ISSN 1000-1239/CN 11-1777/TP
48(9): 1722-1728, 2011

一种基于最大熵模型的加权归纳迁移学习方法

梅灿华　张玉红　胡学钢　李培培
(合肥工业大学计算机科学与技术系　合肥　230009)
(mei_414923794@126.com)

A Weighted Algorithm of Inductive Transfer Learning Based on Maximum Entropy Model

Mei Canhua, Zhang Yuhong, Hu Xuegang, and Li Peipei
(*Department of Computer Science and Technology, Hefei University of Technology, Hefei* 230009)

Abstract Traditional machine learning and data mining algorithms mainly assume that the training and test data must be in the same feature space and follow the same distribution. However, in real applications, the data distributions change frequently, so those two hypotheses are hence difficult to hold. In such cases, most traditional algorithms are no longer applicable, because they usually require re-collecting and re-labeling large amounts of data, which is very expensive and time consuming. As a new framework of learning, transfer learning could effectively solve this problem by transferring the knowledge learned from one or more source domains to a target domain. This paper focuses on one of the important branches in this field, namely inductive transfer learning. Therefore, a weighted algorithm of inductive transfer learning based on maximum entropy model is proposed. It transfers the parameters of model learned from the source domain to the target domain, and meanwhile adjusts the weights of instances in the target domain to obtain the model with higher accuracy. And thus it could speed up learning process and achieve domain adaptation. The experimental results show the effectiveness of this algorithm.

Key words machine learning; data mining; transfer learning; maximum entropy; inductive; AdaBoost

摘　要　传统机器学习和数据挖掘算法主要基于两个假设:训练数据集和测试数据集具有相同的特征空间和数据分布.然而在实际应用中,这两个假设却难以成立,从而导致传统的算法不再适用.迁移学习作为一种新的学习框架能有效地解决该问题.着眼于迁移学习的一个重要分支——归纳迁移学习,提出了一种基于最大熵模型的加权归纳迁移学习算法 WTLME.该算法通过将已训练好的原始领域模型参数迁移到目标领域,并对目标领域实例权重进行调整,从而获得了精度较高的目标领域模型.实验结果表明了该算法的有效性.

关键词　机器学习;数据挖掘;迁移学习;最大熵;归纳式;AdaBoost

中图法分类号　TP181

目前大多数机器学习和数据挖掘算法通常都假设训练数据集和测试数据集具有相同的特征空间和数据分布,然而在现实世界中这两者常常发生变化,从而导致已训练好的模型很容易过时.另外,传统的

收稿日期:2010-01-25;**修回日期**:2011-01-20
基金项目:国家"九七三"重点基础研究发展计划基金项目(2009CB326203);国家自然科学基金项目(60975034);安徽省自然科学基金项目(090412044)

数据挖掘算法通常仅对一个领域建立模型,难以将该模型适用到其他领域.如 Web 文本分类中的一个典型实例[1-2]:若已有某个大学 Web 站点的手工标记的网页,而现需对另一个新建立的 Web 站点构建文本分类器,它的数据特征或分布可能与前者不同,且缺乏类别标记.对于需要大量标记数据来构建新站点分类器的传统文本挖掘技术而言,大多数算法似乎都束手无策.因为事实上不仅数据难以获取,且标记数据所需代价也很大.然而,迁移学习却能有效地解决该问题.

迁移学习[1]指的是一个系统具有识别和应用先前任务中学习到的知识和技巧到新的任务或领域的能力.例如学会了识别苹果可以帮助识别梨子;学会了意大利语学习西班牙语也会容易些.它主要分为归纳迁移学习、直推迁移学习、无监督迁移学习三大分支.目前研究最多的是归纳迁移学习,即原始领域和目标领域数据分布不同,原始领域拥有大量数据而目标领域仅有少量标记数据的情况.

本文同样致力于归纳迁移学习的研究,提出了一种新的归纳迁移学习算法(weighted algorithm of inductive transfer learning based on maximum entropy model, WTLME),该算法基于最大熵(MaxEnt)模型[3],利用实例加权技术,将原始领域学习到的模型参数成功迁移到目标领域,从而减少了重新收集大量目标领域数据并标记数据的代价以及从头开始训练模型的时间,实现了领域适应性.

1 相关工作

迁移学习的提出弥补了传统机器学习和数据挖掘的缺陷,已在文本挖掘[2]、图像挖掘[4]、情感挖掘[5]、用户兴趣挖掘[6]、名词实体识别[7]等领域受到关注.目前大多数的研究工作主要集中在归纳迁移学习分支,采用的基本模型有支持向量机(SVM)[2]、朴素贝叶斯(NB)[8]、最大熵(MaxEnt)[9]等.限于篇幅,以下主要介绍与本文相关的工作.

Dai 等人将传统的 AdaBoost 算法[10]进行扩展,提出了该领域经典的 TrAdaBoost 算法[2],并应用在文本分类上.该算法将原始领域数据和目标领域数据混合在一起训练,并通过调整训练实例的权重来自动过滤分布不同的数据,但其以不同的方式更新原始领域和目标领域中被错误分类的实例权重,且仅在目标领域数据集上计算错误率.

Daumé III 等人提出了 MEGA(maximum entropy genre adaptation)模型[9],并将该模型应用在自动内容抽取(automatic content extraction)领域.该文认为所有的实例特征、参数均拥有两个版本:共有的和特有的.它使用类似于 EM(expectation maximization)的算法来估计所有参数,使用 MaxEnt 来对同一个特征的不同版本分配权重.由于该算法需要对所有的特征、实例、参数的两个版本进行训练,因此算法复杂度较高.

Chelba 和 Acero 结合 MAP(maximum a posterior)和 MEMM(maximum entropy Markov model)技术,提出了适用于大小写标注的 AMEC(adaptation of maximum entropy capitalizer)模型[11].该算法在目标领域和原始领域特征集的并集上建立目标领域的模型,并对共同特征赋予与原始领域相同的权重,对目标领域新特征的初始权重赋为零.该文证明了少量的目标领域数据有助于迁移学习.

以上提及的归纳迁移学习算法,要么是迁移原始领域和目标领域中的加权实例[2],要么是迁移两个领域的模型参数[11],前者旨在有效利用已有数据,后者旨在有效利用已有模型,然而均未同时考虑这两点.因此,本文利用特征权重信息,实现了原始领域到目标领域的模型参数迁移,并调整目标领域实例权重,更有效地利用了目标领域宝贵的标记数据信息,同时也获得了精度较高的目标领域分类器模型.

2 基于最大熵模型的加权归纳迁移学习方法

对于归纳迁移学习而言,要成功实现原始领域到目标领域的迁移主要需要解决两个问题:一是"迁移什么",即利用何种共同信息来连接原始领域和目标领域;二是"如何实现迁移",即在有了迁移的对象之后,应该使用怎样的策略使得共同的信息能够顺利迁移到目标领域,如何有效地利用目标领域的少量数据使迁移的效果更佳.

由于"迁移什么"这个问题与所选取的基本分类模型有很大关系,因此,本文首先选择适合迁移学习的基本分类器模型,然后在此基础上选择参数迁移策略和实例调整方式以解决"如何实现迁移"的问题.

2.1 基本分类器模型的选择

本文选择自然语言处理中功能强大的统计模型——最大熵作为原始领域和目标领域的基本分类器模型,主要基于以下考虑[3,12]:最大熵所采用的特征函数可任意组合,而不会破坏整个模型的一致性,

因而可以灵活地利用特征函数信息进行迁移;同时,它还自然地解决了统计模型中参数平滑的问题,因此,即使目标领域中部分特征未在原始领域出现,也不会影响迁移的效果;另外,最大熵模型无需进行条件独立性假设,对数据的先验分布也不作任何要求,因此,当原始领域和目标领域数据分布不同时,可避免对数据分布作过多的假设.

最大熵的基本思想是在满足所有已知事件的基础上对未知事件作最客观均一的估计[3]. 其模型的形式化表示为

$$p_\Lambda(y \mid x) = \frac{1}{Z(x)}\exp\left(\sum_{j=1}^{m} f_j(x,y)\lambda_{j,y}\right), \quad (1)$$

其中:y 是实例 x 对应的类标签;$f(x,y)$是从训练数据集中抽取出来的约束,称为特征函数(简称特征),所有特征的集合记为 $F=\{f_1,\cdots,f_m\}$,m 为 F 中的特征数目;λ_j 是模型参数,可视为 f_j 所对应的权重,所有特征权重的集合记为 $\Lambda=\{\lambda_{1,y},\cdots,\lambda_{m,y}\}$;$Z(x)$为归一化因子.

在式(1)中,仅 Λ 是未知参数,其他均为已知. 为求解最大熵的模型参数 Λ,本文采用的是 Darroch 和 Ratcliff 提出的 GIS(generalized iterative scaling)算法[13].

2.2 最大熵加权归纳迁移学习算法

为了在原始领域和目标领域进行知识迁移,必须假设两个领域是相关的,因此需要一个知识迁移的桥梁来连接原始领域和目标领域. 受文献[11]启发,本文同样将最大熵模型的参数作为知识迁移的桥梁来连接这两个领域,并认为相似的两个领域其分类器间共享了参数蕴含的信息. 然而不同于文献[11],本文认为由于数据分布不同,共同特征对两个领域的分类作用不能视为等同,且即使是相同的特征,在不同的领域中,每一个特征相对于其他特征而言,对该领域的分类影响度也应有所差异,故本文提出以下策略来解决该问题.

1) 参数传递

为了反映同一个特征对不同领域的作用差异,本文针对原始领域 s 和目标领域 t 中共同特征 f_j,定义了 t 相对于 s 的差异因子 $\delta_j(t,s)$:

$$\delta_j(t,s) = \frac{\tilde{E}^t(f_j^t)}{\tilde{E}^s(f_j^s)}, \quad (2)$$

其中:$\tilde{E}^t(f_j^t)$与 $\tilde{E}^s(f_j^s)$分别为 t 与 s 中特征函数 f_j 的经验期望值.

假设原始领域和目标领域的特征集合分别表示为 F^s 与 F^t,原始领域已经训练好的分类器模型参数为 $\Lambda^s=\{\lambda_{1,y}^s,\cdots,\lambda_{m,y}^s\}$. 对于 F^t 中任意一个特征 f_j^t,若能在 F^s 中找到与其相同的特征,则将相同特征的权重乘以差异因子作为初始权重赋给特征 f_j^t,否则对该特征从头开始学习,权重赋为 0. 即:

$$\lambda_j^t = \begin{cases} \delta_j(t,s)\lambda_i^s, & f_j^t = f_i^s; \\ 0, & \text{其他}. \end{cases} \quad (3)$$

由式(2)(3)可知:若某个特征在目标领域出现频繁而在原始领域出现较少,则赋予其较大的权重;反之,若其在原始领域出现频繁而目标领域出现较少,则赋予其较小的权重. 这使得共有特征的权重能够根据目标领域特征的统计信息从原始领域传递到目标领域.

2) 实例权重调整

由于目标领域数据量较少,为了更合理地利用少量数据蕴含的信息,使训练的模型更为准确,本文采用了类似于 AdaBoost 算法思想来调整实例权重,并根据实例权重来计算先验概率. 假设目标领域数据表示为 $D^t=\{(x_1^t,y_1^t),\cdots,(x_{n_t}^t,y_{n_t}^t)\}$,其中 n_t 为 D^t 中的实例数目. D^t 中实例的权重表示为 $W=\{w_1,\cdots,w_{n_t}\}$. 则对于 D^t 中任意实例 x_j,其先验概率 $\tilde{p}(x_j)$的计算公式为

$$\tilde{p}(x_j) = w_j \Big/ \sum_{i=1}^{n_t} w_i. \quad (4)$$

错误率为

$$e = \sum_{j=1}^{n_t} \tilde{p}(x_j) \mid h(x_j) - y_j \mid, \quad (5)$$

其中,$h(x_j)$为目标预测函数,用来预测实例 x_j 的类标签. 且满足:$h(x_j)-y_j=\begin{cases}1, & h(x_j)=y_j,\\ 0, & h(x_j)\neq y_j.\end{cases}$

令 $\beta=1/2\ln(1-e)/e$,则第 $k+1$ 次迭代的权重为

$$w_j(k+1) = \begin{cases} w_j(k)e^{-\beta}, & h(x_j) = y_j, \\ w_j(k)e^{\beta}, & h(x_j) \neq y_j. \end{cases} \quad (6)$$

文献[10]中对 AdaBoost 方法的稳定性和收敛性进行了详细的理论分析和证明. 该文表明 AdaBoost 的实例加权方法是稳定的、收敛的,尽管其收敛速度受到基分类器性能的影响,但在基分类器性能优于随机分类器条件下该方法总是收敛的.

3) 算法描述

为了便于算法描述,首先介绍所涉及的符号:Λ^s 与 Λ^t 分别表示原始领域和目标领域对应的模型参数集合;D_{train}^t与 D_{test}^t 分别表示目标领域的训练集和测试集;e_{train} 与 e_{test} 分别表示 D_{train}^t 与 D_{test}^t 的错误

率;k 记录 WTLME 算法的迭代次数;θ 为错误率阈值变量,且 $0<\theta\leqslant 0.5$.

算法 1. WTLME 算法.

输入:原始领域 F^s,Λ^s,$\{\tilde{E}^s(f)\mid f\in F^s\}$;

目标领域 F^t,D^t_{train},D^t_{test};

输出:最优化参数 Λ^t 和目标领域分类器 $p^t_{\Lambda^t}$.

① 初始化 W,Λ^t,e_{train},e_{test} 和 k;

② 根据式(4)计算 $\tilde{p}(x_j)$ 其中,$1\leqslant j\leqslant n_t$;令 $k=k+1$;

③ 调用 GIS 算法,在训练数据 D^t_{train} 上,训练分类器 $p^t_{\Lambda^t}(k)$;

④ 根据式(5)计算 $e_{\text{train}}(k)$ 和 $e_{\text{test}}(k)$;

⑤ 根据式(6),更新权重集合 W;

⑥ 如果 $e_{\text{test}}(k)<e_{\text{test}}(k-1)$ 且 $e_{\text{train}}(k)<\theta$,转②;

⑦ 否则,如果 $e_{\text{train}}(k)>e_{\text{train}}(k-1)$,$p^t_{\Lambda^t}=p^t_{\Lambda^t}(k-1)$;

否则 $p^t_{\Lambda^t}=p^t_{\Lambda^t}(k)$;

结束.

由于 WTLME 算法是利用已有的原始领域分类器模型进行迁移学习,故在实际应用中,需要先在原始领域标记数据 $D^s=\{(x^s_1,y^s_1),\cdots,(x^s_{n_s},y^s_{n_s})\}$ 上训练出原始领域分类器模型,再将其模型参数 Λ^s 作为 WTLME 算法的输入.其主要分为三大步骤(如算法 1 所示):第 1 步①,初始化参数 W,Λ^t,e_{train},e_{test} 与 k,其中:$W=\{w_1,\cdots,w_{n_t}\}=\{1,\cdots,1\}$;$\Lambda^t$ 根据式(3)初始化;e_{train} 与 e_{test} 取最大值 1,使算法至少迭代一次;$k=0$.第 2 步②~⑤,首先根据式(4)计算 $\tilde{p}(x_j)(1\leqslant j\leqslant n_t)$,再调用 GIS 算法重新计算分类器模型 $p^t_{\Lambda^t}$,然后根据式(5)(6)分别计算该模型在 D^t_{train} 和 D^t_{test} 上的错误率 e_{train} 和 e_{test} 以及新一轮的实例权重 W.第 3 步⑥⑦,若训练数据集的当前错误率 e_{train} 在可接受的阈值范围 θ 内,且测试数据集的当前错误率 e_{test} 降低,则进行下一轮迭代;否则根据 e_{train} 获得最终模型 $p^t_{\Lambda^t}$:若 e_{train} 增大,则保留上一次迭代获取的模型,否则,取当前模型作为最终模型.综上所述,WTLME 算法每进行一次迭代均会更新 W,以重新计算每一个 $\tilde{p}(x_j)$,从而调整了最大熵分类器模型.

2.3 时间性能分析

训练一个 MaxEnt 基本分类器模型的时间复杂度为线性函数 $O(lncf)$,其中 l 为 MaxEnt 的迭代次数,n 为实例数目,c 为类别数目,f 为特征函数数目.WTLME 算法对 MaxEnt 进行了 k 次迭代,故其时间复杂度也为线性函数 $O(klncf)$.在实际实验中,由于 c 一般较小,迭代次数 k 与 l 一般也不会超过 200,因此,算法时间性能主要由用户定义的特征函数数目 f 和训练实例数目 n 决定.

3 实验结果与分析

3.1 实验数据集

本文实验数据集是文本分类中常用的语料库 20 Newsgroup.它包含 7 个顶级类别、20 个子类别,共近 20 000 个新闻文档.本文使用与文献[2]中相同的采样方法来生成原始领域和目标领域相似但分布不同的数据集,即将顶层类别下不同子类别的数据重新组合来生成原始领域和目标领域数据集.例如:分类任务是判别文本是属于顶层类别中 comp 还是 rec,则可以从它们的子类别 comp.sys.mac.hardware 和 rec.sport.hockey 中选取数据来构成原始领域数据集,从 comp.windows.x 和 rec.sport.baseball 中选取数据来构成目标领域数据集.按此方法,本文将其分成了 5 组数据集:comp vs rec,rec vs sci,comp vs sci,rec vs talk 和 sci vs talk.

为了便于处理,并与文献[11]中算法思想进行比较,本文采用与其相同的 cut-off 特征选择算法来对数据进行预处理.

3.2 实验性能对比

为验证 WTLME 算法的有效性,本文分别从传统机器学习和迁移学习两个方面设计实验进行了对比.一方面,本文选取了传统文本分类中常用的模型 NB,SVM 和自然语言处理中常用的模型 MaxEnt 作对比,以证明迁移学习的必要性;另一方面,与文献[11]中 AMEC 模型的迁移算法思想进行对比(本文以最大熵模型替代其中的最大熵马尔可夫模型),以证明本文算法的优越性.以上算法均是在 Windows XP 操作系统、1 GB 内存、Inter P4 2.8 GHz CPU、Java 编程环境 Eclipse3.4 下实现.

对于原始领域训练数据数目 n_s 和目标领域训练数据数目 n_t,迁移学习中一般要求 $n_s\gg n_t\geqslant 0$[1].故在实验中,需满足 n_s 至少高于 n_t 一个数量级.若设数据比例 $r=n_s/n_t$,则 $r\geqslant 10$.本文对每组数据集分别取 $r=10,30,50,80,100$ 的数据比例进行了实验,且所有实验结果均为运行 10 次所得的平均值.

由于 WTLME 算法采用了类似于 AdaBoost 的迭代框架,同 AdaBoost 算法一样,基本分类器模型

仅需是弱分类器,故实验中错误率阈值变量 θ 可取最大值 0.5,且基本分类器 MaxEnt 的迭代次数 l 也无需设置过大,以避免对目标领域少量的训练数据过拟合.在本文实验中,最佳分类效果时的 WTLME 的参数 l 取值均低于 50,为便于比较本文取 $l=50$,这略低于通常迭代次数(100～200 次)[14].

1) 分类精度对比

表 1 展示了传统机器学习算法和迁移学习算法在不同数据集上的分类错误率.由表 1 可知:传统机器学习算法在对分布不同的任务进行分类时效果较差,准确率一般在 50%～70%之间.而迁移学习算法则能够预测新领域知识,准确率基本可达 80%以上,这说明 WTLME 和 AMEC 算法均能提高新领域模型的预测准确率,具有领域适用性.另外,与 AMEC 相比,WTLME 也具有一定的优势,除在 rec vs talk 数据集上,WTLME 的准确率仅高于 AMEC 0.9%(在该数据集上,AMEC 准确率已较高,提升空间很小),而在其他数据集上,WTLME 的准确率基本能比 AMEC 提高 2%～5%.

Table 1 Classification Error Rates of Different Algorithms ($r=50$)

表 1 不同算法的分类错误率($r=50$)

Data Set	No Transfer			Transfer	
	NB	SVM	MaxEnt	AMEC	WTLME
comp vs rec	0.367	0.218	0.224	0.172	0.149
rec vs sci	0.287	0.342	0.318	0.175	0.145
comp vs sci	0.324	0.323	0.279	0.242	0.187
rec vs talk	0.509	0.285	0.332	0.089	0.080
sci vs talk	0.343	0.332	0.346	0.158	0.142

图 1 以数据集 comp vs rec 为例,展示了 3 种基于最大熵技术的分类算法:无迁移的 MaxEnt 和有迁移的 AMEC 与 WTLME 随 r 的降低分类错误率的变化趋势.由图 1 可知:在无迁移时,MaxEnt 的准确率保持在 65%～70%之间;当数据量充足时($r<10$),WTLME 与 AMEC 相当,准确率比无迁移的 MaxEnt 提高 20%左右;而当数据量不足时($r\geqslant10$),与无迁移的 MaxEnt 相比,WTLME 的准确率提高幅度超过 AMEC,且随着 r 的增大其提升幅度逐渐增大,在 $r=100$ 时,WTLME 的准确率比 AMEC 提高了 10%左右,其主要原因在于 WTLME 相较 AMEC 而言多了一个实例加权迭代训练的过程,弥补了 AMEC 未充分利用目标领域实例信息的缺陷,从而进一步提高了目标领域的分类精度,故 WTLME 算法在处理少量数据时更具优越性.

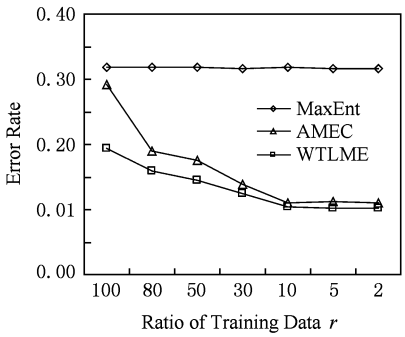


Fig. 1 Error rate changes varying with the values of r.

图 1 3 种算法的错误率随 r 变化趋势

2) 时间性能对比

根据 2.3 节中时间性能分析可知:在原始领域模型已训练好的情况下,WTLME,AMEC 的迁移时间主要取决于特征数目和实例数目,而对特定数据集而言,实验中的特征数目均相等,故 WTLME 和 AMEC 的迁移时间实际上主要取决于训练实例数目,且与实例数目成正比;而若原始领域模型尚未创建,则两种算法均需先训练原始领域模型再进行迁移学习,故最终获得目标领域模型的总时间为原始领域模型训练时间与迁移时间之和,由于原始领域模型训练时间与具体迁移算法无关,故两种算法中原始领域训练时间理论上应该完全相等.为验证以上理论分析,本文设计以下实验来对比 WTLME 与 AMEC 的迁移时间和训练总时间,如表 2、表 3 所示:

Table 2 Comparison on the Transfer Time of AMEC and WTLME

表 2 AMEC 与 WTLME 迁移时间对比(comp vs rec) ms

Algorithm	$n_s=4\,909$				
	$n_t=490$	$n_t=163$	$n_t=98$	$n_t=61$	$n_t=49$
AMEC	221	76	48	31	21
WTLME	193	70	58	39	31

Table 3 Comparison on the Total Time of AMEC and WTLME

表 3 AMEC 与 WTLME 总时间对比(comp vs rec) ms

Algorithm	$n_s+n_t=4\,909+490$	$n_s+n_t=4\,909+163$	$n_s+n_t=4\,909+98$	$n_s+n_t=4\,909+61$	$n_s+n_t=4\,909+49$
AMEC	3 797+221	4 099+76	3 942+48	3 997+31	4 032+21
WTLME	3 903+193	4 073+70	3 862+58	4 012+39	4 086+31

在本文实验中,$r=10,30,50,80,100$ 的数据集获取方法为(以 comp vs rec 为例):保持原始领域训练数据数目 $n_s=4\,909$ 不变,逐步减少目标领域训练数据数目,$n_t=490,163,98,61,49$. 表 2 展示了 WTLME,AMEC 的迁移时间与 n_t 的关系. 由表 2 可知,WTLME,AMEC 的迁移时间均随 n_t 的减少而减少,这与前面的理论分析一致;且在 n_t 相同时,WTLME 与 AMEC 实际的迁移时间也相差不大,基本处于同一数量级. 表 3 则展示了两种算法总训练时间与总训练实例数目(n_s+n_t)的关系. 由表 3 可知:WTLME 与 AMEC 总训练时间(原始领域模型训练时间+迁移时间)基本不随目标领域训练数据量变化,这主要是因为 $n_s\gg n_t$ 导致 n_t 可忽略不计,故总训练时间主要取决于原始领域模型的训练时间,因为其远远大于迁移时间,这也正说明了利用已有知识进行迁移学习比从头学习更快(另外,由于原始领域实例数目与特征数目保持不变,故 AMEC 与 WTLME 的原始领域模型训练时间理论上应完全相同,而该表中却表现出一定的差异,这应该是环境和机器时间差异造成的,应视为在一定偏差范围内两者还是相等的). 综上所述可知:AMEC 与 WTLME 时间性能相当,且迁移学习可以比从头学起更有效率.

4 结束语

本文基于最大熵技术提出了一种新的归纳迁移学习算法 WTLME. 该算法有效地利用了已有数据和已有模型,弥补了传统机器学习算法不具领域适应性的缺陷,以较少的时间代价获得了分类性能较高的目标领域模型. 然而 WTLME 算法仅考虑了从一个原始领域迁移到目标领域的情况,如何将多个原始领域知识迁移到目标领域,并将迁移学习与实际应用相结合是未来工作的重点.

参 考 文 献

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

[2] 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

[3] Berger A L, Pietra S, PIetra V. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 38-73

[4] Raina R, Battle A, Lee H, et al. Self-taught learning: Transfer learning from unlabeled data [C] //Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 759-766

[5] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification [C] //Proc of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: ACL Press, 2007: 432-439

[6] Cao B, Liu N, Yang Q. Transfer learning for collective link prediction in multiple heterogenous domains [C] //Proc of the 27th Int Conf on Machine Learning. Haifa, Israel: Omnipress, 2010: 159-166

[7] Andrew A, Ramesh N, William W C. A comparative study of methods for transductive transfer learning [C] //Proc of the 7th IEEE Int Conf on Data Mining Workshops. Los Alamitos, CA: IEEE Computer Society, 2007: 77-82

[8] Dai W, Xue G, Yang Q, et al. Transferring naive Bayes classifiers for text classification [C] //Proc of the 22nd AAAI Conf on Artificial Intelligence. Vancouver, British Columbia: AAAI Press, 2007: 540-545

[9] Daumé III H, Marcu D. Domain adaptation for statistical classifiers [J]. Journal of Artificial Intelligence Research, 2006, 26: 101-126

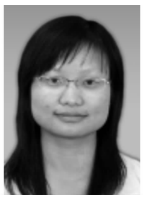
[10] Freund Y, Schapire R E. A short introduction to boosting [J]. Journal of Japanese Society for Artificial Intelligence, 1999, 14(5): 771-780

[11] Chelba C, Acero A. Adaptation of maximum entropy capitalizer: Little data can help a lot [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Barcelona, Spain: ACL Press, 2004: 285-292

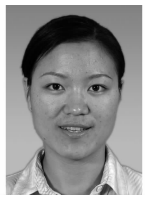
[12] Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing, 9708 [R]. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania, 1997

[13] Darroch J, Ratcliff D. Generalized iterative scaling for log-linear models [J]. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480

[14] Zhou Yaqian, Guo Yikun, Huang Xuanjing, et al. Chinese and English baseNP recognition based on a maximum entropy model [J]. Journal of Computer Research and Development, 2003, 40(3): 440-446 (in Chinese)
(周雅倩, 郭以昆, 黄萱菁, 等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003, 40(3): 440-446)



Mei Canhua, born in 1985. MSc candidate at Hefei University of Technology. Her main research interests are data mining and artificial intelligence.



Zhang Yuhong, born in 1979. PhD candidate and lecturer. Her main research interests include data mining and artificial intelligence(zhangyuhong99@163.com).

Hu Xuegang, born in 1961. PhD, Professor and PhD supervisor of Hefei University of Technology. His main research interests include data mining and artificial intelligence.

Li Peipei, born in 1983. PhD candidate. Her main research interests include data mining and artificial intelligence.