文章编号:2095-6134(2015)01-0121-06

基于平移不变核的异构迁移学习*

关增达[1], 程 立[2,3], 朱廷劭[4†]

(1 中国科学院大学计算机与控制学院, 北京 101408; 2 新加坡科技研究局生物信息研究所, 新加坡 138671;
3 新加坡国立大学计算机学院, 新加坡 119077; 4 中国科学院心理研究所, 北京 100101)

(2013 年 12 月 25 日收稿; 2014 年 3 月 19 日收修改稿)

Guan Z D, Cheng L, Zhu T S. Heterogeneous transfer learning based on translation invariant kernels[J]. Journal of University of Chinese Academy of Sciences, 2015,32(1):121-126.

摘 要 提出一种新的异构迁移学习方法.利用与目标数据集相关的异构特征数据集.通过把目标集和异构集的数据使用平移不变核(欧式距离核和径向基函数核),映射到一个新的再生核希尔伯特空间上.在新空间中 2 个数据集的特征相同,特征维度相等,分布接近,且保持数据的拓扑性质不变.实验证明,该方法特别是基于欧式距离核的方法取得了较好的效果,在目标训练集的标注数据较少时,有大于 5% 甚至超过 10% 的精度提高.

关键词 异构迁移学习; 平移不变核; RKHS

中图分类号:TP301 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2015.01.020

Heterogeneous transfer learning based on translation invariant kernels

GUAN Zengda[1], CHENG Li[2,3], ZHU Tingshao[4]

(1 *School of Computer and Control, University of Chinese Academy of Sciences, Beijing* 101408, *China*;
2 *Bioinformatics Institute, A*STAR, Singapore* 138671; 3 *School of Computing, National University of Singapore, Singapore* 119077; 4 *Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract We propose a new heterogeneous transfer learning method, which uses related heterogeneous feature dataset. We use translation invariant kernels (Euclidean kernels and RBF kernels) to map the target dataset and the related dataset to a new reproducing kernel Hilbert space, in which the two datasets have equal feature dimensions and similar distributions and reserve their topological property. The experimental results show that our method works well and the method based on the Euclidean kernel improves accuracy by more than 5% ~10%.

Key words heterogeneous transfer learning; translation invariant kernel; RKHS

当目标数据集(本文中简称目标集)没有足够的标注时,监督学习的效果会比较差,而利用大量与目标集具有相同标注类别的其他来源的相关数据集(本文中简称源集),通过迁移学习,常常

*国家重点基础研究发展(973)计划(2014CB744600)、中国科学院重点项目(KJZD-EW-L04)、中国科学院战略性先导科技专项(XDA06030800)资助

†通信作者,E-mail: tszhu@psych.ac.cn

能帮助目标集训练出更好的模型. 近些年来,迁移学习逐步发展起来,并且在文本处理、图象处理等许多领域获得了实际应用[1].

在迁移学习中,有一类情况,当目标集与源集数据的特征不同且特征的维度也不相等时,目标集与源集的特征不满足一一对应的条件,传统的迁移学习方法就无法发挥作用. 这样的问题被称为异构迁移学习. 例如,要对一些西班牙语文本分类,但是只有很少的标注数据,因此难以训练出好的监督学习模型,而同时我们有大量的英语标注文本,尽管西班牙语与英语文本的特征(文档关键词)不同,但仍可以考虑迁移这些英语文本上的预测模型知识用于西班牙语文本分类,帮助得出一个更好的机器学习模型来. 不要求目标集与源集数据特征的维度数相同,特征一一对应,使异构迁移学习具有远比一般迁移学习更加广阔的应用空间.

异构迁移学习研究中,一般通过发现目标集与源集数据间的相关性,以及它们预测函数间的相关性,把原来的异构数据映射为同维的中间数据,来实现源集和目标集之间的迁移. 目前的异构迁移学习研究还不够成熟,已有的研究中,文献[2]提出一种扩展的概率隐语义分析模型(probabilistic latent semantic analysis),利用异构数据集帮助改善图像聚类的效果. 文献[3]提出使用集体矩阵因子分解(collective matrix factorization)技术,通过寻找辅助的异构数据集与目标数据集的共同语义表示改进图像分类. 文献[4]的方法通过一种语言模型来关联和"翻译"源集和目标集的特征集,以改善跨语言文本分类等. 文献[5]对多个数据集使用面向标签的流形对齐方法,结合现有的迁移学习方法改善目标集上的学习效果,但是这种两阶段的迁移学习过程增加了误差产生的可能. 文献[6]提出一种称作 heterogeneous feature augmentation 的方法,通过对源集和目标集数据进行异构变换得到二者的共同特征,然后新数据与原数据联合训练得到好的监督学习模型. 文献[7]使用一种非对称非线性的方法,将相关数据集往目标集所在的空间上迁移,并对数据集之间语义相似与不相似的情况分别做处理. 不过,文献[6]和文献[7]中的方法没有考虑在迁移变换过程中保持原数据某些有用的性质,比如拓扑性质等. 文献[8]则利用线性核作异构变换,并在保持原始数据拓扑性质的条件下做异构迁移学习,但是该方法难以满足许多情况下非线性变换的要求. 而在传统的迁移学习方法中,文献[9]和文献[10]则分别考虑了在迁移过程中保持原数据的拓扑性和方差约束,但是他们的方法不能直接用于异构迁移学习.

针对上面方法的不足,本文提出一种基于平移不变核(包括欧式距离核和 RBF 核)的异构迁移学习算法,以非线性的方式将源集和目标集的原数据,映射到一个分布接近的新特征空间,并在过程中保持数据的相对距离和拓扑性质,以便在目标集标注量较少时实现更好的模型预测效果.

1 模型

首先定义本文使用的各种数学符号. 源集和目标集分别记作 D_s 和 D_t. $(x_{si},y_{si})|_{i=1}^{n_s}$ 定义为源集的数据及其标注,x_{si} 和 y_{si} 分别是源集中数据样本及对应标注,n_s 为源数据集的样本数,d_s 是源集的特征维度数,y_{si} 的取值范围是 $\{-1,1\}$. 通过类似的方式定义目标数据集的 $(x_{ti},y_{ti})|_{i=1}^{n_t}$,$d_t$,以及 y_{ti} 的取值范围 $\{-1,1\}$ 等. 值得注意的是,由于是异构迁移学习问题,这里 $d_s \neq d_t$.

本文采用 Borgwardt 等提出的最大平均差异(maximum mean discrepancy, MMD)[11] 度量源集与目标集分布之间的距离. 该方法主要通过把原始数据映射到一个表达力足够丰富(universal)的数学空间——压缩核希尔伯特空间(reduced kernel Hilbert space, RKHS)中,在这个新空间里,如果 MMD 值足够小,则可以认为这 2 个数据集属于相同分布. 在本文的方法中,用 $X_{si}\boldsymbol{W}_s$ 和 $X_{ti}\boldsymbol{W}_t$ 把源集和目标集数据转变为相同维度,然后映射到一个 RKHS 空间上. $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 称作特征变换矩阵. 源集和目标集的 MMD 的经验估计方法如下:

$$\mathrm{Dist}(X_s,X_t) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\boldsymbol{\Phi}(X_{si}\boldsymbol{W}_s) - \frac{1}{n_t}\sum_{i=1}^{n_t}\boldsymbol{\Phi}(X_{ti}\boldsymbol{W}_t)\right\|_H^2. \quad (1)$$

(1)式中,$\boldsymbol{\Phi}(X_{si}\boldsymbol{W}_s)$ 和 $\boldsymbol{\Phi}(X_{ti}\boldsymbol{W}_t)$ 表示 X_{si} 和 X_{ti} 的转换函数,$\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 把不同维度的源集和目标集数据变换为相同的维度,然后 $\boldsymbol{\Phi}(\,)$ 将源集和目标集的数据映射到新的 RKHS 特征空间 H 中.

采用文献[9]中的化简方式,可以推导得到 $\min \mathrm{Dist}(X_s,X_t) = \min \mathrm{tr}(\boldsymbol{KL})$,

其中, $\boldsymbol{K}=\begin{bmatrix}\boldsymbol{K}_{ss} & \boldsymbol{K}_{st}\\ \boldsymbol{K}_{ts} & \boldsymbol{K}_{tt}\end{bmatrix}$, $\boldsymbol{L}=\begin{bmatrix}\boldsymbol{L}_{ss} & \boldsymbol{L}_{st}\\ \boldsymbol{L}_{ts} & \boldsymbol{L}_{tt}\end{bmatrix}$,

$$\boldsymbol{K}_{ij}=<\boldsymbol{\Phi}(x_i\boldsymbol{W}_{s/t}),\boldsymbol{\Phi}(x_j\boldsymbol{W}_{s/t})>. \tag{2}$$

(2)式中,$\boldsymbol{K}$ 是一个 $(n_s+n_t)\times(n_s+n_t)$ 矩阵,$\boldsymbol{K}_{ss}$、$\boldsymbol{K}_{st}$、$\boldsymbol{K}_{ts}$ 以及 $\boldsymbol{K}_{tt}$ 分别是源集之间、源集与目标集、目标集与源集以及目标集之间的 kernel Gram 矩阵,其中的元素 $\boldsymbol{K}_{ij}:=k(x_i,x_j)$. $\boldsymbol{L}$ 同样也是一个 $(n_s+n_t)\times(n_s+n_t)$ 矩阵,$\boldsymbol{L}_{ss}=1/n_s^2$,$\boldsymbol{L}_{st}=-1/(n_s\times n_t)$,$\boldsymbol{L}_{ts}=-1/(n_s\times n_t)$,$\boldsymbol{L}_{tt}=1/n_t^2$. $\mathrm{tr}(\boldsymbol{KL})$表示 $\boldsymbol{K}$ 和 $\boldsymbol{L}$ 的乘积矩阵的迹. 通过求解(2)式可以得到 $\boldsymbol{W}_s$和 $\boldsymbol{W}_t$.

这里选用平移不变核函数,包括欧氏距离核和径向基函数核(radial basis function kernel,简称 RBF 核). 这些平移不变核相比线性核,能实现非线性映射,且具有性质 $k(x_i,x_j)=k(x_i-x,x_j-x)$,这对于在异构转换过程中保持数据点的相对距离不变是重要的. 而 RBF 核作为支持向量机中常用的核函数,更具有优良的函数表达能力,当给出的数据缺少先验知识时,RBF 核往往是一种较好的选择,能够给出平滑的估计. 其表达式是 $k(x_i,x_j)=\exp(-\gamma\|x_i\boldsymbol{W}_{s/t}-x_j\boldsymbol{W}_{s/t}\|^2)$. 对于欧氏距离核,$k(x_i,x_j)=(x_i\boldsymbol{W}_{s/t}-x_j\boldsymbol{W}_{s/t})^2$,则具有计算简单的优点,也具有一定的适应性. 当然,这里的 2 次方项,根据实际情况也可设为其他幂指数.

为了在转换过程中使源集和目标集中实例的拓扑性质尽量保持不变,以免转换后源集和目标集中实例的相邻关系变化过大,我们使用文献[9]中的 Graph Regulation,如下:

$$\begin{aligned}\Omega(\phi)&=\frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{U}_{ij}\left\|\frac{\boldsymbol{\Phi}(x_i\boldsymbol{W}_{s/t})}{\sqrt{D_{ii}}}-\frac{\boldsymbol{\Phi}(x_j\boldsymbol{W}_{s/t})}{\sqrt{D_{jj}}}\right\|_2^2\\&=\mathrm{tr}(\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{\Phi}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})=\mathrm{tr}(\boldsymbol{KG}),\end{aligned} \tag{3}$$

(3)式中,$\boldsymbol{U}$ 表示邻接矩阵,$\boldsymbol{U}_{ij}$ 表示 2 个数据点 X_i 和 X_j 间的相似度量,$\boldsymbol{D}$ 是对角矩阵,并且 $\boldsymbol{D}=\sum_j U_{ij}$,$\boldsymbol{G}=\boldsymbol{D}-\boldsymbol{U}$ 是拉普拉斯矩阵.

有效地利用数据集的标注信息,一般能较大地提高监督学习的效果,这里我们以与文献[8]相同的方式,最小化同标注数据之间的 MMD,最大化不同标注数据之间的 MMD,同时考虑图正规化因素,用下面的函数表示:

$$\begin{aligned}f(\boldsymbol{W}_s,\boldsymbol{W}_t)&=\lambda_1\sum_i^{N_=}\mathrm{tr}(\boldsymbol{K}_=\boldsymbol{L}_=)+\\&\lambda_2\sum_j^{N_\neq}\mathrm{tr}(\boldsymbol{K}_\neq\boldsymbol{L}_\neq)+\lambda_3\sum_k^{N}\mathrm{tr}(\boldsymbol{KG});\end{aligned}$$

同时,约束 $\boldsymbol{W}_s$和 $\boldsymbol{W}_t$的取值范围,得到下面的优化公式:

$$\arg\min_{\boldsymbol{W}_s,\boldsymbol{W}_t}f(\boldsymbol{W}_s,\boldsymbol{W}_t),$$

其中,$\|\boldsymbol{W}_s\|<m_1$,$\|\boldsymbol{W}_t\|<m_2$. (4)

在上两式中,"=" 表示 2 个实例的标注相同的情况,"≠" 表示 2 个实例的标注不相同的情况,$N_=$表示 2 个实例的标注相同情况的总数量,$N_\neq$表示 2 个实例的标注不相同情况的总数量,N 在这里表示不区分标注相同与不同情况时的数量(即 $N=1$),当区分标注相同与不同时则 $N=N_=+N_\neq$,$\boldsymbol{KG}$ 也做相应变化;同时以范数的形式约束 $\boldsymbol{W}_s$和 $\boldsymbol{W}_t$的取值范围. 公式中通过 λ_1、λ_2 以及 λ_3 可以设置各部分的权重,根据具体的核函数设置符号的正负,特别注意的,λ_2 和 λ_1 符号相反. 与文献[8]不同的是,这里的核函数采用了平移不变核.

由于对(4)式难以计算全局最优点,我们使用梯度下降方法寻找满足条件的局部最优点.

2　求解与算法

使用梯度下降方法求解. 通过对(4)式求解 $\boldsymbol{W}_t$ 和 $\boldsymbol{W}_t$ 的偏导数,得到 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 的更新公式:

$$\boldsymbol{W}_s^{(r+1)}=\boldsymbol{W}_s^{(r)}-\eta\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial\boldsymbol{W}_s}, \tag{5}$$

$$\boldsymbol{W}_t^{(r+1)}=\boldsymbol{W}_t^{(r)}-\eta\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial\boldsymbol{W}_t}. \tag{6}$$

上式中,(r)和$(r+1)$表示第 r 次和第 $r+1$ 次迭代,η 是计算 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 时梯度下降的步长.

对于欧式距离核,

$$\begin{aligned}&\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial\boldsymbol{W}_s}\\&=\sum_{k=1}^{3}\lambda_k\sum_{l=1}^{N_k}[2\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{ks}}\lg_{1kij}(\boldsymbol{x}_{ksi}-\boldsymbol{x}_{ksj})^{\mathrm{T}}(\boldsymbol{x}_{ksi}-\boldsymbol{x}_{ksj})\boldsymbol{W}_s^{(r)}+\\&4\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{kt}}\lg_{2kij}(\boldsymbol{x}_{ksi}^{\mathrm{T}}\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)}-\boldsymbol{x}_{ksi}^{\mathrm{T}}\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)})].\end{aligned} \tag{7}$$

其中,$k=1$ 表示 2 个实例标注相同的情况,即同一种标注的全部实例,$k=2$ 表示 2 个实例标注不相同的情况,$k=3$ 表示 Graph Regulation 的情况. N_k 在 $k=1,2,3$ 时分别表示 $N_=$,$N_\neq$ 以及 N,在 N_k 右侧的中括号内的 n,lg 和 x 都对应各自的情况 l,因为式中符号过多,在不影响理解的情况下省略

下标 l,下同. 对应的,n_{ks}(即 n_{kls},下同)分别表示 $k=1,2,3$ 时 l 情况的源集实例数量,n_{kt} 分别表示 $k=1,2,3$ 时 l 情况的目标集实例数量. 当 $k=1$ 或 2 时,$\lg_{1kij}$ 表示 $l/(n_{ks}\times n_{ks})$,$\lg_{2kij}$ 表示 $-1/(n_{ks}\times n_{kt})$,当 $k=3$ 时,$\lg_{1kij}=G_{ij}$,$\lg_{2kij}=G_{i(n_s+j)}=0$. x_{ksi},x_{kti} 等作类似看待.

$$\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial \boldsymbol{W}_t}=\sum_{k=1}^{3}\lambda_k\sum_{l=1}^{N_k}\Big[2\sum_{i=1}^{n_{kt}}\sum_{j=1}^{n_{kt}}\lg_{1kij}(\boldsymbol{x}_{kti}-\boldsymbol{x}_{ktj})^{\mathrm{T}}(\boldsymbol{x}_{kti}-\boldsymbol{x}_{ktj})\boldsymbol{W}_t^{(r)}+4\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{kt}}\lg_{2kij}(\boldsymbol{x}_{ktj}^{\mathrm{T}}\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)}-\boldsymbol{x}_{ktj}^{\mathrm{T}}\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)})\Big]. \tag{8}$$

其中,当 $k=1$ 或 2 时,$\lg_{1kij}$ 表示 $1/(n_{kt}\times n_{kt})$,$\lg_{2kij}$ 表示 $-1/(n_{ks}\times n_{kt})$,当 $k=3$ 时,$\lg_{1kij}=G_{(n_s+i)(n_s+j)}$,$\lg_{2kij}=G_{(n_s+i)j}=0$. 其他情况与式(7)相同.

对于 RBF 核,

$$\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial \boldsymbol{W}_s}=\sum_{k=1}^{3}\lambda_k\sum_{l=1}^{N_k}\Big[-2\gamma\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{ks}}\lg_{1kij}\exp(-\gamma\|(\boldsymbol{x}_{ksi}-\boldsymbol{x}_{ksj})\boldsymbol{W}_s^{(r)}\|_2^2)(\boldsymbol{x}_{ksi}-\boldsymbol{x}_{ksj})^{\mathrm{T}}(\boldsymbol{x}_{ksi}-\boldsymbol{x}_{ksj})\boldsymbol{W}_s^{(r)}-4\gamma\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{kt}}\lg_{2kij}\exp(-\gamma\|\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)}-\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)}\|_2^2)(\boldsymbol{x}_{ksi}^{\mathrm{T}}\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)}-\boldsymbol{x}_{ksi}^{\mathrm{T}}\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)})\Big]. \tag{9}$$

$$\frac{\partial f(\boldsymbol{W}_s,\boldsymbol{W}_t)}{\partial \boldsymbol{W}_t}=\sum_{k=1}^{3}\lambda_k\sum_{l=1}^{N_k}\Big[-2\gamma\sum_{i=1}^{n_{kt}}\sum_{j=1}^{n_{kt}}\lg_{1kij}\exp(-\gamma\|(\boldsymbol{x}_{kti}-\boldsymbol{x}_{ktj})\boldsymbol{W}_t^{(r)}\|_2^2)(\boldsymbol{x}_{kti}-\boldsymbol{x}_{ktj})^{\mathrm{T}}(\boldsymbol{x}_{kti}-\boldsymbol{x}_{ktj})\boldsymbol{W}_t^{(r)}-4\gamma\sum_{i=1}^{n_{ks}}\sum_{j=1}^{n_{kt}}\lg_{2kij}\exp(-\gamma\|\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)}-\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)}\|_2^2)(\boldsymbol{x}_{ktj}^{\mathrm{T}}\boldsymbol{x}_{ktj}\boldsymbol{W}_t^{(r)}-\boldsymbol{x}_{ktj}^{\mathrm{T}}\boldsymbol{x}_{ksi}\boldsymbol{W}_s^{(r)})\Big]. \tag{10}$$

上面(9)、(10)式里面的各项符号定义与式(7)、(8)中相同.

利用上面的结果,得到如下算法:

算法　基于平移不变核的异构迁移算法

输入　源集数据 X_s 及其标注 Y_s,目标集数据 X_t 及其标注 Y_t,待预测数据集 X_p,$\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 分别初始化为(0,1)范围内取值的矩阵,根据实际具体情况设置参数 $\lambda_1,\lambda_2,\lambda_3,m_1,m_2,r_{\max}$ 以及 τ 的值.

输出　数据集 X_p 的预测值 Y_p.

1)根据 k-最近邻算法计算每个数据点最相邻的 k 个邻接点,得到邻接矩阵,然后根据公式(3),计算相关的 Laplacian 矩阵 $\boldsymbol{G}$;

2)通过公式(4)—(10),利用梯度下降方法计算 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$,r 初始值为 1:

while $r<r_{\max}$.

根据公式(5)、(6)以及(7)、(8)(或(9)、(10)),通过梯度下降算法更新 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$,并且保证 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 的 Frobenius-范数分别小于 m_1 和 m_2;

利用新的 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 计算公式(4)的新结果;

If 公式(4)的新结果与它上一次结果的差绝对值小于 τ,

Then 认为结果已经收敛,退出循环;

Else $r\leftarrow r+1$;

end

3)利用上一步计算得到的 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 计算新的数据集 $X_s\boldsymbol{W}_s$ 和 $X_t\boldsymbol{W}_t$,并且把它们放在一起训练,得到一个监督学习的模型,然后预测 X_p 的值 Y_p,计算精度.

3　实验与分析

实验中我们比较了无迁移、线性核异构迁移、欧式距离核,以及 RBF 核异构迁移 4 种情况,以及欧式距离核与 RBF 核异构迁移中参数设置和样本数量对预测结果精度的影响.

我们的实验选择了一个文本分类的公开数据集 Reuters multilingual dataset①,通过对 Reuters RCV1 和 RCV2 数据集使用部分取样获得. 我们分别随机选取 200 篇英语和西班牙语的文档,都分为 2 个类别. 然后用标注了类别的英语文档和部分的西班牙语文档做训练,剩余的西班牙语文档做测试,即英语文档作为源集,西班牙语文档作为目标集,用于训练的西班牙语文档分别选 10,14,20,30 与 40 篇,其中 2 个类别的文档各占 50%. 因为英语文档与西班

① http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm

牙语文档都含有非常多的关键词,所以先用 PCA 技术降维,选取其中最相关的特征. 降维后英语文档与西班牙语文档分别是 218 维和 186 维,使算法的时间和空间复杂度能满足在一台普通工作站上训练和学习. 另外,对于方法中的图邻接矩阵,相似度计算采用余弦相似度,k 最近邻中的近邻点则取所有点.

我们使用基线方法、非迁移方法 non-transfer)和线性核异构迁移方法(linear kernel)[8],与本文欧式距离核方法(Euclidean kernel)和 RBF 方法(RBF kernel)进行比较(表 1). 对于非迁移方法以及各种方法异构转换后数据的处理,我们使用数据挖掘软件 Weka① 做监督学习的训练和测试. 经过实验,选择表现较好的基于多项式核的 SVM 分类方法. 具体过程是在英文文档和一部分标注的西班牙语文档上训练,并执行 10 - fold 交叉检验,然后利用剩余的西班牙语文档测试. 每一个数据都通过 10 次实验取均值.

表 1　异构迁移方法的比较

Table 1　Comparison among heterogeneous transfer learning methods　　%

	10	14	20	30	40
non-transfer	54.84	65.38	67.00	73.06	74.00
linear kernel	54.89	69.95	69.89	75.53	77.69
Euclidean kernel	67.05	70.86	74.33	76.06	77.56
RBF kernel	64.95	68.44	72.50	72.82	71.56

在表 1 中,顶栏中的数值表示目标集也就是西班牙语文档中用于训练的标注文本数量,左栏的"不迁移"方法表示直接使用标注文本进行训练,而不进行异构迁移学习. 从表 1 中,可以看到,欧式距离核的方法各种情况下都表现最优,或与最优水平相差很小. 同时可以看到,使用异构迁移学习的方法比起不使用的情况,在目标集训练样本不太多的情况下(10,14 以及 20 个训练样本),都有精度的提高. 而随着样本数量的增多(30、40 个样本时),线性核方法和欧氏距离核方法仍然表现有精度提高,而 RBF 核方法则不如非迁移方法. 这可能是因为 RBF 核方法在该批数据上适应性不好. 而欧式距离核的方法的健壮性则较好,但是其精度提高的幅度也在收窄. 总体上,表现最好的方法是欧式距离核方法. 在各种异构迁移学习算法的计算时间复杂度方面,线性核方法最快,欧式距离核方法其次,RBF 核方法最慢. 这主要是因为,用来做异构迁移的核函数在计算中,RBF 核涉及指数运算,时间复杂度最高,欧氏距离核属于幂运算,时间复杂度次之,而线性核主要是内积运算,时间复杂度最低.

对于欧式距离核和 RBF 核的方法,它们的原始数据在异构转换后的新特征的维度对预测结果的精度有较大影响. 我们分别测试了在新特征维度是 50、100、150、200 和 250 的情况. 这里使用的样本数量都是 20. 各个结果仍然是测量 10 次取平均值,结果如图 1 所示.

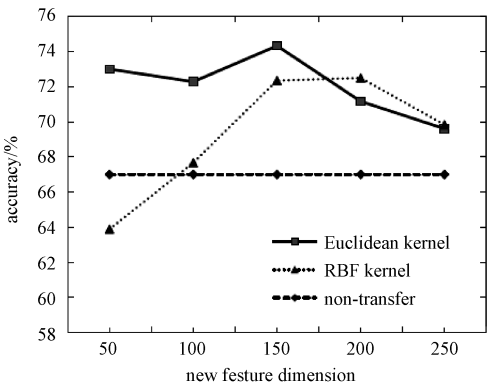

图 1　不同的新特征维度对异构迁移方法的影响

Fig. 1　Influence of different new feature dimensions on heterogeneous transfer learning

可以看到,在新特征维度过小和过大时,都没有取得最优的预测精度,甚至可能表现得不如非迁移时的情况. 异构转换后的新特征维度只有在取合适值时,才会取得最佳的预测精度.

4　结论

本文提出一种新的基于平移不变核函数的异构迁移学习方法,并分别实现了利用欧式距离核和 RBF 核的异构迁移学习算法. 通过将源集和目标集的异构数据映射到一个新特征空间 RKHS 中,同时尽量保持原数据间的拓扑相邻性质,最后通过使用梯度下降方法解优化问题得到特征变换

① Weka3.7,来自 http://www.cs.waikato.ac.nz/ml/weka/

矩阵. 实验结果证明了该方法的有效性和优越性. 未来,我们将考察更多的核函数对异构变换的作用,并研究异构迁移学习中的回归问题.

附录

公式(9)和(10)的推导如下:

由公式(4),

$$\arg\min_{\boldsymbol{W}_s,\boldsymbol{W}_t}\lambda_1\sum_i^{N_=}\mathrm{tr}(\boldsymbol{K}_=\boldsymbol{L}_=)+\lambda_2\sum_j^{N_{\neq}}\mathrm{tr}(\boldsymbol{K}_{\neq}\boldsymbol{L}_{\neq})+\lambda_3\sum_k^{N}\mathrm{tr}(\boldsymbol{KG}),$$

取公式中第 1 项,分别对 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 求偏导,其中,

$$\frac{\partial\mathrm{tr}(\boldsymbol{K}_=\boldsymbol{L}_=)}{\partial\boldsymbol{W}_s}=\frac{\partial(\boldsymbol{K}_{=ss}\boldsymbol{L}_{=ss}+\boldsymbol{K}_{=st}\boldsymbol{L}_{=ts}+\boldsymbol{K}_{=ts}\boldsymbol{L}_{=st}+\boldsymbol{K}_{=tt}\boldsymbol{L}_{=tt})}{\partial\boldsymbol{W}_s},$$

由于平移不变核, $\boldsymbol{K}_{=st}=\boldsymbol{K}_{=ts}$, 又 $\boldsymbol{L}_{=ts}=\boldsymbol{L}_{=st}$, 而且 $\boldsymbol{K}_{=tt}\boldsymbol{L}_{=tt}$ 与 $\boldsymbol{W}_s$ 无关,所以上式变为

$$\frac{\partial\mathrm{tr}(\boldsymbol{K}_=\boldsymbol{L}_=)}{\partial\boldsymbol{W}_s}=\frac{\partial(\boldsymbol{K}_{=ss}\boldsymbol{L}_{=ss}+2\boldsymbol{K}_{=st}\boldsymbol{L}_{=ts})}{\partial\boldsymbol{W}_s}$$

$$=-\frac{2\gamma}{n_{=s}^2}\sum_{i,j=1}^{n=s}\mathrm{e}^{-\gamma\|(\boldsymbol{x}_{=si}-\boldsymbol{x}_{=sj})\boldsymbol{W}_s\|_2^2}(\boldsymbol{x}_{=si}-\boldsymbol{x}_{=sj})^{\mathrm{T}}(\boldsymbol{x}_{=si}-\boldsymbol{x}_{=sj})\boldsymbol{W}_s+\frac{4\gamma}{n_{=s}n_{=t}}\sum_{i=1}^{n=s}\sum_{j=1}^{n=t}\mathrm{e}^{-\gamma\|\mathrm{x}_{=si}\boldsymbol{W}_s-\mathrm{x}_{=tj}\boldsymbol{W}_t\|_2^2}((\boldsymbol{x}_{=si})^{\mathrm{T}}\boldsymbol{x}_{=st}\boldsymbol{W}_s-(\boldsymbol{x}_{=si})^{\mathrm{T}}\boldsymbol{x}_{\mathrm{tj}}\boldsymbol{W}_t);$$

同理,得到

$$\frac{\partial\mathrm{tr}(\boldsymbol{K}_=\boldsymbol{L}_=)}{\partial\boldsymbol{W}_t}$$

$$=-\frac{2\gamma}{n_{=t}^2}\sum_{i,j=1}^{n=t}\mathrm{e}^{-\gamma\|(\boldsymbol{x}_{=ti}-\boldsymbol{x}_{=tj})\boldsymbol{W}_t\|_2^2}(\boldsymbol{x}_{=ti}-\boldsymbol{x}_{=tj})^{\mathrm{T}}(\boldsymbol{x}_{=ti}-\boldsymbol{x}_{=tj})\boldsymbol{W}_t+\frac{4\gamma}{n_{=s}n_{=t}}\sum_{i=1}^{n=s}\sum_{j=1}^{n=t}\mathrm{e}^{-\gamma\|\boldsymbol{x}_{=si}\boldsymbol{W}_s-\boldsymbol{x}_{=tj}\boldsymbol{W}_t\|_2^2}((\boldsymbol{x}_{=tj})^{\mathrm{T}}\boldsymbol{x}_{=tj}\boldsymbol{W}_t-(\boldsymbol{x}_{=tj})^{\mathrm{T}}\boldsymbol{x}_{=si}\boldsymbol{W}_s).$$

对于公式中的第 2 项,可以类似处理.

对于公式中的第 3 项,类似处理,得

$$\frac{\partial\mathrm{tr}(\boldsymbol{KG})}{\partial\boldsymbol{W}_s}=\frac{\partial\sum_{i,j=1}^{n}\boldsymbol{K}_{ij}\boldsymbol{G}_{ij}}{\partial\boldsymbol{W}_s}$$

$$=-2\gamma\sum_{i,j=1}^{n_s}\boldsymbol{G}_{ij}\mathrm{e}^{-\gamma\|(\boldsymbol{x}_{si}-\boldsymbol{x}_{sj})\boldsymbol{W}_s\|_2^2}(\boldsymbol{x}_{si}-\boldsymbol{x}_{sj})^{\mathrm{T}}(\boldsymbol{x}_{si}-\boldsymbol{x}_{sj})\boldsymbol{W}_s+4\gamma\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\boldsymbol{G}_{i(n_s+j)}\mathrm{e}^{-\gamma\|\boldsymbol{x}_{si}\boldsymbol{W}_s-\boldsymbol{x}_{tj}\boldsymbol{W}_t\|_2^2}((\boldsymbol{x}_{si})^{\mathrm{T}}\boldsymbol{x}_{si}\boldsymbol{W}_s-(\boldsymbol{x}_{si})^{\mathrm{T}}\boldsymbol{x}_{tj}\boldsymbol{W}_t),$$

$$\frac{\partial\mathrm{tr}(\boldsymbol{KG})}{\partial\boldsymbol{W}_t}=\frac{\partial\sum_{i,j=1}^{n}\boldsymbol{K}_{ij}\boldsymbol{G}_{ij}}{\partial\boldsymbol{W}_t}$$

$$=-2\gamma\sum_{i,j=1}^{n_t}\boldsymbol{G}_{(n_s+i)(n_s+j)}\mathrm{e}^{-\gamma\|(\boldsymbol{x}_{ti}-\boldsymbol{x}_{tj})\boldsymbol{W}_t\|_2^2}(\boldsymbol{x}_{ti}-\boldsymbol{x}_{tj})^{\mathrm{T}}(\boldsymbol{x}_{ti}-\boldsymbol{x}_{tj})\boldsymbol{W}_t+4\gamma\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\boldsymbol{G}_{(n_s+i)j}\mathrm{e}^{-\gamma\|\boldsymbol{x}_{si}\boldsymbol{W}_s-\boldsymbol{x}_{tj}\boldsymbol{W}_t\|_2^2}((\boldsymbol{x}_{tj})^{\mathrm{T}}\boldsymbol{x}_{tj}\boldsymbol{W}_t-(\boldsymbol{x}_{tj})^{\mathrm{T}}\boldsymbol{x}_{si}\boldsymbol{W}_s),$$

因此,分别把关于 $\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 的 3 项相加,即得到公式(9)和(10). 注意, $\boldsymbol{G}_i(n_s+j)$ 与 $\boldsymbol{G}(n_s+i)j$ 均为 0.

对于公式(7)和(8),推导类同,且比公式(9)和(10)更易处理,所以省略.

参考文献

[1] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1 345-1 359.

[2] Yang Q, Chen Y, Xue G, et al. Heterogeneous transfer learning for image clustering via the social web[C]// Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Suntec, Singapore, 2009: 1-9.

[3] Zhu Y, Chen Y, Lu Z, et al. Heterogeneous transfer learning for image classification[C]// Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 1 304-1 309.

[4] Dai W Y, Chen Y Q, Xue G R, et al. Translated learning: transfer learning across different feature spaces[C]// Proc 21st Ann Conf Neural Information Processing Systems. Vancouver, Canada, 2008.

[5] Wang C, Mahadevan S. Heterogeneous domain adaptation using manifold alignment[C]// Proc 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1 541-1 546.

[6] Duan L, Xu D, Tsang I W. Learning with augmented features for heterogeneous domain adaptation[C]// Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, UK, 2012.

[7] Kulis B, Saenko K, Darrell T. What you saw is not what you get: domain adaptation using asymmetric kernel transforms[C]// Computer Vision and Pattern Recognition, Colorado. USA, 2011: 1 785-1 792.

[8] Guan Z D, Bai S B, Zhu T S. Heterogeneous domain adaptation using linear kernel[C]// ICPCA-SWS. Vina del Mar, Chile, Springer, 2013.

[9] Gu Q Q, Li Z H, Han J W. Learning a kernel for multi-task clustering[C]// Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011.

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

[11] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel method for the two-sample-problem[C]// Proceedings of the 2006 Conference Advances in Neural Information Processing Systems 19. Vancouver, Canada, MIT Press, 2006, 20: 513-520.