Computer Science and Application 计算机科学与应用, 2020, 10(2), 340-349
Published Online February 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.102035

Hans汉斯

A Label Proportion Information-Based Transfer Learning Algorithm

Huaipei Wang[1], Yanshan Xiao[1], Bo Liu[2]

[1]School of Computers, Guangdong University of Technology, Guangzhou Guangdong
[2]School of Automation, Guangdong University of Technology, Guangzhou Guangdong
Email: gdsgwhp@163.com

Received: Feb. 4th, 2020; accepted: Feb. 18th, 2020; published: Feb. 25th, 2020

Abstract

The learning with label proportions problem is a learning task that only uses bag's label proportions information to build a classification model. Due to insufficient training samples, the existing methods that viewed the above problem as single task did not perform well in text classification. To some extent, transfer learning can solve the problem of insufficient training data, the problem that how to use historical data (the original task data) to help the newly generated data (target task data) to classify becomes extremely important. This paper presents a label proportion information-based transfer learning approach to transfer knowledge from the source task to the target task, helping the target task to build a classifier. In order to obtain the transfer learning model, this method converted the original optimization problem into a convex optimization problem, and then solved the dual optimization problem to establish an accurate classifier for the target task. Extensive experiments have shown that the proposed method outperforms the traditional methods.

Keywords

Learning with Label Proportions, Data Mining, Transfer Learning

一种基于标签比例信息的迁移学习算法

汪槐沛[1]，肖燕珊[1]，刘　波[2]

[1]广东工业大学计算机学院，广东 广州
[2]广东工业大学自动化学院学院，广东 广州
Email: gdsgwhp@163.com

收稿日期：2020年2月4日；录用日期：2020年2月18日；发布日期：2020年2月25日

文章引用：汪槐沛，肖燕珊，刘波. 一种基于标签比例信息的迁移学习算法[J]. 计算机科学与应用，2020，10(2): 340-349. DOI: 10.12677/csa.2020.102035

摘　要

标签比例学习问题是一项仅使用样本标签比例信息去构建分类模型的挖掘任务，由于训练样本不充分，现有方法将该问题视为单一任务，在文本分类中的表现并不理想。考虑到迁移学习在一定程度上能解决训练数据不充分的问题，于是如何利用历史数据(原任务数据)帮助新产生的数据(目标任务数据)进行分类显得异常重要。本文提出了一种基于标签比例信息的迁移学习算法，将知识从原任务迁移到目标任务，帮助目标任务更好构建分类器。为了获得迁移学习模型，该方法将原始优化问题转换为凸优化问题，然后解决对偶优化问题为目标任务建立准确的分类器。实验结果表明，大部分条件下所提算法性能优于传统方法。

关键词

标签比例学习，数据挖掘，迁移学习

1. 引言

在传统监督学习的分类问题中，已知所有样本的标签，分类器可以通过大量样本属性及其标签学习得到，进而利用学习得到的分类器对未知标签的样本进行预测。但在实际应用中，通过人工标注获取样本标签需要较高成本，或者受限于隐私等客观条件，有时无法获取所有样本的标签，而仅仅已知各类样本的标签比例信息，比如在匿名投票中，只能知道反对票和赞成票的比例。因此，在已知样本标签比例信息的前提下，以多个样本组成的包为单位，基于包内样本和包的标签比例信息来训练从而获取样本层面的分类器，更加具有实用价值。

近年来，标签比例学习[1] [2] (Learning with Label Proportions, LLP)在数据挖掘引起了广泛的关注，并成功应用于现实生活中的许多领域，如欺诈识别、银行重要客户识别、垃圾邮件过滤、视频事件检测、收入预测、视觉特征建模等。在标签比例学习问题中，只知道每个包中属于不同类别的样本的比例，但是样本的标签是未知的，它基于包层面的标签比例信息解决了样本层面的分类问题。

迁移学习(Transfer Learning) [3] [4]是可以将知识从原任务(Source task)迁移到目标任务(Target task)的一种新的机器学习方法，其运用已存有的知识对不同但相关领域问题进行求解，迁移的知识可以帮助目标任务建立迁移学习分类器以进行预测。然而，大多数现有的方法都没有考虑实践中从原任务到目标任务的知识迁移，将标签比例学习视为单一任务，无法解决迁移学习问题。

综上所述，本文针对标签比例学习问题，为了训练得到更准确的分类器，提出了一种基于标签比例信息的迁移学习算法(label proportion information-based transfer learning method, LPI-TL)，该方法可以利用迁移学习将知识从原任务迁移到目标任务，帮助目标任务构建分类器。首先为了帮助目标任务学习预测模型，本文提出了一种迁移学习模型，然后使用拉格朗日方法将方法的原始问题转换为凸优化问题并求解，最后获得目标任务的预测分类器。实验结果表明，本文方法在标签比例问题上能取得更好的性能。

本文主要贡献如下：

1) 结合支持向量回归算法提出了基于标签比例信息的迁移学习模型，该模型可以利用迁移学习将知识从原任务迁移到目标任务。

2) 利用拉格朗日方法将原始目标模型转换为凸优化问题，并获得原任务和目标任务的预测模型。

3) 在多个数据集上进行广泛实验，并与现有算法进行对比，验证了提出算法的有效性。

2. 问题描述与相关工作

2.1. 问题描述

在标签比例学习问题中，一个包内含有多个样本，仅知道包中不同类别样本的标签比例信息。本文定义包的标签比例为包中正样本的比例。假设给定的原任务数据集为 $D=\{x_1,x_2,\cdots,x_n\}$，则每个样本 x_i 所对应的标签 y_i 未知，数据会被分为 t_1 个互相独立的包 $\left(B_I^s,P_I^s\right),I=1,2,\cdots,t_i$，其中 B_I^s 和 P_I^s 分别表示原任务数据集的第 I 个包和包中正样本的比例 $P_I^s=\left|\left\{x_i\in B_I^s:y_i=1\right\}\right|\Big/\left|B_I^s\right|$，同理，目标任务数据集用 $\left(B_J^t,P_J^t\right),J=1,2,\cdots,t_2$ 表示。

对于二元分类问题，标签比例学习任务是学习一个分类器将未知标签样本分为正类或负类。如图 1 所示：图左边的黑色椭圆表示包，黑色圆圈表示未标记的样本。在图的右边，加号“+”和减号“−”分别表示分类后的正样本和负样本，实线表示由标签比例和未标记的样本训练得到的分类器。

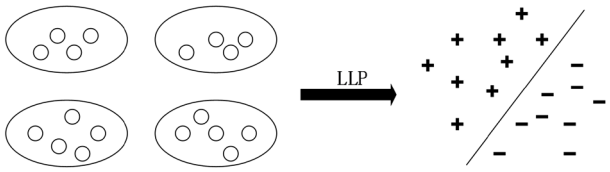

Figure 1. Two-class label proportions learning problem
图 1. 二分类标签比例学习问题

2.2. 相关工作

本目前，国内外有多种标签比例学习分类方法的研究，主要分为概率模型和支持向量机模型两类。

基于概率模型的分类方法。文献[5]假设给定类变量的预测变量之间具有条件独立性，并采用了三种期望最大化算法来学习朴素贝叶斯模型。文献[6]通过估计条件类密度来估计后验概率，从贝叶斯角度提出一种新的学习框架，并且利用估计对数概率的网络模型来求解分类问题的后验概率。文献[7]构建了一个应用于美国总统大选的概率方法，该方法使用基数势在学习过程中对潜在变量进行推理，并引入了一种新的消息传递算法，将基数势扩展到多变量概率模型。文献[8]开发模型并使用 Twitter 数据来估算美国总统大选期间政治情绪与人口统计之间的关系。

基于支持向量机(Support Vector Machine, SVM)的分类方法。文献[9]提出 InvCal 算法，该方法可以利用样本的比例信息，反推出分类器，使得分类器预测出的样本比例与实际比例相近。文献[10]提出了 Alter-SVM 算法并通过交替优化的方法最小化损失函数。该方法在标准 SVM 模型的基础上加入了比例损失的约束项，使得模型得到每个包的比例与实际比例尽可能接近。在文献[10]的基础上，文献[11]提出了一种基于二支持向量机的分类模型，模型被转换为两个更小的二分类问题求解。文献[12]首先分析了比例学习问题中的结构化信息，并利用数据点的几何信息引入拉普拉斯项并且讨论了如何将比例学习框架与拉普拉斯项结合。文献[13]提出了一种基于非平行支持向量机的解决方案并将分类器改进为一对非平行分类超平面。

尽管对标签比例学习的研究已经比较深入，然而大部分研究仅将该问题视为单一任务，没有利用历

史数据弥补训练样本不充分的不足，不能很好地体现出分类器的效果。不同于大部分已有的工作，本文从迁移学习角度出发，利用迁移学习可以在相似领域中帮助新领域目标任务学习的特性，提出了一种基于标签比例信息的迁移学习方法，该方法基于支持向量回归方法并结合样本组成的包构建模型，并给出了目标方程从而解决了迁移学习方法运用于标签比例学习分类的问题。

3. 标签比例学习算法

由于只通过标签比例信息无法直接训练分类函数，参考 InvCal [9]中利用标签比例信息反推出分类器的操作，提出的方法使用 Platt 尺度函数[14]并反解求出 y：

$$y=-\log\left(\frac{1}{p}-1\right) \tag{1}$$

则每个包预测的 y 值可表示为：

$$\forall_i:\frac{1}{|B_i|}\sum_{j\in B_i}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j+b\right)=y_i\text{。} \tag{2}$$

3.1. 目标函数

除对于具有相关性的原任务数据集和目标任务数据集，该算法使用 f_1 和 f_2 分别表示原任务和目标任务分类器：

$$f_1(\boldsymbol{x})=\boldsymbol{w}_1^{\mathrm{T}}\cdot\boldsymbol{x}+b_1 \tag{3}$$

$$f_2(\boldsymbol{x})=\boldsymbol{w}_2^{\mathrm{T}}\cdot\boldsymbol{x}+b_2 \tag{4}$$

其中 $\boldsymbol{w}_1=\boldsymbol{w}_0+\boldsymbol{v}_1$，$\boldsymbol{w}_2=\boldsymbol{w}_0+\boldsymbol{v}_2$，$\boldsymbol{w}_0$ 表示分离超平面的权向量公共参数，$\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 为增量参数，原任务和目标任务越相似，则 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 越“小”，b_1 和 b_2 是偏差。

本文对每个包中所有样本求平均值得到一个平均数样本来代表包，并利用支持向量回归算法对所得到的平均数样本求回归方程，于是 LLP 问题转换为求解如下目标函数：

$$\min\quad\frac{1}{2}\|\boldsymbol{w}_0\|^2+\frac{\lambda_1}{2}\|\boldsymbol{v}_1\|^2+\frac{\lambda_2}{2}\|\boldsymbol{v}_2\|^2+C_1\sum_{i=1}^{t_1}\left(\xi_{1i}+\xi_{1i}^*\right)+C_2\sum_{m=1}^{t_2}\left(\xi_{2m}+\xi_{2m}^*\right) \tag{5}$$

约束条件：

$$\forall_{i=1}^{t_1}:\begin{array}{l}\dfrac{1}{\left|B_i^s\right|}\displaystyle\sum_{j\in B_i^s}\left(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j+b_1\right)-y_i\le\varepsilon_{1i}+\xi_{1i}\\ y_i-\dfrac{1}{\left|B_i^s\right|}\displaystyle\sum_{j\in B_i^s}\left(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j+b_1\right)\le\varepsilon_{1i}+\xi_{1i}^*\end{array}$$

$$\forall_{m=1}^{t_2}:\begin{array}{l}\dfrac{1}{\left|B_m^t\right|}\displaystyle\sum_{n\in B_m^t}\sum_{n\in B_m^t}\left(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_n+b_2\right)-y_m\le\varepsilon_{2m}+\xi_{2m}\\ y_m-\dfrac{1}{\left|B_m^t\right|}\displaystyle\sum_{n\in B_m^t}\left(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_n+b_2\right)\le\varepsilon_{2m}+\xi_{2m}^*\end{array}$$

$$\xi_{1n},\xi_{1n}^*\ge0\left(n=1,\cdots,t_1\right)$$

$$\xi_{2m},\xi_{2m}^*\ge0\left(m=1,\cdots,t_2\right)$$

其中：ξ_{ti} 和 $\xi_{ti}^*\left(t=1,2\right)$ 为训练误差，ε_{1i} 和 ε_{2m} 为可容忍损失，控制 ε 损失带的大小；参数 λ_1 和 λ_2 用来控

制原任务和目标任务权重，C_1 和 C_2 是边缘与经验损失的权衡参数。

为了更好地理解迁移学习，图 2 示意了知识迁移的基本思想。图 2 中圆圈表示样本，圆圈内数字表示样本所属包的编号，每个包都有两个未知标签的样本，只知道包中正样本的比例为 50%；实线和虚线则分别表示分类器和间隔边界。在图的右边，红色圆圈表示目标任务样本，当仅利用目标任务样本构建分类。

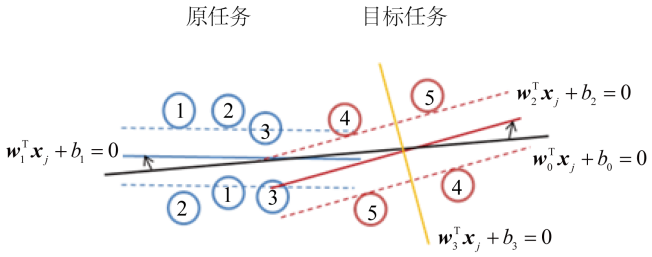

Figure 2. Transfer knowledge from the source task to the target task
图 2. 原任务迁移知识到目标任务

3.2. 对偶问题

由于目标函数公式 5 为凸函数，引入拉格朗日乘子，公式 5 转换为其对偶问题，等同为最小化下式：

$$\begin{aligned}
&\frac{1+\lambda_1}{2\lambda_1}\sum_{i,j=1}^{t_1}\frac{\left(\alpha_{1i}^*-\alpha_{1i}\right)\left(\alpha_{1j}^*-\alpha_{1j}\right)}{\left|B_i^s\right|\left|B_j^s\right|}K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\\
&+\frac{1+\lambda_2}{2\lambda_2}\sum_{m,n=1}^{t_2}\frac{\left(\alpha_{2m}^*-\alpha_{2m}\right)\left(\alpha_{2n}^*-\alpha_{2n}\right)}{\left|B_m^t\right|\left|B_n^t\right|}K\left(\boldsymbol{x}_m,\boldsymbol{x}_n\right)\\
&+\sum_{i=1}^{t_1}\sum_{m=1}^{t_2}\frac{\left(\alpha_{1i}^*-\alpha_{1i}\right)\left(\alpha_{2m}^*-\alpha_{2m}\right)}{\left|B_i^s\right|\left|B_m^t\right|}K\left(\boldsymbol{x}_i,\boldsymbol{x}_m\right)\\
&-\sum_{i=1}^{t_1}\left(y_i\left(\alpha_{1i}^*-\alpha_{1i}\right)-\varepsilon_{1i}\left(\alpha_{1i}^*+\alpha_{1i}\right)\right)\\
&-\sum_{m=1}^{t_2}\left(y_m\left(\alpha_{2m}^*-\alpha_{2m}\right)-\varepsilon_{2m}\left(\alpha_{2m}^*+\alpha_{2m}\right)\right)
\end{aligned}\tag{6}$$

约束条件：

$$\sum_{i=1}^{t_1}\left(a_{1i}-a_{1i}^*\right)+\sum_{m=1}^{t_2}\left(a_{2m}-a_{2m}^*\right)=0$$

$$\forall_{i=1}^{t_1}:0\le\alpha_{1i},\alpha_{1i}^*\le C_1$$

$$\forall_{m=1}^{t_2}:0\le\alpha_{2m},\alpha_{2m}^*\le C_2$$

其中 K 是核函数，$\alpha=\left[\alpha_{1i},\alpha_{1i}^*,a_{2m},a_{2m}^*\right]\ge 0$ 为拉格朗日乘子。公式 6 为凸函数可用现有方法直接求解出拉格朗日乘子 a，则目标任务分类器中 w_0 和 v_2 可通过下式求得：

$$\boldsymbol{w}_0=\sum_{i=1}^{t_1}\left(a_{1i}^*-a_{1i}\right)\frac{1}{\left|B_i^s\right|}\sum_{j\in B_i^s}\boldsymbol{x}_j+\sum_{m=1}^{t_2}\left(a_{2m}^*-a_{2m}\right)\frac{1}{\left|B_m^t\right|}\sum_{j\in B_m^t}\boldsymbol{x}_n\tag{7}$$

$$\boldsymbol{v}_1=\frac{1}{\lambda_1}\sum_{i=1}^{t_1}\left(\alpha_{1i}^*-\alpha_{1i}\right)\frac{1}{\left|B_i^s\right|}\sum_{j\in B_i^s}\boldsymbol{x}_j\tag{8}$$

$$\boldsymbol{v}_2 = \frac{1}{\lambda_2}\sum_{m=1}^{t_2}\left(\alpha_{2m}^{*} - \alpha_{2m}\right)\frac{1}{\left|B_m^t\right|}\sum_{n\in B_m^t}\boldsymbol{x}_n \tag{9}$$

具体算法求解过程如表 1 所示。

Table 1. LPI-TL Algorithm
表 1. LPI-TL 算法流程

算法：LPI-TL
输入：$\left(B_I^s, P_I^s\right), I = 1,2,\cdots,t_1$，$\left(B_J^t, P_J^t\right), J = 1,2,\cdots,t_2$，$\lambda_1, \lambda_2, C_1, C_2, \varepsilon$
输出：$\boldsymbol{w}_1, \boldsymbol{w}_2, b_1, b_2$
方法：
a)：**For** i=1:(t_1+t_2)
b)：$y_i = -\log\left(\frac{1}{p_i} - 1\right)$
c): **End**
d)：调用 Matlab 的 CVX 包求解公式 6，解得拉格朗日乘子 $\boldsymbol{a}$
e)：将 $\boldsymbol{a}$ 代入公式(7)-(9)求得 $\boldsymbol{w}_1$，$\boldsymbol{w}_2$
f)：将 $\boldsymbol{w}_1$，$\boldsymbol{w}_2$，y_i 代入公式 3 和公式 4 可解得 b_1，b_2

3.3. 时间复杂度分析

对于提出的 LPI-TL 算法，假设求解标准 SVM 的时间复杂度为$O\left(\left[\text{数据量}\right]^2\right)$，则在本文中解决公式 6 等同于求解一个数据量为 M 个原任务样本和 N 个目标任务样本的标准 SVM 问题，则最终的时间复杂度为$O\left((M+N)^2\right)$。

4. 实验与分析

为验证所提出算法的有效性，本文设计了数据实验来验证所述方法，并采用 Inv-Cal [8]，Alter-SVM [9]和 p-NPSVM [12]作为对比方法。

4.1. 实验数据

论本文采用 SRAA[1] 和 20 Newsgroups[2] 数据集，这两个数据集是用于文本分类、文本挖掘和信息检索研究的国际标准数据集之一,并被广泛用于迁移学习实验。SRAA 数据集包含来自四个讨论组的 7327 个 UseNet 文章，里面包含着模拟赛车，模拟航空，真实汽车，真实航空四个主题的数据。20 Newsgroups 数据集收集了大约 20,000 左右的新闻组文档，每个顶级类别下都有 20 个子类别，每个子类别都有 1000 个样本。有一些新闻之间是相关的，比如 Sci.elec vs. Sci.med；有一些新闻是不相关的，比如 Sci.cryptvs.Alt.atheism。

由于上述两个数据集不是专门为 LLP 问题设置，需要将文本数据集重新组织为适用于标签比例学习问题的数据集。采用文献[15] [16]处理上述数据集的方法，我们根据数据集的顶级类别重新组织 LLP 数据集。首先，我们从顶级类别(A)中选择一个子类别 $a(1)$作为正类别，因此将该子类别 $a(1)$中的每个样本视为正类样本，其他顶级类别作为负类别。其次，对于原任务，我们从正子类别 $a(1)$中随机选择多个样本作为正样本，并从其他类别中随机选取相同数目的样本作为负样本，并将它们组成为原任务数据集。对目标任务执行相同的操作以形成正类样本和负类样本。为了使两个任务相关，我们让原任务和目标任

[1] http://www.iesl.cs.umass.edu/datasets.html。
[2] http://qwone.com/~jason/20Newsgroups/。

务的正类具有相同的顶级类别，例如原任务为 $a(1)$，则目标任务为 $a(2)$。在不损失有效性的情况下，我们仅保留具有较高文档频率的单词以减少维数，并且每个样本均由特征表示。最后随机选取同等比例的正样本和负样本生成 5 个数据集，如表 2 所示。

Table 2. The list of data sets
表 2. 数据集列表

编号	原任务	样本个数	目标任务	样本个数
1	Simauto	3000	Realauto	800
2	Auto	3000	Aviation	800
3	Sci-elec	5000	Sci-med	1500
4	Rec.autos	5000	Rec.motor	1500
5	Spo.baseball	5000	Spo.hockey	1500

4.2. 实验设置

为减少包中样本数量对实验结果的影响，本实验从每个数据集中分别随机选择 20，40 和 60 个样本组成包，并分别对大小不同的包依次进行实验，其中对四个算法的参数设置如下：

Inv-Cal：$C \in \left[2^{-2}, 2^{5}\right], \varepsilon \in [0.01, 0.1]$。

Alter-SVM：$C \in \left[2^{-2}, 2^{5}\right], C_p \in \left[2^{-2}, 2^{7}\right]$。

p-NPSVM：$C_i \in \left[2^{-5}, 2^{5}\right] (i = 1,2,3,4), C_p \in \{0.1, 1, 10\}$。

LPI-TL：$C_i \in \left[2^{-2}, 2^{7}\right] (i = 1,2), \varepsilon \in [0,1]$。

其中本文方法使用线性核函数 $K(x_1, x_2) = x_1 * x_2$，采用五折交叉验证法进行实验。由于 Inv-Cal，Alter-SVM 和 p-NPSVM 为单一任务算法，对其只在目标数据集上进行实验，对提出的算法则使用原任务数据集和目标任务数据集进行实验。

4.3. 实验结果分析

首先，利用本文提出的基于标签比例信息的迁移学习算法迁移原任务数据知识来对目标任务数据集进行实验，并采用准确率、精度、召回率和 F1 值等评价指标与 Inv-Cal，Alter-SVM 和 p-NPSVM 对比。

五个数据集具体的平均测试准确率和标准差实验结果如表 3 所示。表 3 表明，在数据集 3 包的大小为 20 以及数据集 4 包的大小为 40 和 60 中，本文提出算法比其他方法略低外，在其他实验结果中均能取得最高准确率和较低的标准差。图 3 展示了这四个算法的平均准确率的柱状图，其中四个算法的平均准确率分别为：62.44%、63.09%、63.38%和 67.77%，可见提出的算法 LPI-TL 的平均准确率高于 Inv-Cal，Alter-SVM 和 p-NPSVM 三个算法。这表明提出的方法在通过迁移原任务数据知识在分类准确率上能取得较高且稳定的性能。

另外，为了将其他算法与提出的 LPI-TL 方法进一步比较，将对上面实验得到的准确率做 Wilcoxon 符号秩和检验[17] [18]。通常，如果测试值 p 值低于置信度 0.05，则 LPI-TL 与所比较的方法之间存在显著差异。对于每种方法，其与 LPI-TL 之间的测试结果列在表 4 中。从表中可以看出，每种方法对 LPI-TL 的 p 值均小于置信度 0.05，这意味着在统计视角中，本文提出的算法 LPI-TL 比其他三个方法在准确率上取得显著性提高。

为了进一步验证本文提出算法的性能，对 5 个数据集在包为 20 的情况下分别计算准确率、召回率和 F1 值等指标的平均值。实验结果如表 5 所示。从表中可以看出，LPI-TL 算法平均召回率比 Alter-SVM 算

Table 3. Experimental accuracy and standard deviation Statistics
表 3. 实验准确率和标准差结果统计

数据集	算法	20	40	60
1	Inv-Cal	64.87 ± 1.60	60.42 ± 1.04	56.31 ± 2.00
	Alter-SVM	62.53 ± 1.31	60.13 ± 1.20	60.72 ± 1.45
	p-NPSVM	66.72 ± 0.47	63.23 ± 0.96	60.02 ± 1.52
	LPI-TL	**70.97 ± 0.37**	**68.60 ± 0.75**	**65.87 ± 1.22**
2	Inv-Cal	71.47 ± 0.70	68.37 ± 1.26	63.62 ± 1.06
	Alter-SVM	70.03 ± 0.95	66.47 ± 0.97	63.03 ± 0.98
	p-NPSVM	71.30 ± 1.02	65.50 ± 1.11	65.31 ± 1.00
	LPI-TL	**72.65 ± 0.41**	**72.57 ± 0.95**	**69.79 ± 0.82**
3	Inv-Cal	62.23 ± 0.73	59.52 ± 2.31	55.91 ± 1.79
	Alter-SVM	60.52 ± 1.28	58.68 ± 1.45	56.76 ± 1.21
	p-NPSVM	**63.22 ± 1.52**	62.07 ± 1.35	57.82 ± 0.48
	LPI-TL	63.00 ± 0.73	**64.32 ± 1.02**	**63.02 ± 0.75**
4	Inv-Cal	60.81 ± 1.47	62.02 ± 1.72	59.91 ± 1.79
	Alter-SVM	64.07 ± 1.08	**62.59 ± 0.93**	**60.06 ± 1.41**
	p-NPSVM	61.19 ± 1.11	58.65 ± 1.03	56.02 ± 0.88
	LPI-TL	**64.25 ± 0.92**	62.03 ± 0.87	59.82 ± 0.75
5	Inv-Cal	65.28 ± 1.32	63.56 ± 0.52	62.24 ± 1.24
	Alter-SVM	70.32 ± 0.96	65.27 ± 0.76	65.21 ± 0.91
	p-NPSVM	68.02 ± 1.23	65.42 ± 1.04	66.13 ± 1.08
	LPI-TL	**75.33 ± 0.62**	**73.56 ± 0.52**	**70.82 ± 0.53**

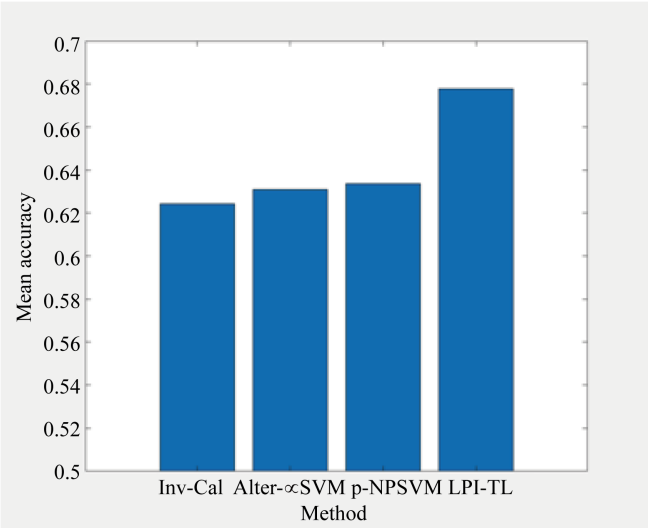

Figure 3. The mean accuracy
图 3. 平均准确率

Table 4. Wilcoxon signed ranks test.
表 4. Wilcoxon 符号秩和检验

LPI-TL	R+	R-	p-value
vs. Inv-Cal	118	2	0.0062
vs. Alter-SVM	115	5	0.0144
vs. p-NPSVM	119	1	0.0380

Table 5. Performance comparison of each algorithm
表 5. 各个算法性能对比

算法	平均精度	平均召回率	平均 F1 值
InvCal	0.501	0.648	0.565
Alter-SVM	0.462	0.672	0.548
p-NPSVM	0.499	0.662	0.569
LPI-TL	0.554	0.667	0.606

法低，这是由于 Alter-SVM 算法找出更多的负样本，所以召回率更高；相比之下，LPI-TL 算法取得更大的精度和 F1 值。总体来看，本文提出的 LPI-TL 算法的结果比对比方法更佳。

5. 结束语

本文对基于标签比例信息的迁移学习进行了研究，为了目标任务能更有效的学习预测模型，本文提出了一种迁移学习的模型用于从标签比例信息中学习分类器，该方法能将知识从原任务迁移到目标任务，并可以帮助目标任务构建分类器。本文实施了大量的实验去研究该方法的性能，实验表明该方法优于现有的 LLP 方法。该方法的不足之处在于仅能处理二分类问题还无法处理多分类数据，将来希望将该方法应用于多分类问题，这个问题值得后续进一步研究。

基金项目

国家自然科学基金资助项目(61876044)。

参考文献

[1] Kyuck, H. and de Freitas, N. (2005) Learning about Individuals from Group Statistics. In: *Proceedings of the* 21*st Conference on Uncertainty in Artificial Intelligence*, AUAI Press, New York, 332-339.

[2] Chen, Z., Shi, Y. and Qi, Z. (2019) Constrained Matrix Factorization for Semi-Weakly Learning with Label Proportions. *Pattern Recognition*, **91**, 13-24. https://doi.org/10.1016/j.patcog.2019.01.016

[3] Tan, B., Song, Y., Zhong, E. and Qiang, Y. (2015) Transitive Transfer Learning. *Proceedings of the* 21*st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2015, 1155-1164. https://doi.org/10.1145/2783258.2783295

[4] Pan, S.J. and Yang, Q. (2009) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[5] Hernández, J. and Inza, I. (2011) Learning Naive Bayes Models for Multiple-Instance Learning with Label Proportions. In: *Proceedings of Conference of the Spanish Association for Artificial Intelligence*, Springer, Berlin, Heidelberg, 134-144. https://doi.org/10.1007/978-3-642-25274-7_14

[6] Fan, K., Zhang, H., Yan, S., *et al*. (2014) Learning a Generative Classifier from Label Proportions. *Neurocomputing*, **139**, 47-55. https://doi.org/10.1016/j.neucom.2013.09.057

[7] Sun, T., Sheldon, D. and O'Connor, B. (2017) A Probabilistic Approach for Learning with Label Proportions Applied to the US Presidential Election. 2017 *IEEE International Conference on Data Mining* (*ICDM*), New Orleans, LA,

18-21 November 2017, 445-454. https://doi.org/10.1109/ICDM.2017.54

[8] Ardehaly, E.M. and Culotta, A. (2017) Mining the Demographics of Political Sentiment from Twitter Using Learning from Label Proportions. 2017 *IEEE International Conference on Data Mining* (*ICDM*), New Orleans, LA, 18-21 November 2017, 733-738. https://doi.org/10.1109/ICDM.2017.84

[9] Rueping, S. (2010) SVM Classifier Estimation from Group Probabilities. *Proceedings of the 27th International Conference on Machine Learning* (*ICML*-10), Haifa, Israel, 21-24 June 2010, 911-918.

[10] Yu, F.X., Liu, D., Kumar, S., *et al*. (2013) $\propto$ SVM for Learning with Label Proportions. arXiv Preprint arXiv:1306.0886.

[11] Wang, B., Chen, Z. and Qi, Z. (2015) Linear Twin SVM for Learning from Label Proportions. 2015 *IEEE*/*WIC*/*ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT), Singapore, 6-9 December 2015, 56-59. https://doi.org/10.1109/WI-IAT.2015.130

[12] Cui, L., Chen, Z., Meng, F. and Shi, Y. (2016) Laplacian SVM for Learning from Label Proportions. 2016 *IEEE* 16*th International Conference on Data Mining Workshops* (*ICDMW*), Barcelona, 12-15 December, 2016, 847-852. https://doi.org/10.1109/ICDMW.2016.0125

[13] Chen, Z., Qi, Z., Wang, B., *et al*. (2017) Learning with Label Proportions Based on Nonparallel Support Vector Machines. *Knowledge-Based Systems*, **119**, 126-141. https://doi.org/10.1016/j.knosys.2016.12.007

[14] Platt, J. (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, **10**, 61-74.

[15] Zhang, M.L. and Zhou, Z.H. (2008) M3MIML: A Maximum Margin Method for Multi-Instance Multi-Label Learning. 2008 *Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 15-19 December 2008, 688-697. https://doi.org/10.1109/ICDM.2008.27

[16] Liu, B., Xiao, Y. and Hao, Z. (2018) A Selective Multiple Instance Transfer Learning Method for Text Categorization Problems. *Knowledge-Based Systems*, **141**, 178-187. https://doi.org/10.1016/j.knosys.2017.11.019

[17] Demšar, J. (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1-30.

[18] Derrac, J., García, S., Molina, D., *et al*. (2011) A Practical Tutorial on the Use of Nonparametric Statistical Tests as a Methodology for Comparing Evolutionary and Swarm Intelligence Algorithms. *Swarm and Evolutionary Computation*, **1**, 3-18. https://doi.org/10.1016/j.swevo.2011.02.002