器, 其计算时间复杂度约为 $O(m^2)$. 因此算法 1 主要的计算时间复杂度约为 $O((D+r)(m+n)^2+(m+n)Dd)$.

标准核矩阵的计算复杂度是其样本规模的二次阶, Long 等 [5] 提出其简化计算方法, 核心是对数据采样计算核矩阵再利用 Nyström 方法进行近似, 该技术手段也可嵌入至本文方法以降低计算复杂度.

3 实验

本节将本文算法运用到公开的文本和图像迁移学习基准数据集上, 通过与多个经典方法和前沿成果的对比实验和参数分析, 验证了本文算法的有效性, 这包括迁移学习算法的有效性和差分隐私保护程序的有效性.

3.1 迁移学习算法有效性

3.1.1 实验数据

文本数据集: Reuters-21578[3] 该语料库收录了 1987 年的路透社新闻文章, 是目前使用最广泛的文本分类标准测试集之一. Reuters-21578 包含多个大类和子类, 其中最大的 3 个大类为 orgs, people 和 place. 将 3 个大类两两随机组合作为源域和目标域, 可生成 6 个跨域二分类任务 orgs vs. people, orgs vs. place, people vs. place. 为了公平的比较, 均采用 Gao 等 [35] 发布的预处理集.

图像数据集: Office 和 Caltech-256[4] Office 是图像迁移学习的主流基准数据集, 由 Amazon (在线电商图片)、DSLR (单反相机拍摄的高解析度图片) 和 Webcam (网络摄像头拍摄的低解析度图片) 3 个对象领域组成, 共计 31 个类别. Caltech-256 是对象识别的基准数据集, 包括 1 个对象领域 Caltech, 共计 256 个类别, Caltech-256 和 Office 库有 10 个共同的类, 这两个库总共可以分成 4 个域: A (Amazon), C (Caltech-256), D (DSLR) 和 W (Webcam). 将这 4 个域两两随机组合作为源域和目标域, 共可构造 12 个跨域图像识别任务. 实验采用了 Gong 等 [36] 发布的 Office+Caltech 预处理数据集. 对每张图片提取 SURF 特征并向量化为 800 维的直方图表征, 所有直方图向量都进行了标准化处理, 直方图码表由 K 均值聚类算法在 Amazon 子域上生成. 上述实验数据集的统计信息见表 1.

3) http://www.daviddlewis.com/resources/testcollections/reuters21578/.

4) https://github.com/jindongwang/transferlearning/blob/master/data.

表 1 跨域文本、图像数据集统计信息

Table 1 Statistics of cross-domain text and image data sets

Type	Data set	Subsets	#Instance	#Feature	#Class
Text	Reuters-21578	Orgs	1237	4771	2
		People	1208		
		Place	1016		
Image	Office	Amazon	958	800	10
		Webcam	295		
		DSLR	157		
	Caltech-256	Caltech-256	1123		

3.1.2 性能比较

与本文所提迁移学习方法比较的经典方法和前沿成果包括: (1) 传统的分类学习算法 SVM; (2) 半监督学习算法拉普拉斯支持向量机 (Laplacian support vector machine, LapSVM) [37]; (3) 非核学习迁移学习方法: 跨领域谱分类 (cross-domain spectral classification, CDSC) [38], 谱特征对齐 (spectral feature alignment, SFA) [39], 核均值匹配 (kernel mean matching, KMM) [40], TCA [9]; (4) 核学习迁移学习方法: 领域多核学习 (domain transfer multiple kernel learning, DTMKL) [41], 测地线流式核 (geodesic flow kernel, GFK) [36], TKL [5].

为保证对比实验的公平性, 采用相关参考文献 [5, 9, 41] 的测试协议, 统一将 SVM 作为基准分类器, 采用目标域的分类准确率作为各算法的评价指标. 由于目标域数据均无标签, 在实验中无法利用交叉验证 (cross validation) 来选择最优的参数, 因此通过对各方法的参数空间, 如 SVM 的正则参数 C, CDSC, SFA, TCA, GFK 的子空间参数 k, 进行网格搜索找到其最佳参数, 将每种算法在各种参数设置下的最佳效果用于性能对比. TKL 和本文方法有共同的模型参数: 特征普阻尼系数 ζ 和 SVM 正则参数 C, 统一设置为文本实验中 $\zeta = 2.0$, $C = 5.0$, 图像实验中 $\zeta = 1.1$, $C = 10.0$. 此外, 核矩阵保留的最大特征系统个数 r 均设置为 $\min\{200, n\}$, 其中 n 表示目标域样本数. 对所有核方法, 统一在文本数据集上采用线性核 $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{y}$, 在图像数据集上采用高斯核 $k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2}$, 高斯核参数设置为 $\gamma = \frac{1}{A}$, 其中 A 是所有训练样本的平均欧式距离 [3]. 本文自适应权衡参数 μ 由线性 SVM 分类器计算得到, 同类标记矩阵 $\boldsymbol{U}(a, b)$ 的参数统一设置为 $a = 1$, $b = 0.9$.

比较各算法在 18 个分类问题上的性能, 如表 2 所示[5)]. 显然, 传统分类算法 SVM 未能在大多数任务上取得足够高的准确率, 主要原因是训练数据和测试数据来自不同概率分布, 使得标准分类学习的可靠性下降. 流形正则化半监督学习 LapSVM 通过目标域无标数据挖掘目标域的判别结构, 取得了比传统分类方法更佳的效果. 其次, 迁移学习方法通常较标准学习问题性能更优, 但是在文本和图像数据集实验中的表现不尽相同.

文本数据集. 由 Reuters-21578 数据集构造了 3 个跨域分类任务组, 每组包含 2 个文本分类任务. 每个任务组的平均分类准确率如表 2 所示. 可观察到本文方法在所有 6 个分类任务的平均准确率为 77.78%, 相比于标准算法 SVM 提升了 9.38%, 仅次于 TKL 的 77.79%. 事实上, Reuters-21578 数据集的每个大类都包含多样化的子类, 导致 CDSC, SFA, GFK 等方法将难以抽取领域间共享的隐含结构. 利用 MMD 准则对齐分布的 KMM, TCA 和 DTMKL 方法, 由于不能适配复杂结构, 没有取得理想的分类效果. TKL 和本文方法通过将目标域插值得到的核矩阵与源域真实的核矩阵进行匹配, 增大领域

5) 部分实验结果数据引用自 Long 等 [5].

表 2 文本和图象数据集上的平均分类正确率 (%)

Table 2 Average classification accuracy(%) on the text and image data set

Data set	Standard learning		Non-kernel learning				Kernel learning			
	SVM	LapSVM	CDSC	SFA	KMM	TCA	DTMKL	GFK	TKL	Our
Orgs vs. people	78.55	82.68	80.97	77.20	80.48	81.58	81.19	81.00	**84.79**	84.71
Orgs vs. place	66.71	68.67	70.62	74.59	68.47	68.15	69.20	76.31	**80.75**	**80.75**
People vs. place	59.94	60.68	64.53	67.08	57.33	57.61	57.80	58.50	67.83	**67.87**
Mean±S.T.	68.40 ±9.42	70.68 ±11.14	72.04 ±8.31	72.96 ±5.25	68.76 ±11.58	69.11 ±12.01	69.40 ±11.70	71.94 ±11.87	**77.79** ±8.86	77.78 ±8.80
C→A	55.64	**56.27**	52.16	49.32	48.32	54.70	54.33	55.95	54.28	54.70
C→W	45.22	45.80	38.54	39.31	45.78	40.76	42.04	42.68	51.19	**52.20**
C→D	43.73	43.73	43.64	41.96	**53.53**	46.44	44.74	48.81	46.50	46.50
A→C	45.77	44.23	42.28	42.33	42.21	45.33	45.01	43.28	45.59	**45.95**
A→W	42.04	42.74	34.94	34.94	42.38	36.31	36.94	42.04	**49.04**	45.76
A→D	39.66	39.79	37.81	36.86	42.72	39.32	40.85	41.36	46.44	**49.68**
W→C	31.43	31.99	32.28	32.50	29.01	33.66	32.50	27.52	34.82	**35.00**
W→A	34.76	34.77	35.73	34.72	31.94	38.00	36.53	34.34	**40.92**	40.71
W→D	82.80	83.43	81.80	83.38	71.98	87.90	**88.85**	79.62	83.44	83.44
D→C	29.39	29.49	33.33	30.50	31.61	33.84	32.10	35.26	35.80	**36.06**
D→A	26.62	27.37	35.88	29.41	32.20	37.79	34.03	37.68	**40.71**	40.61
D→W	63.39	64.31	80.76	68.14	72.88	82.37	81.69	77.29	84.75	**85.08**
Mean±S.T.	45.04 ±15.93	45.32 ±16.06	45.76 ±17.45	43.61 ±16.34	45.38 ±14.70	48.03 ±18.35	47.47 ±18.80	47.15 ±16.29	51.12 ±16.45	**51.31** ±16.50

不变的特征谱, 减小领域特定的特征谱, 比 MMD 适配更有利于探索领域间的多样化结构, 因此具有优良的性能.

图像数据集. 图像迁移学习较文本更具挑战性, 因为图像的简单底层特征与丰富高层语义之间存在 "语义鸿沟" 问题, 因此依赖特征相关性的 SFA 产生负迁移, 分类效果低于 SVM. 此外, 图像迁移任务中, 领域间的概率分布差异通常较大, CDSC 没有显式地最小化领域间的概率分布差异, 也就没有在该任务上取得较好的分类效果. 分布对齐方法 TCA 和 DTMKL 显著优于基准分类器 SVM, 体现了分布对齐对图像迁移学习的重要性. 不过由于 MMD 仅能适配概率分布的各阶统计量, 不能十分有效地对图像数据进行重构. 由表 2 可知, 本文方法获得了图像分类任务的最高平均准确率, 为 51.31%, 相比于次优算法 TKL 提升了 0.19%. 在多分类任务中, 源域和目标领域之间的判别分类面极有可能并不相同, 因此核空间同时对齐边缘分布和条件分布的策略, 在该迁移任务中体现出了相对于 TKL 的优越性.

3.1.3 考虑条件分布适配的有效性验证与参数敏感性分析

考虑条件分布适配的有效性验证. 为验证考虑条件分布适配是否真正提升了分类准确率, 在 18 组分类任务中又进行了仅对齐条件分布 (记为 Conditional)、仅对齐边缘分布 (TKL)、对齐联合分布 (记为 Our) 的对比实验, 表 3 和 4 分别展现了文本和图像数据集上的对比结果. 可见: 对齐条件分布是有效的, Conditional 方法在文本和图像数据集的多个分类任务中达到了最佳的正确率, 甚至在文本数据集

表 3 文本数据集上的联合分布适配有效性验证

Table 3 Validation of joint distribution adaptation on the text data set

Method	Orgs→people	People→orgs	Orgs→place	Place→orgs	People→place	Place→people	Mean±S.T.
TKL	81.46	88.12	78.43	83.07	69.92	**65.74**	77.79± 8.43
Conditional	**81.62**	**88.20**	**78.52**	**83.17**	**70.01**	65.65	**77.86**±8.49
Our	81.46	87.95	**78.52**	82.97	**70.01**	**65.74**	77.78±8.37

表 4 图像数据集上的联合分布适配有效性验证

Table 4 Validation of joint distribution adaptation on the image data set

Method	C→A	C→W	C→D	A→C	A→W	A→D	W→C	W→A	W→D	D→C	D→A	D→W	Mean±S.T.
TKL	54.28	51.19	46.50	45.59	**46.44**	49.04	34.82	**40.92**	**83.44**	35.80	**40.71**	84.75	51.12±16.45
Conditional	53.97	**53.22**	**49.04**	45.86	44.41	47.77	**35.35**	40.19	**83.44**	35.62	40.50	**85.08**	51.20±16.57
Our	**54.70**	52.20	46.50	**45.95**	45.76	**49.68**	35.00	40.71	**83.44**	**36.06**	40.61	**85.08**	**51.31**±16.50

上取得了最佳的平均分类精度. 注意到实验中设置同类标记矩阵中的参数 $0 < b < 1$, 表示 Conditional 并非完全忽略两个域之间不同类数据的结构信息, 只是加强了同类数据的结构适配. 考虑到同类标记矩阵参数的取值不当可能会损害迁移学习的效果, 因此从提升分类准确率的稳定性来看, 同时进行边缘和条件分布适配的本文方法相较于其他两种单独适配的方法会更为有效.

参数敏感性分析. 本文迁移算法主要包含 2 个可调参数: 特征谱阻尼系数 ζ 和 SVM 正则参数 C. 理论上, ζ 的取值可由输入核矩阵的特征谱衰减趋势大致确定; C 表示 SVM 模型对误差的惩罚系数, C 越高, 说明越不能容忍出现误差, 分类模型容易过拟合, C 越小, 容易欠拟合. 但实验表明, 只要 ζ 和 C 在可行的范围内, 它们的取值对本文方法的分类性能仅有轻微影响. 为了证明这些参数的影响, 从合理的离散集 $\zeta \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 2.5, 3, 5\}$, $C \in \{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$ 中取值, 在全部文本和图像数据集上计算所有参数组合的平均分类准确率. 各参数值组合的分类精度如图 1 所示. 可以看出, (1) 对于 ζ, 文本和图像迁移任务中的最佳取值范围分别是 $[1.2, 5.0]$ 和 $[1.1, 1.5]$. 由于图像数据的特征谱衰减速度较慢, 因此 ζ 的最佳取值范围也就相对较窄. (2) 对于 C, 文本数据的最佳取值范围是 $[1, 100]$, 图像数据的最佳取值范围是 $[2, 100]$. 由于在很大的参数取值范围内分类精度大致一致, 故可知本文算法对参数 ζ 和 C 在相当大取值范围内不敏感, 而且在这些取值范围内, 均可以取得超越很多基准方法的分类正确率.

3.2 差分隐私程序分析

施加隐私保护要求必然会降低分类器的性能, 该部分利用文本和图像数据集展示了分类算法的准确性如何随着随机空间维度 D、隐私预算 ε, 以及类别数而变化. 为保护数据的隐私, 该部分高斯核参数 γ 由数据持有方利用源数据直接计算并发送给聚合器.

随机空间维度 D, 隐私预算 ε 影响. 保持 ζ 和 C 的取值同 3.1.2 小节, 设差分隐私 SVM 中的 Huber 常数 $h = 0.5$ [22], $\Lambda = 0.1$. 设置随机空间维度 D 和隐私预算 ε 分别取 $\{500, 1000\}$ 与 $\{0.01, 0.1\}$, 在文本数据集 Reuters-21578 上执行本文完整算法, 并与差分隐私 SVM 直接跨域预测的结果进行对比, 得到实验数据如表 5 所示[6]. 可以观察到: (1) 加入隐私程序后, SVM 的分类精度明显下降, 本文方法的分类精度仅有非常小幅的下降, 而且在每种参数设置下仍保持比 SVM 高至少 20% 的分类精

6) 表中 "(dp)" 表示添加差分隐私保护的算法 (表 6 同). 为了公平地对比, "Our" 部分列出的是使用高斯核计算得到的准确率.

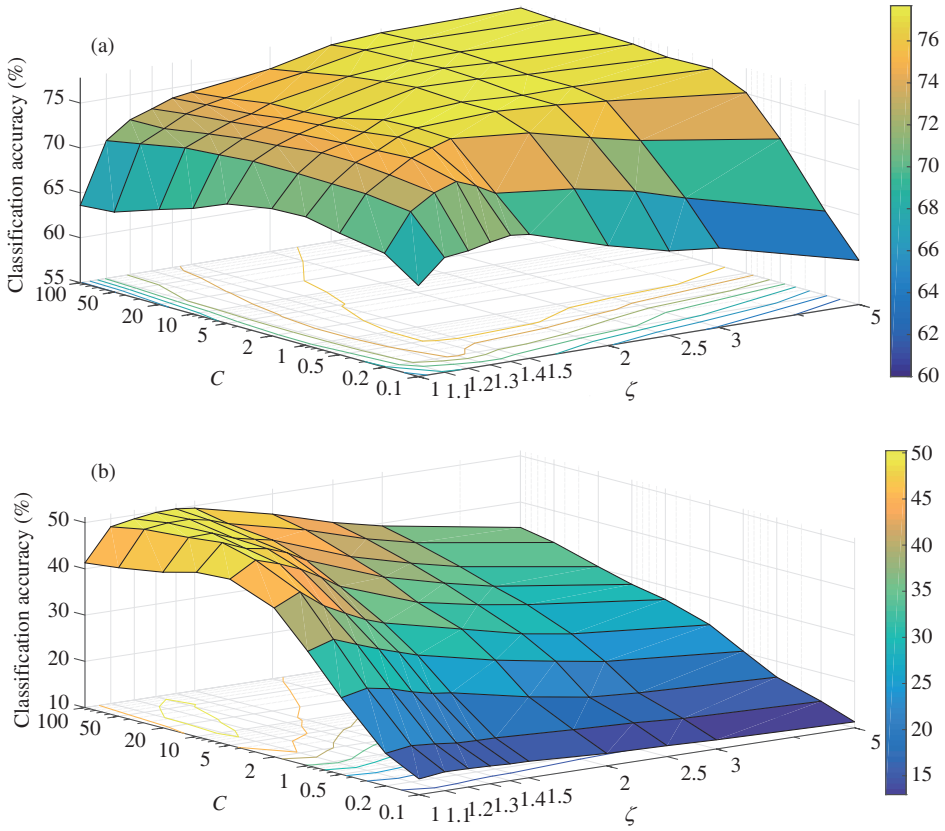

图 1 (网络版彩图) 参数敏感性分析

Figure 1 (Color online) Parameter sensitivity. (a) On Reuters-21578 data set; (b) on office and Caltech-256 data sets

度. (2) 特征空间维度增加, 本文算法的的分类准确率提升, 这是因为 Rahimi-Recht 随机特征算法 [30] 的核近似性能随着投影空间的维数 D 增加而得到了提升. (3) 理论上在满足 ε 差分隐私时, ε 越小, 隐私保护的级别越高, 加入的噪声越多, 实验中 ε 取值为 0.1 时, 差分隐私 SVM 确实在绝大多数任务中取得了较 $\varepsilon = 0.01$ 时更好的实验结果, 但是注意到本文方法的分类精度受到 ε 取值的影响较小, 这允许我们在构建初始源域隐私分类器时设置较高的隐私保护级别.

分类类别数影响. 固定 $D = 500$, $\varepsilon = 0.01$, $\Lambda = 0.01$, 探究在图像数据集的多分类任务中数据类别的多少对目标域分类结果的影响. 选取 C→A (Caltech-256→Amazon) 作为实验数据, 实验结果如表 6 所示. 对于该分类任务, 非隐私版本的本文方法与 SVM 在分类精度上并无明显差别, 但是在添加了隐私程序后, SVM 的分类正确率大大下降, 每个分类类别数对应的降幅均高达 50% 左右, 而本文方法的精度降幅最高不超过 15%, 特别是在低数量类别的分类任务中, 仍能保持与非隐私版本较为接近的正确率.

4 总结

针对跨机构的机器学习服务中存在的数据分布不同的问题, 本文提出了一种基于联合分布核适配的无监督迁移学习方法, 将数据持有方的特征数据和样本标签用于对查询方无标数据的预测. 多个文本和图像数据集上的迁移实验表明, 该方法在较大的参数范围内均能稳定地超越许多基准方法. 此外,

表 5 文本数据集上的随机空间维度与隐私预算影响

Table 5 Impact of random space dimension and privacy budget on the text data set

Data set	Public version		$D = 500$				$D = 1000$			
			$\varepsilon = 0.01$		$\varepsilon = 0.1$		$\varepsilon = 0.01$		$\varepsilon = 0.1$	
	SVM	Our	SVM(dp)	Our(dp)	SVM(dp)	Our(dp)	SVM(dp)	Our(dp)	SVM(dp)	Our(dp)
Orgs→people	75.24	79.97	48.43	79.39	48.18	79.47	46.36	80.96	46.44	80.88
People→orgs	77.12	87.63	47.37	79.14	47.78	79.14	52.71	77.45	53.19	77.36
Orgs→place	70.18	77.95	48.99	67.69	50.05	67.59	46.88	71.91	47.56	71.72
Place→orgs	63.78	83.76	54.23	72.54	53.64	72.54	51.67	82.09	51.87	82.09
People→place	60.63	70.57	44.01	59.42	45.59	59.42	53.39	67.04	52.00	67.04
Place→people	57.94	66.39	48.65	62.02	49.12	62.12	46.80	63.60	47.63	63.51
Mean±S.T.	67.48 ±7.90	77.71 ±7.99	48.62 ±3.30	70.03 ±8.47	49.06 ±2.70	70.05 ±8.48	49.63 ±3.29	73.84 ±7.57	49.78 ±2.89	73.77 ±7.58

表 6 图像迁移任务 C→A 上的分类类别数影响

Table 6 Impact of class number on image transfer task C→A

Method	#Class									Mean±S.T.
	2	3	4	5	6	7	8	9	10	
SVM	96.55	96.27	89.10	77.73	70.90	64.41	60.97	57.79	55.64	74.37±16.26
Our	90.23	92.91	89.37	76.87	72.84	65.32	61.88	59.07	53.97	73.61±14.63
SVM(dp)	55.75	36.57	25.34	20.34	16.40	14.11	12.01	10.70	9.81	22.34±15.14
Our(dp)	93.68	89.55	81.47	67.45	60.49	55.41	49.35	45.23	41.96	64.90±19.39

通过设计让数据源访问随机函数、聚合器发布差分隐私 SVM 分类器, 一定程度上保护了双方数据的隐私安全. 实验结果显示, 即使受非常严格的隐私保护约束, 本文方法也可以取得较高的分类精度.

迁移学习中的数据隐私保护是一个重要问题, 对于机构间的多方合作训练模型具有重要意义. 本文要求源域和目标域数据集中到一个聚合器上进行集中的模型训练, 如何放松这一要求, 实现数据不出本地就能协作训练模型, 并且将双方的学习扩展至多方协作, 仍是有待研究的课题.

参考文献

1 Hu H F, Zheng M, Wu W J, et al. Protein function prediction through multi-instance multi-label transfer learning. Sci Sin Inform, 2017, 47: 1538–1550 [胡海峰, 郑茂, 吴伟坚, 等. 基于多示例多标记迁移学习的蛋白质功能预测. 中国科学: 信息科学, 2017, 47: 1538–1550]

2 Pan S J, Ni X C, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, 2010. 751–760

3 Duan L X, Xu D, Tsang I W H, et al. Visual event recognition in videos by learning from web data. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 1667–1680

4 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359

5 Long M S, Wang J M, Sun J G, et al. Domain invariant transfer kernel learning. IEEE Trans Knowl Data Eng, 2015, 27: 1519–1532

6 Cao X D, Wipf D, Wen F, et al. A practical transfer learning algorithm for face verification. In: Proceedings of IEEE International Conference on Computer Vision, Sydney, 2013. 3208–3215

7 Si S, Tao D C, Geng B. Bregman divergence-based regularization for transfer subspace learning. IEEE Trans Knowl

Data Eng, 2010, 22: 929–942
8 Gretton A, Borgwardt K M, Rasch M, et al. A kernel method for the two-sample problem. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2007. 513–520
9 Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence, Pasadena, 2009. 1187–1192
10 Long M S, Wang J M, Ding G G, et al. Transfer feature learning with joint distribution adaptation. In: Proceedings of IEEE International Conference on Computer Vision, Sydney, 2013. 2200–2207
11 Long M S, Wang J M, Ding G G, et al. Adaptation regularization: a general framework for transfer learning. IEEE Trans Knowl Data Eng, 2014, 26: 1076–1089
12 Hou C A, Tsai Y H H, Yeh Y R, et al. Unsupervised domain adaptation with label and structural consistency. IEEE Trans Image Process, 2016, 25: 5552–5562
13 Zhang J, Li W Q, Ogunbona P. Joint geometrical and statistical alignment for visual domain adaptation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 5150–5158
14 Long M S, Cao Y, Wang J M, et al. Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on Machine Learning, Lile, 2015. 97–105
15 Yan H L, Ding Y K, Li P H, et al. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 945–954
16 Long M S, Zhu H, Wang J M, et al. Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 3470–3479
17 Kamvar S D, Klein D, Manning C D. Spectral learning. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, 2003. 561–566
18 Yin H W, Li F Z. Survey on spectral machine learning. J Comput Sci Tech-Ch, 2015, 9: 1409–1419 [尹宏伟, 李凡长. 谱机器学习研究综述. 计算机科学与探索, 2015, 9: 1409–1419]
19 Azar Y, Fiat A, Karlin A. Spectral analysis of data. In: Proceedings the 33rd Annual ACM Symposium on Theory of Computing, Creta, 2001. 619–626
20 Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Theory of Cryptography Conference, New York, 2006. 265–284
21 Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately? In: Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, Philadelphia, 2008. 531–540
22 Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. J Mach Learn Res, 2011, 12: 1069–1109
23 Wang P Y, Zhang H. Differential privacy for sparse classification learning. Neurocomputing, 2020, 375: 91–101
24 Zhang T, Zhu Q. Dynamic differential privacy for ADMM-based distributed classification learning. IEEE Trans Inform Forensic Secur, 2017, 12: 172–187
25 Zhang X R, Khalili M M, Liu M Y. Improving the privacy and accuracy of ADMM-based distributed algorithms. In: Proceedings of the 35th International Conference on Machine Learning, Stockholm, 2018. 9221–9230
26 Wang Y, Gu Q Q, Brown D. Differentially private hypothesis transfer learning. In: Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Dublin, 2018. 811–826
27 Xie L Y, Baytas I M, Lin K X, et al. Privacy-preserving distributed multi-task learning with asynchronous updates. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, 2017. 1195–1204
28 Sarpatwar K, Shanmugam K, Ganapavarapu V S, et al. Differentially private distributed data summarization under covariate shift. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 14432–14442
29 Wang J D, Feng W J, Chen Y Q, et al. Visual domain adaptation with manifold embedded distribution alignment. In: Proceedings of the 26th ACM International Conference on Multimedia, Seoul, 2018. 402–410
30 Rahimi A, Recht B. Random features for large-scale kernel machines. In: Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, 2007. 1177–1184
31 Williams C K I, Seeger M. Using the Nyström method to speed up kernel machines. In: Proceedings of Advances in Neural Information Processing Systems, Denver, 2000. 682–688

32 Zhang T, Ando R K. Analysis of spectral kernel design based semi-supervised learning. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2005. 1601–1608
33 Jin R, Yang T B, Mahdavi M, et al. Improved bounds for the Nyström method with application to kernel classification. IEEE Trans Inform Theory, 2013, 59: 6939–6949
34 Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res, 2008, 9: 1871–1874
35 Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 283–291
36 Gong B Q, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 2066–2073
37 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res, 2006, 7: 2399–2434
38 Ling X, Dai W Y, Xue G R, et al. Spectral domain-transfer learning. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 488–496
39 Chen B, Lam W, Tsang I, et al. Extracting discriminative concepts for domain adaptation in text mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 179–187
40 Huang J Y, Smola A J, Gretton A, et al. Correcting sample selection bias by unlabeled data. In: Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, 2006. 601–608
41 Duan L X, Tsang I W, Xu D. Domain transfer multiple kernel learning. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 465–479

Transfer learning based on joint distribution kernel adaptation and its privacy protection

Xuanming NI[1], Xinyuan SHEN[1] & Hai ZHANG[2,3*]

1. *School of Software and Microelectronics, Peking University, Beijing* 100871*, China*;
2. *School of Mathematics, Northwest University, Xi'an* 710127*, China*;
3. *Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, East China Normal University, Shanghai* 200062*, China*
* Corresponding author. E-mail: zhanghai@nwu.edu.cn

Abstract Transfer learning solves a learning problem in a target domain by utilizing the labeled data in a different but related source domain. Most prior methods to reduce the distribution difference between domains depend on the maximum mean discrepancy (MMD) distance, but MMD can only match the moments of data distributions between domains. In addition, the increasing privacy protection awareness restricts access to data sources and poses new challenges for the development of transfer learning. This paper proposes a transfer learning method based on joint distribution kernel adaptation and puts forward its privacy protection approach. We learn a domain-invariant kernel by directly matching both the marginal distribution and conditional distribution in the reproducing kernel Hilbert space. Besides, both data sources are designed to access the same random projection function firstly, then the aggregator is set to release a differential privacy kernel classifier based on objective perturbation. While implementing kernel-based joint distribution adaptation, it avoids the data source and the aggregator from directly sharing the original feature data. Comparative experiments and parameter analysis on multiple text and image transfer benchmark data sets verify the effectiveness of the proposed method.

Keywords transfer learning, privacy protection, distribution adaptation, spectral learning, differential privacy

Xuanming NI was born in 1984. He received his Ph.D. degree in quantitative economics from Tsinghua University, Beijing, in 2015. Currently, he is an M.S. supervisor at the School of Software and Microelectronics at Peking University, Beijing. His research interests include Fintech, quantitative economics, and population economics.

Hai ZHANG was born in 1975. He received his Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, in 2012. He is currently a professor at Northwest University. His research interests include statistical machine learning, high-dimensional statistics, and social network analysis.

Xinyuan SHEN was born in 1996. Currently, she is an M.S. candidate in computer technology at Peking University, Beijing. Her research interests include distributed learning and Fintech.