

基于分布和逆文本类别指数的特征迁移加权算法

邱云飞¹, 刘世兴^{1*}, 林明明¹, 邵良杉²

(1. 辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105; 2. 辽宁工程技术大学 系统工程研究所, 辽宁 葫芦岛 125105)

(* 通信作者电子邮箱 494784913@qq.com)

摘要:传统机器学习面临一个难题, 即当训练数据与测试数据不再服从相同分布时, 由训练集得到的分类器无法对测试集文本准确分类。针对该问题, 根据迁移学习原理, 在源领域和目标领域的交集特征中, 依据改进的特征分布相似度进行特征加权; 在非交集特征中, 引入语义近似度和新提出的逆文本类别指数(TF-ICF), 对特征在源领域内进行加权计算, 充分利用大量已标记的源领域数据和少量已标记的目标领域数据获得所需特征, 以便快速构建分类器。在文本数据集 20Newsgroups 和非文本数据集 UCI 中的实验结果表明, 基于分布和逆文本类别指数的特征迁移加权算法能够在保证精度的前提下对特征快速迁移并加权。

关键词:迁移学习; 特征分布; 逆文本类别指数; 语义近似度; 特征加权

中图分类号: TP391.1 **文献标志码:** A

Feature transfer weighting algorithm based on distribution and term frequency-inverse class frequency

QIU Yunfei¹, LIU Shixing^{1*}, LIN Mingming¹, SHAO Liangshan²

(1. School of Software, Liaoning Technical University, Huludao Liaoning 125105, China;

2. System Engineering Institute, Liaoning Technical University, Huludao Liaoning 125105, China)

Abstract: Traditional machine learning faces a problem: when the training data and test data no longer obey the same distribution, the classifier trained by training data can't classify test data accurately. To solve this problem, according to the transfer learning principle, the features were weighted according to the improved distribution similarity of source domain and target domain's intersection features. The semantic similarity and Term Frequency-Inverse Class Frequency (TF-ICF) were used to weight non-intersection features in source domain. Lots of labeled source domain data and a little labeled target domain were used to obtain the required features for building text classifier quickly. The experimental results on test dataset 20Newsgroups and non-text dataset UCI show that feature transfer weighting algorithm based on distribution and TF-ICF can transfer and weight features rapidly while guaranteeing precision.

Key words: transfer learning; feature distribution; Term Frequency-Inverse Class Frequency (TF-ICF); semantic similarity; feature weighting

0 引言

随着数据的快速更新, 当测试数据与训练数据的分布不再相同时, 训练得到的分类器难以对不同分布的测试文本准确分类, 传统机器学习中, 搜集语料—人工标注—训练分类器的模式逐渐跟不上信息的更新速度。对此, 研究人员希望在训练分类器时更多地利用已标记好的历史数据, 以便快速得到分类器。迁移学习可以充分利用先前学习任务的特性引起人们的重视。

关于迁移学习, 已有诸多相关算法及应用提出。在算法方面: 文献[1]提出一种基于实例迁移学习的文本分类算法, 该算法是对 Adaboost 算法的拓展, 利用其他相似源领域的标记数据集扩充目标领域数据集, 以生成精度较高的分类器; 文献[2]提出基于支持向量机的迁移学习(Transfer Learning-Support Vector Machine, TL-SVM)算法, 利用少量标记的目标

领域数据和大量源领域数据构建分类模型; 文献[3]提出基于单源及多源的迁移学习算法, 解决了 TrAdaBoost 算法中源领域与目标领域样本权值之间易出现两极分化的问题; 文献[4]提出基于相似度衡量的决策树自适应迁移算法, 以解决迁移学习中的负迁移问题; 文献[5]引入“属性”这一概念, 将先前的训练集结果迁移到具有相同“属性”的测试任务中; 文献[6]提出状态行为概率, 用这种新的衡量方式提高迁移学习算法的效率; 文献[7]利用共享子空间对跨域的迁移学习进行改进; 文献[8]利用特征的不同配制, 提出了半监督跨域学习算法, 以克服训练集中的错误标记和不可用的外域样本点。在应用方面: 文献[9-11]提出利用迁移学习原理来学习跨领域文本数据; 文献[12]用直推式迁移学习方法解决命名实体识别问题; 文献[13]将主动学习算法与迁移学习相结合, 以提高主动学习算法性能; 文献[14]将贝叶斯网络与迁移学习结合, 提高其鲁棒性; 文献[15]使用不充分的目标领

收稿日期: 2014-12-22; **修回日期:** 2015-03-17。 **基金项目:** 国家自然科学基金资助项目(70971059); 辽宁省创新团队项目(2009T045); 辽宁省高等学校杰出青年学者成长计划项目(LJQ2012027)。

作者简介: 邱云飞(1976-), 男, 辽宁阜新人, 教授, 博士, CCF 会员, 主要研究方向: 数据挖掘、情感分析; 刘世兴(1990-), 男, 辽宁丹东人, 硕士研究生, 主要研究方向: 数据挖掘、特征选择; 林明明(1989-), 女, 辽宁大连人, 硕士研究生, 主要研究方向: 数据挖掘、情感分析; 邵良杉(1961-), 男, 辽宁凌源人, 教授, 博士, 主要研究方向: 数据挖掘、情感分析。

域的数据和大量低质量的源领域的数据解决图像分类问题。综上所述,迁移学习已经广泛应用于文本数据挖掘^[16-18]、文本分类^[19-21]和命名实体识别^[22-24]等领域。目前迁移学习相关算法大多是基于 Adaboost 算法,对 n 个弱分类器进行加权联合,少有关于迁移学习的特征加权方法,而在较短时间内,获取高质量特征是训练分类器的关键。

本文根据源领域和目标领域特征的不同特性,针对交集特征的分布情况进行迁移加权,而非交集特征提出一种逆文本类别指数,结合语义近似度对源领域和目标领域的非交集特征进行迁移加权,从而在较短时间对特征完成选择和加权,适用于对实时性要求较高的文本分类任务。在文本数据集 20NewsGroups 和非文本数据集 UCI 中的实验表明,基于分布和逆文本类别指数的特征迁移加权算法能够在大量历史数据和少量实时数据的情况下,快速筛选出高质量特征,并在支持向量机(Support Vector Machine, SVM)中检验特征的分类效果,相比基于 Adaboost 的算法准确率更高、耗时更短。

1 特征迁移加权算法

用源领域数据结合目标领域数据进行特征迁移加权时,数据可分为两部分:一部分是新数据,即目标领域数据,这部分数据量少,但与测试数据分布相同;另一部分是历史数据,即源领域数据,这部分数据数量多,有类别标记,但与测试数据分布有差异。在对两种数据进行特征迁移加权时,存在六种情况:

- 1) 在源领域中出现概率较大,在目标领域中出现概率较小;
- 2) 在源领域中出现概率较小,在目标领域中出现概率较大;
- 3) 在源领域出现概率较大,在目标领域中出现概率较大;
- 4) 在源领域中出现概率较小,在新目标领域出现概率较小;
- 5) 在源领域中出现,在目标领域中未出现;
- 6) 在源领域中未出现,在目标领域出现。

在上述六种情况中,当充分利用源领域的“多”和目标领域的“准”时,能快速准确地对特征迁移加权。

本文使用的基本符号如下:

- 1) $Data_s$: 源领域数据,其类别已经标记。
- 2) $Data_t$: 目标领域数据,数量远少于源领域数据,已标记。
- 3) $Term_s$: $Data_s$ 中筛选出来的特征。
- 4) $Term_t$: $Data_t$ 中筛选出来的特征。

1.1 交集特征迁移加权算法

交集特征是指既出现在源领域,也出现在目标领域中的特征。它们要经过迁移加权才能作为分类模型的特征,其原因如下:

当情况 1) 作为分类特征时,因为源领域数据量多且出现概率大,导致整个分类器分类时向源领域偏倚,忽略了目标领域。由于源领域与测试数据分布有差异,得到的分类效果必然不理想。

当情况 2) 作为分类特征时,虽然特征在目标领域出现概率大,但由于这部分数据量很少,特征仅能代表部分测试数据特征,很可能出现对于有些测试类别不能识别的现象。

当情况 3) 作为分类特征时,要充分利用这部分特征,因为它们既包含了源领域的“广”,又包含了目标领域的“准”。

由于情况 4) 在两个领域的出现概率均较小,作为分类特征时,会出现稀疏数据,分类准确率降低,故不予考虑。

由上述分析可知,交集特征的迁移加权算法前提是特征

在源领域和目标领域出现概率都较大,其他三种会出现数据偏倚、类别表达不全和稀疏数据的情况,如果作为特征参与分类模型的训练会严重影响分类性能。

确定范围后,要对情况 3) 中的特征做迁移加权。加权依据为特征在源领域和目标领域分布相似或相同,这类特征既包含广泛的类别信息,弥补了目标领域所选特征可能存在的片面性,又能保证数据分布与测试数据保持一致。式(1)为本文提出经过改进的特征分布计算方法。

$$dist(C_i, t) = \frac{1}{realC_i / allC_i} \sqrt{\sum_{j=1}^n (x_i - \bar{x})^2 / n} \quad (1)$$

式(1)用于衡量特征在源领域和目标领域中每个类的分布情况。其中: x_i 为 C_i 类中第 i 篇文本所含特征 t 的数量; \bar{x} 为特征 t 在 C_i 类中所有文本出现的平均次数; $\sqrt{\sum_{j=1}^n (x_i - \bar{x})^2 / n}$ 即为特征 t 在 C_i 类中的标准差,虽然能表示特征在 C_i 类的分布特点,但仍需要分母 $realC_i / allC_i$ 作限制; $realC_i$ 为 C_i 类中特征数大于阈值 ε ($\varepsilon \geq 0$) 的文本数; $allC_i$ 为 C_i 类内所有文本数。对为何加分母的限制给出如下说明:

假设有一个类 C , 包含五篇文本,特征 t 在源领域的 C 类中分布为 $d_{source} = \{x, 0, 0, 0, 0\}$ ($x > 0$), 在目标领域的 C 类中分布为 $d_{target} = \{x + y, y, y, y, y\}$ ($y > 0$)。虽然所得方差均为 $2x/\sqrt{5}$, 但从特征在 5 篇文本的分布来看明显不同,可见分子的标准差只表现了特征在类内的波动状况。若加入分母部分,当 ε 为 0 时,得到的分布结果为 $10x/\sqrt{5}$ 和 $2x/\sqrt{5}$, 表明在 C 中,特征 t 在源领域和目标领域的分布不同,不能给其赋过高的权值。

得到交集特征在源领域和目标领域中每个类的分布后,利用式(2)对特征在源领域和目标领域的分布相似度对其进行加权。

$$w(t) = \sum_{i=1}^n \frac{1}{|(dist_s(C_i, t) / dist_t(C_i, t)) - 1| + \delta} \quad (2)$$

式(2)为交集特征的加权计算方法。其中: $dist_s(C_i, t)$ 为特征 t 在源领域的分布; $dist_t(C_i, t)$ 为特征 t 在目标领域的分布; δ 为数据平滑处理,防止分母出现 0, 取值大于 0 小于 1。若特征 t 在源领域和目标领域分布越近似,则分母越趋近于 0, 特征 t 的权值越大; 否则 t 的权值将越小。交集特征迁移加权具体算法如下:

算法 1 交集特征迁移加权具体算法。

输入 源领域文本的特征集合 $Term_s = \{t_1, t_2, \dots, t_m\}$, 目标领域的特征集合 $Term_t = \{t_1, t_2, \dots, t_l\}$, 类别个数 n , 特征频率筛选阈值 δ 。

输出 交集特征迁移加权后的权值集合 $weight = \{weight_1, weight_2, \dots, weight_n\}$ 。

begin

$Term_s' = \{t_i \mid t_i.freq > \delta, t_i \in Term_s\}$;

$Term_t' = \{t_i \mid t_i.freq > \delta, t_i \in Term_t\}$;

$Term_{Intersection} = Term_s' \cap Term_t'$;

for i : 1 to n

for j : 1 to $Term_{Intersection}.size$

$weight = weight + weight(t_j)$;

end for

```

end for
return weight;
end

```

算法中 $Term_s'$ 和 $Term_t'$ 为经过特征频率(Text Frequency, TF) 值筛选后留下的特征, $Term_{Intersection}$ 为要进行迁移加权的交集特征。

1.2 非交集特征迁移加权算法

在特征迁移加权的過程中, 经过迁移加权的交集特征具有与测试集数据分布相同的优点, 但由于数据量少, 包含的类别信息有限, 导致特征难以准确表示各类文本, 这时就要考虑非交集特征, 即在源领域中出现, 在目标领域中未出现和在源领域中未出现, 在目标领域中出现的情况。利用源领域大量已标记的数据, 弥补交集特征类别相关信息不足的缺陷。结合语义近似度和逆文本类别指数, 提出非交集特征情况下的迁移特征加权算法。假设有源领域数据集 $Data_s$ 和目标领域数据集 $Data_t$ 。性质 1 为逆文本类别指数计算公式的前提条件, 下面给出性质 1 及其证明。

性质 1 迁移特征加权的权值正比于特征的出现频率, 反比于与特征高相关的类别数占有类别的比例。

证明 若特征 t 在文本中出现频率很小, 则它仅能对极少量文本进行表示, 导致对文本进行量化表示时出现稀疏数据, 分类准确率降低; 若特征 t 出现频率很高, 则表明它在文本分类中含有的信息较多。故权值与特征频率成正比, 性质 1 前半部分得证。若特征 t 与每个类都高度相关, 则说明该特征无法对类别进行区分, 在分类时起不到任何作用, 与特征高相关的类别数占有类别比例为 1 时, 说明该特征对分类没有任何意义; 相反, 若特征 t 仅与个别类高度相关, 占有类别数的比例较低, 则特征对分类帮助较大, 具有较高权值。故权值反比于与特征高相关类别占有类别的比例, 性质得证。

根据性质 1, 本文提出式 (3), 即逆文本类别指数计算公式, 利用该公式对源领域和目标领域的非交集特征进行迁移加权。

$$weight(term) = -tf * \lg\left(\frac{|class'| + \theta}{|class|}\right) \quad (3)$$

其中: tf 为特征频率, $|class|$ 为源领域中所有类别数量, $|class'|$ 为特征与类别相关度大于阈值 ν 的类别数, 分子 θ 为数据平滑处理。

非交集特征迁移加重的具体过程如下:

1) 根据特征频率(TF) 值从源领域和目标领域筛选特征集 $Term_s$ 和 $Term_t$ 。

2) 利用语义让每个 $Term_t$ 中的特征与 $Term_s$ 中的特征作语义近似度计算 $Sim(Term_t, Term_s)$, 保留 $Term_s$ 中 $Sim(Term_t, Term_s)$ 大于阈值 γ 小于 1 的特征。

3) 用互信息计算公式:

$$MI(class, t) = p(class, t) \lg \frac{p(class, t)}{p(class)p(t)} \quad (4)$$

计算出每个特征 t 与源领域中各个类别的相关性。

4) 根据性质 1 提出的逆文本类别指数, 设定相关性过滤阈值 τ , 对 $Term_s$ 中的特征加权。

非交集特征迁移加权算法具体如下:

算法 2 非交集特征迁移加权算法。

输入 源领域文本 $Data_s$, 目标领域文本 $Data_t$ 。

输出 非交集特征迁移加权后的权值集合 $weight = \{weight_1, weight_2, \dots, weight_n\}$ 。

begin

通过 tf 值从 $Data_s$ 和 $Data_t$ 中过滤低频特征, 得到特征集合 $Term_s$ 和 $Term_t$;

通过语义相似度计得到特征集 $Term_s', Term_t' = \{t \mid 1 > Sim(Term_s, Term_t) > \gamma\}$;

for $i: 1$ to $Term_s'$. size

for $j: 1$ to $class$. size

if $MI(Term_s'(i), class) > \delta$
| $class'|++$;

end for

$weight(Term_s) = -tf * \lg\left(\frac{|class'| + \theta}{|class|}\right)$;

end for

return $weight$;

end

1.3 基于分布和逆文本类别指数的特征迁移加权算法

本文根据特征的分布情况和逆文本类别指数, 分别对源领域和目标领域的交集特征和非交集特征作筛选加权。通过结合上述两种算法, 提出基于分布和逆文本类别指数的特征迁移加权算法。该算法分别在交集特征和非交集特征作迁移加权, 充分利用标记好的源领域, 节省了标记新文本类别耗费的时间, 适用于对实时性要求较高的文本分类任务。

基于分布和逆文本类别指数的特征迁移加权算法具体如下:

算法 3 基于分布和逆文本类别指数的特征迁移加权算法。

输入 大量已标记源领域文本 $Data_s$, 少量已标记目标领域文本 $Data_t$ 。

输出 源领域与目标领域的交集和非交集特征加权后的权值集合 $weight$ 。

1) 交集特征迁移加权算法得到 $weight_1$;

2) 非交集特征迁移加权算法得到 $weight_2$;

3) $weight = weight_1 \cup weight_2$;

4) return $weight$ 。

2 实验结果与分析

本文实验采用的数据集分为两类: 1) 非文本数据集 UCI; 2) 文本数据集 20Newsgroups。实验使用支持向量机(Support Vector Machine, SVM)对迁移加权后的特征分类性能进行评价, 评价指标包括准确率、召回率和 F 值^[25]。

为验证本文算法的准确性和可行性, 在 UCI 数据集中, 对比本文算法与其他算法在分类精度、耗时以及不同领域相似度对分类性能的影响; 在 20Newsgroups 数据集中, 对比本文算法与其他算法在不同数量的目标领域特征下的分类性能。

2.1 UCI 数据集的实验

2.1.1 分类精度与耗时对比

采用 UCI 机器学习库中的 16 个数据集进行实验, 数据集分属 4 个不同领域, 分别为生命科学: Iris、Mushroom、Stalog (Heart)、Ecoli、Acute Inflammation、Haberman's survival、Mammographic mass 和 SPECT heart; 社会科学: Balance scale、Nursery、Hayes-roth、Teaching assistant evaluation、Car evaluation 和 MONK's problem; 竞技体育: Chess (KR vs KP); 物质结构: Wine。其中, 属性跨度: 3 ~ 36, 类别跨度: 2 ~ 8, 样本跨度: 120 ~ 12960。

实验在 UCI 数据集中, 将本文方法与 TrAdaBoost 算法、TL-

SVM 算法、多源动态 TrAdaboost (Multi-Source dynamic TrAdaboost,MSDT)算法和相似度衡量的决策树自适应迁移 (Self-adaptive Transfer for Decision Tree,STDt)算法在分类精度和效率上进行比较,选用 SVM 作为 TrAdaboost 的基准分类器。表 1 为五种算法在 UCI 数据集中的分类精度和分类耗时。

由表 1 可以看出,五种算法对于分类精度而言,以本文算法和 STDt 算法为优;而对于复杂度,从表 1 的分类耗时上来看,本文算法略优于 STDt 及其他三种算法。本文算法复杂度与类别数和非交集特征数相关;而对比的四种算法的复杂度不仅依赖于源领域特征数,还与其最大迭代次数相关。

表 1 UCI 数据集分类性能对比

UCI 数据集	TrAdaboost		TL-SVM		STDt		MSDT		本文算法	
	精度	耗时/s	精度	耗时/s	精度	耗时/s	精度	耗时/s	精度	耗时/s
Iris	0.710	0.016	0.794	0.016	0.855	0.015	0.732	0.016	0.871	0.014
Balance scale	0.682	0.009	0.773	0.008	0.803	0.007	0.715	0.008	0.811	0.006
Car evaluation	0.718	0.021	0.695	0.022	0.778	0.021	0.701	0.024	0.792	0.019
Chess	0.753	0.390	0.838	0.349	0.925	0.345	0.793	0.431	0.931	0.301
Mushroom	0.700	0.259	0.691	0.287	0.712	0.285	0.683	0.301	0.745	0.253
Nursery	0.726	0.120	0.817	0.106	0.892	0.107	0.786	0.130	0.905	0.113
Statlog	0.753	0.090	0.738	0.098	0.784	0.090	0.742	0.091	0.796	0.085
Wine	0.813	0.384	0.735	0.400	0.813	0.397	0.729	0.387	0.857	0.329
Hayes-roth	0.672	0.023	0.673	0.019	0.708	0.018	0.697	0.019	0.734	0.020
SPECT heart	0.694	0.039	0.700	0.045	0.759	0.046	0.718	0.047	0.770	0.038
TAE	0.858	0.019	0.856	0.020	0.926	0.020	0.878	0.021	0.941	0.018

2.1.2 领域相似度对分类性能的影响

从 UCI 数据集中筛选出两部分数据,一部分领域相似度较高,另一部分领域相似度较低。每部分数据中均包含 UCI 的 16 个子类,将每个子类的数据集分为两组,数量多的作为源领域特征,数量少的作为目标领域特征。通过上述两部分数据对本文方法、TrAdaboost 算法、TL-SVM 算法、MSDT 算法和 STDt 算法在不同领域相似度下的分类性能进行对比。

图 1(a)~(c)是利用 UCI 数据集在源领域与目标领域相似度高时的准确率、召回率和 F 值对比。从图中可以看到,

在领域相似度较高时,本文方法除在 Car、Wine 和 Hayes 数据集中准确率略高于 STDt 算法,在其他数据集里,本文方法均明显好于 STDt 算法。

图 2(a)~(c)是利用 UCI 数据集,在源领域与目标领域相似度较低时的准确率、召回率和 F 值对比。从图中可以看到,当领域相关度较低时,本文方法在 Iris、Balance scale、Wine、Hayes-roth、Mushroom、Acute inflammations 和 Haberman's survival 数据集中的准确率比 STDt 算法提高 5% 以上,在其他数据集中有少量提高或持平。

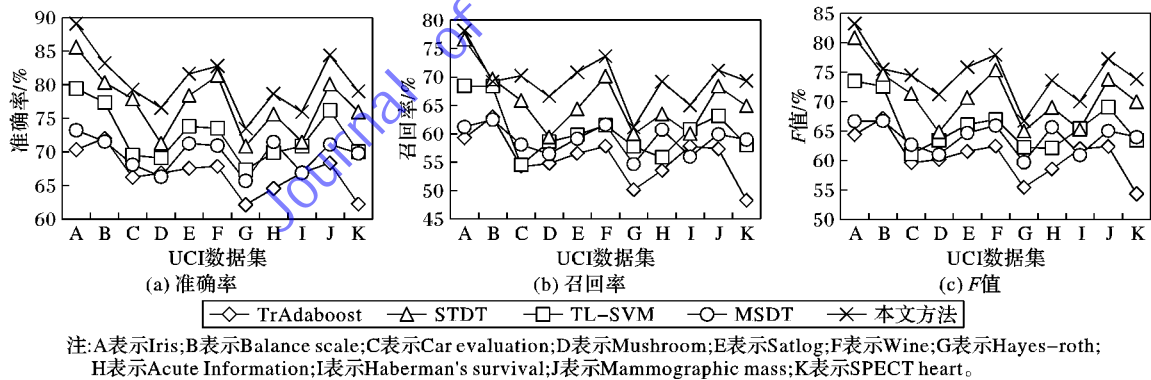


图 1 高领域相关度下的分类性能对比

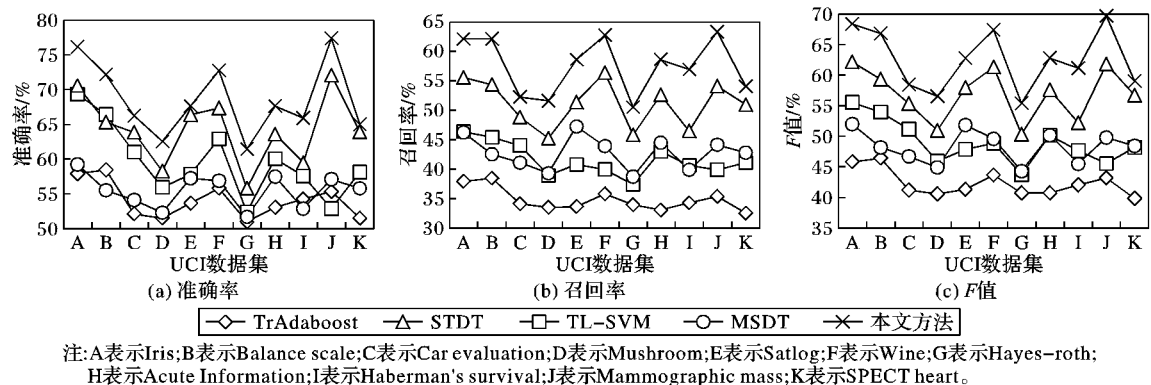


图 2 低领域相关度下的分类性能对比

综上所述,在源领域和目标领域相似度较高的前提下,STDT 算法和本文方法均有较好的迁移学习能力,在 UCI 中的平均分类准确率分别为 81.12% 和 83.53%;而在相似度较低的情况下,本文方法和 STDT 算法的平均准确率分别为 71.84% 和 67.68%。因此,在源领域与目标领域相似度较高时,本文方法和 STDT 算法均有较好的迁移能力;而在领域相似度较低时,即源领域与目标领域的非交集特征较多时,本文方法通过特征迁移加权从源领域和目标领域中挖掘出更多对目标领域分类有用的特征,在支持向量机中得到获得更好的分类效果。

2.2 20Newsgroups 文本集的实验

20Newsgroups 是由 7 大类 20 小类组成的报文数据集,本文选用其中 comp、rec、sci 和 talk 四大类进行实验,将大类中子类分别作为源领域和目标领域本文,对五种算法在不同数量的目标领域数据集下的分类性能进行实验,实验结果如图

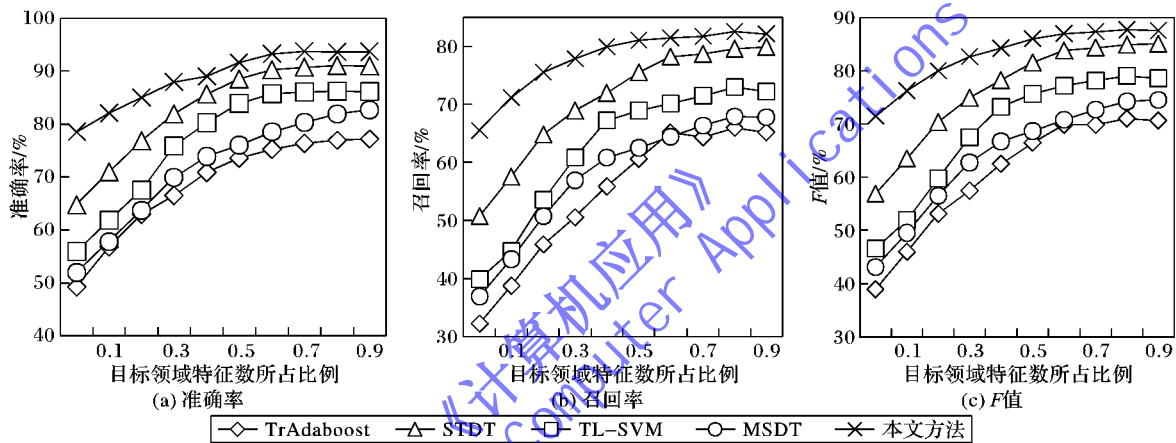


图3 目标领域特征数量变化对分类性能的影响

3 结语

本文提出一种基于分布和逆文本类别指数的特征迁移加权算法,针对源领域和目标领域的交集特征在各自领域的分布,提出交集特征的迁移加权算法。针对非交集特征,通过语义近似度将源领域和目标领域联系起来,根据本文提出的逆文本类别指数对非交集特征进行迁移加权。为验证有效性,对本文算法及其他四种迁移算法在分类准确度和耗时进行了比较;同时,在不同领域相似度和不同目标领域样本数情况下,对五种算法的分类性能作了对比。由实验结果可以看到本文算法与其他迁移算法相比有以下特点:1)本文算法复杂度仅与源领域特征数相关,而基于 Adaboost 的迁移学习分类算法复杂度不仅与特征数相关,还要进行多次遍历才能找到最优分类器,复杂度较高,耗时较多;2)本文提出的逆文本类别指数使分类准确度有所提升;3)由于引入了语义近似度,本文算法适用于源领域与目标领域相关性较低和目标领域数据量不足的情况,相比 STDT 算法,适用范围更广。

参考文献:

[1] LIU X. Text classification research based on instance transfer learning [D]. Changchun: Jilin University, 2014. (刘晓明. 基于实例迁移学习的文本分类研究[D]. 长春: 吉林大学, 2014.)
 [2] XU M, WANG S, GU X. TL-SVM: A transfer learning algorithm

3 所示。

由图 3(a) ~ (c) 可知,随着目标领域数据集数量的增加,五种算法的分类性能呈上升趋势。本文方法在比例达到 60% 时,分类性能趋于稳定;STDT 和 TL-SVM 算法在比例达到 70% 时性能趋于稳定;TrAdaboost 和 MSDT 算法在比例达到 80% 时性能趋于稳定。

当目标领域样本数量较少时,本文的特征迁移加权算法在分类性能上效果显著。原因在于本文方法引入了语义近似度,即便拥有少量的目标领域数据,也能从源领域中挖掘出更多对新文本分类有用的信息,供目标领域特征参考,并进行加权。其他算法要想具有较好分类性能,都需要有足够的目标领域数据进行支持,而获得足够多的目标领域训练样本代价较高,标记目标领域数据需要耗费大量的时间,对于实时性要求较高的文本分类任务并不适用。

[J]. Control and Decision, 2014, 29(1): 141-146. (许敏,王士同,顾鑫. TL-SVM: 一种迁移学习算法[J]. 控制与决策, 2014, 29(1): 141-146.)
 [3] GUO Y. Research of transfer learning based on single-source and mutli-source [D]. Xi'an: Xidian University, 2013. (郭勇. 基于单源及多源的迁移学习方法研究[D]. 西安: 西安电子科技大学, 2013.)
 [4] WANG X, PAN J, CHENG Y, et al. Self-adaptive transfer for decision trees based on similarity metric [J]. Acta Automatic Sinica, 2013, 39(12): 2186-2192. (王雪松,潘杰,程玉虎,等. 基于相似度衡量的决策树自适应迁移[J]. 自动化学报, 2013, 39(12): 2186-2192.)
 [5] KALE D, YAN L. Transfer learning based on the observation probability of each attribute [C]// Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2014: 3627-3631.
 [6] WANG H, FAN S, SONG J, et al. Reinforcement learning transfer based on subgoal discovery and subtask similarity [J]. IEEE/CAA Journal of Automatica Sinica, 2014, 1(3): 257-266.
 [7] FANG Z, ZHANG Z. Cross domain shared subspace learning for unsupervised transfer classification [C]// Proceedings of the 2014 22nd International Conference on Pattern Recognition. Stockholm: ICPR, 2014: 3927-3932.
 [8] ZHU X. Cross-domain semi-supervised learning using feature formu-

- lation [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012; 6(41): 1627–1638.
- [9] RAINA R, NG A Y, KOLLER D. Constructing informative priors using transfer learning [C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 713–720.
- [10] DAI W, XUE G, YANG Q, *et al.* Transferring native Bayes classifiers for text classification [C]// Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2007: 540–545.
- [11] DAI W, XUE G, YANG Q, *et al.* Co-clustering based classification for out-of-domain documents [C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 210–219.
- [12] ARNOLD A, NALLAPATI R, COHEN W. A comparative study of methods for transductive transfer learning [C]// Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Washington, DC: IEEE Computer Society, 2007: 77–82.
- [13] KALE D, LIU Y. Accelerating active learning with transfer learning [C]// Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2013: 1085–1090.
- [14] OYEN D, LANE T. Bayesian discovery of multiple Bayesian networks via transfer learning [C]// Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2013: 577–586.
- [15] WUP, DIETTERIEH G. Improving svm accuracy by training on auxiliary data sources [C]// Proceedings of the 21st International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2004: 871–878.
- [16] ANDO S, SUZUKI E. Unsupervised cross-domain learning by interaction information co-clustering [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Piscataway: IEEE, 2008: 13–22.
- [17] KAMISHIMA T, HAMASAKI M, AKAHO S. TrBagg: A simple transfer learning method and its application to personalization in collaborative tagging [C]// Proceedings of the 9th IEEE International Conference on Data Mining. Piscataway: IEEE, 2009: 219–228.
- [18] ZHANG Y, YEUNG D. Transfer metric learning by learning task relationships [C]// Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 1199–1208.
- [19] CHEN B, LAM W, TSANG I, *et al.* Extracting discriminative concepts for domain adaptation in text mining [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 179–188.
- [20] WANG P, DOMENICONI C, HU J. Using Wikipedia for co-clustering based cross-domain text classification [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Piscataway: IEEE, 2008: 1085–1090.
- [21] GAO J, FAN W, JIANG J, *et al.* Knowledge transfer via multiple model local structure mapping [C]// Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 283–291.
- [22] ARNOLD A, NALLAPATI R, COHEN W. Exploiting feature hierarchy for transfer learning in named entity recognition [C]// Proceedings of the 2008 Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2008: 245–253.
- [23] RICHMAN A, SCHONE P. Mining wiki resources for multilingual named entity recognition [C]// Proceedings of the 2008 Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2008: 1–9.
- [24] GOLDWASSER D, ROTH D. Active sample selection for named entity transliteration [C]// Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2008: 53–56.
- [25] PING Y. Research on clustering and text categorization based on support vector machine [D]. Beijing: Beijing University of Posts and Telecommunications, 2012. (平源. 基于支持向量机的聚类及文本分类研究[D]. 北京: 北京邮电大学, 2012.)

(上接第1632页)

- [15] ZHU J, LIU S, LI H, *et al.* Aggregation approach of multiple stages multiple judgment preferences styles in group decision making [J]. Control and Decision, 2008, 23(7): 730–734. (朱建军, 刘思峰, 李洪伟, 等. 群决策中多阶段多元判断偏好的集结方法研究[J]. 控制与决策, 2008, 23(7): 730–734.)
- [16] ZHANG F, GUO Y, YI P. Multi-phase interactive group evaluation method based on rank correlation analysis [J]. Journal of Systems Engineering, 2011, 26(5): 702–709. (张发明, 郭亚军, 易平涛. 序关系分析下的多阶段交互式群体评价方法[J]. 系统工程学报, 2011, 26(5): 702–709.)
- [17] LAI Y J, LIU T Y, HWANG C L. TOPSIS for MODM [J]. European Journal of Operational Research, 1994, 76(3): 486–500.
- [18] YAGER R R. On ordered weighted averaging aggregation operators in multi-criteria decision making [J]. IEEE Transactions on Systems, Man and Cybernetics, 1988, 18(1): 183–190.
- [19] WANG Y-M, LUO Y, HUA Z. Aggregating preference rankings using OWA operator weights [J]. Information Sciences, 2007, 177(16): 3356–3363.
- [20] WANG Y-M, PARKAN C. A minimax disparity approach for obtaining OWA operator weights [J]. Information Sciences, 2005, 175(1/2): 20–29.
- [21] HAO J, ZHU J, LIU S. Aggregation of multi-stage uncertain linguistic evaluation information based on Orness [J]. System Engineering – Theory and Practice, 2013, 33(11): 2866–2873. (郝晶晶, 朱建军, 刘思峰. 基于 Orness 测度的多阶段不确定语言信息优化集结[J]. 系统工程理论与实践, 2013, 33(11): 2866–2873.)
- [22] LI D. Fuzzy multiobjective many-person decision making and game [M]. Beijing: National Defense Industry Press, 2003: 75–77. (李登峰. 模糊多目标多人决策与对策[M]. 北京: 国防工业出版社, 2003: 75–77)
- [23] WANG Y-M, PARKAN C. A preemptive goal programming method for aggregating OWA operator weights in group decision making [J]. Information Sciences, 2007, 177(8): 1867–1877.