计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140156
52(3):629-638,2015

基于分组提升集成的跨领域文本情感分类

赵传君[1] 王素格[1,2] 李德玉[1,2] 李 欣[1]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(wsg@sxu.edu.cn)

Cross-Domain Text Sentiment Classification Based on Grouping-AdaBoost Ensemble

Zhao Chuanjun[1], Wang Suge[1,2], Li Deyu[1,2], and Li Xin[1]
[1](*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)

Abstract In the cross-domain sentiment classification, the labeled data in the target domain is often scarce and precious. To solve this problem, this paper proposes a grouping-AdaBoost ensemble classifier method by comprehensively using the strategies and techniques of semi-supervised learning, Bootstrapping, data grouping, AdaBoost, ensemble learning. Firstly, we adopt a small amount of labeled data in the target domain to generate a number of virtual data by using synthetic minority over-sampling technique. On this basis, we can obtain more data with high credibility label in the target domain by using Bootstrapping method. In the aspect of classifier construction, we firstly make an equivalent quantity partition to the labeled data in the source domain, and combine each part with the labeled data in the target domain to form the corresponding combined data sets. Corresponding to each combined data set, a classifier is trained, and it is then promoted by AdaBoost method. At last, these classifiers corresponding to the combined data sets are linearly integrated into an ensemble classifier. The experimental results on four data sets from Amazon online shopping reviews corpora indicate that the proposed method can improve the accuracy of cross-domain sentiment transformation effectively.

Key words sentiment classification; cross-domain; synthetic minority over-sampling technique; grouping-AdaBoost; ensemble classifier

摘 要 针对目标领域带标签数据偏少的问题,综合运用半监督学习、BootStrapping、数据分组、AdaBoost、集成学习等策略与技术,提出了一种基于分组提升集成的跨领域文本情感分类方法.该方法首先利用少量人工标注的目标领域数据,基于合成过抽样技术产生一定数量的虚拟数据.在此基础上,采用 BootStrapping 方法获得更多目标领域高可信度的带标签数据.在分类器的构建方面,首先将源领域的带标签数据等量分割,并分别与目标领域带标签数据组合,在每个组合数据块上运用 AdaBoost 方法提升地训练多个分类器,并将这些分类器线性地集成为一个分类器.在亚马逊购物网站 4 个领域的情感数据集上的实验表明,基于分组提升集成的跨领域文本情感分类方法一定程度上提高了跨领域文本情感分类的精度.

收稿日期:2014-02-28;**修回日期**:2014-09-16
基金项目:国家自然科学基金项目(61175067,61272095,61405109);国家"八六三"高技术研究发展计划基金项目(2015AA015407);山西省回国留学人员科研项目(2013-014);山西省自然科学基金项目(2013011066-4);山西省科技攻关项目(20110321027-02)
通信作者:王素格(wsg@sxu.edu.cn)

关键词 情感分类;跨领域;合成过抽样技术;分组提升;集成分类器

中图法分类号 TP391

Web2.0 技术使 Internet 由一个静态信息媒介变成了一个动态信息交流平台.越来越多的用户乐意针对公共事件借助微博、论坛等社交媒体发表自己的言论,表达自己的观点和立场;众多消费者倾向于在亚马逊(Amazon)等购物网站上表达对某些产品或服务的使用感受[1].如何高效地分析、处理和利用广泛分布于网上的海量观点信息逐渐受到自然语言处理、情感分析、电子商务等领域专家学者的高度关注[2].

文本情感倾向分类是一类重要的情感分析问题.通常情况下,基于对文本内容的分析,情感倾向分类自动地将评论文本分为正面(支持)和负面(反对)2 类[3-4],大多已有的结果都假定训练集和测试集源于同一领域[5-6].在实际应用中,现有的方法至少在以下 2 个方面面临挑战:1)网上的评论文本往往涉及多个领域,例如,新闻评论一般涉及到政治、经济、体育、娱乐等领域;产品评论可能涉及到汽车、电子产品、书籍、股票等领域.众所周知,有监督机器学习理论是建立在所谓的"独立同分布"先验假设之上的.当一个领域中训练的情感分类器被直接用于另一个领域时,由于源领域和目标领域的特征分布不同,会导致分类器的分类精度下降,甚至失效[7-8].2)训练分类器需要大量已标注数据,这项工作耗时费力,这也正是目前半监督机器学习兴起的一个重要原因.因此,利用已有的领域标注数据辅助其他领域的文本情感倾向性分析任务(领域情感移植)不仅可以减少新领域数据标注工作量,而且可以提高源领域标注资源的利用率.目前,领域情感迁移已成为自然语言处理和机器学习领域中值得关注的研究课题[9].

关于跨领域文本情感倾向性分析,已有的成果主要从实例迁移和特征迁移 2 个角度开展研究[9].就实例迁移而言,在源领域和目标领域数据分布差异明显的情况下,需要解决 2 个问题:1)在目标领域有少量带标签数据的条件下,如何从目标领域无标签数据中选取更多的对目标领域分类有作用的数据;2)在带标签的源领域数据比较多而目标领域较少的情况下,如何组织源领域和目标领域带标签的数据,使它们共同为目标领域的情感分类提供有价值的信息[8].

为了解决上述 2 个问题,本文提出了一种基于半监督的 2 阶段集成分类器的跨领域文本情感分类方法.首先,采用合成过抽样技术(synthetic minority over-sampling technique, SMOTE)[10]对目标领域少量的人工标注数据进行随机线性扩充;设计了目标领域数据初始标签算法(initial data labeling algorithm, IDLA),它利用经合成过抽样技术扩充后的标注数据集,迭代地获得一定规模的目标领域中具有较高标签可信度的训练数据.在此基础上,利用数据分组、AdaBoost 和集成学习等策略与技术,提出一种基于数据分组提升集成的学习算法(grouping AdaBoost ensemble learning, GAdaBEL),它通过融合源领域和目标领域中带有标签的数据,提升目标领域数据的标注效果,实现跨领域文本情感分类.在亚马逊购物网站上的 4 个跨领域的英文语料上,和已有相关方法进行的比较实验表明,本文提出的方法在一定程度上提高了跨领域情感分类的正确率.

1 相关研究

如引言中所述,关于跨领域文本情感分类,已有的工作主要从特征迁移和实例迁移 2 个方面开展研究.

在特征迁移方面,研究的主要思路是寻找源领域和目标领域之间的关联特征(共享特征),旨在构建跨领域数据的统一特征表示空间[11-13].Blitzer 等人[11]提出了一种自称为结构一致学习(structural correspondence learning, SCL)的领域适应性学习方法,该方法首先选取同时与源领域和目标领域经常共现的若干个枢纽特征,然后计算每个枢纽特征与 2 个领域中其他非枢纽特征的相关性,构建相关关系矩阵,再将奇异值分解用于求解相关关系矩阵的低维线性近似,最终的特征集由最初的特征集和新抽取的实值特征所组成.Pan 等人[12]通过使用谱特征对齐算法(spectral feature alignment, SFA)构建不同领域间的连接关系.该方法中,这种连接关系体现为领域独立词和领域关联词间的带权二部图,通过词与词的共现信息获得二部图中的边,边的权重为 2 个词的共现频次.作者采用谱聚类算法,使关联度高的领域独立词(领域关联词)聚在一个簇下,最终的特征集由原始特征集和生成的簇所构成.Xia 等人[13]在跨领域情感分类问题研究中发现,某些特

定词性的特征具有领域依赖性,而其他词性的特征则与领域无关.鉴于此,他们提出了一种基于词性的集成分类模型用于整合不同词性的特征以提高分类的效果.

在实例迁移方面,研究的主要思路是从源领域已标记数据中选取那些对目标领域分类有价值的实例[14-23],用于辅助目标领域文本情感分类.Tan 等人[14]用源领域的样本训练分类器,对目标领域的未标记样本进行分类,再选择目标领域中高可信度的样本重新训练分类器.Tan 等人[15]提出一种基于随机游走模型的跨领域倾向性分析方法,该方法利用源领域和目标领域的文本与词之间的关联关系,利用迭代的方法来计算目标领域中未标记样本的情感分,并以情感分值来判断其情感倾向性.吴琼等人[16]提出了一种基于图排序的情感倾向性分类算法,该算法在图排序的基础上,利用源领域的标签和目标领域的伪标签进行迭代,将目标领域中标签较为准确的样本作为种子,通过最大期望算法(expectation maximization algorithm, EM)进行迭代以实现跨领域的倾向性分类.吴琼等人[17-18]提出了一种面向跨领域情感分类的多阶段框架,该方法首先利用源领域的带标签文本预测目标领域文本的初始标签,然后在目标领域中构建一个加权网络,把目标领域中高可信度文本作为热传导中的源点和汇点,利用物理学中热传导的思想不断进行迭代,最终实现跨领域的情感分类.Liao 等人[19]提出了 Migratory-Logit 方法,通过引入一个阈值评估源领域数据中每一个样本对目标领域样本分类的贡献程度,在此基础上,结合主动学习方法挑选出最优的一部分源领域数据构建适合目标领域数据分类的分类器.Jiang 等人[20]从减小领域分布差异的角度,提出了一种实例权重框架来解决领域适应性问题.首先利用目标领域数据的信息对源领域数据进行评估,将源领域数据中被认为具有"误导"作用的样例剔除.在训练分类器时,赋予目标领域带标签样本较高的权重,源领域带标签样本较低的权重,用得到的预测标签来选择部分高可信度的目标领域的实例加入到训练集中.Anthony 等人[21]提出了 4 种解决领域适应性的方法:1)从多个源领域中选取有效数据形成混合数据集作为训练集,训练分类器来对目标领域数据集进行分类;2)训练一个同 1)中情形的分类器,但特征限定从目标领域中选取;3)从多个源领域中选取有效的数据训练各自的分类器组成 1 个联合分类器,用集成学习的方法对目标领域数据集进行分类;4)把少量的源领域带标签数据和目标领域大量未带标签数据进行混合,在混合数据集上运用 EM 算法来学习 1 个朴素贝叶斯分类器.Xia 等人[22]在实例选择和实例权重的基础上通过 PU 学习(positive and unlabeled learning, PU)提出了解决跨领域情感分类问题一种策略.首先通过 PU 学习得到训练集中每个样本的目标领域概率,在此基础上提出了 2 个模型.实例选择模型(positive and unlabeled instance selection, PUIS)通过选择训练集中的高概率样本作为训练数据集.实例权重确定模型(positive and unlabeled instance weighting, PUIW)首先调整训练集中样本的目标领域概率到合适的水准,在最大权重似然估计的基础上,用校准后的权重作为样本权重来训练朴素贝叶斯模型.Li 等人[23]首先采取主动学习的策略选取目标领域的少量带标签数据,从源领域和目标领域中的带标签数据训练 2 个独立的分类器,然后采用委员会投票选择算法(query-by-committee learning algorithm, QBC)联合 2 个分类器作出最后的决策.

2 目标领域文本初始标签算法

机器学习的"独立同分布"先验假设意味着要获得一个泛化能力较强的分类器需要一定规模训练数据.通常情况下,带标签的数据规模越大,训练的分类器对未带标签的数据的分类精度越高[10].然而,对许多应用来讲,人工标注大量带标签的数据,费时费力,尤其在海量数据背景下,人工标注能反映数据总体分布的大规模训练数据几乎是不可能的.针对跨领域文本情感分类,为了获得较多的目标领域的标签具有高可信度的文本(种子文本),本文首先人工标注少量的目标领域文本,然后采用 SMOTE 和 BootStrapping 方法,设计目标领域文本初始标签算法(IDLA),用于获得一定数量的目标领域中带标签的数据.

本文采用向量空间模型表示文本,设 $F=\{t_1,t_2,\cdots,t_n\}$ 为用于文本表示的特征空间,则一个文本 x 被表示为 $\boldsymbol{V}(x)=((t_1,w_1(x)),\cdots,(t_i,w_i(x)),\cdots,(t_n,w_n(x)))$,这里 $w_i(x)$ 表示特征 t_i 在文本 x 中的权重.

几类重要的文本集记号:

D^t—— 目标领域中人工标注了情感类别的文本集;

D^S—— 对 D^t 利用 SMOTE,在目标领域生成

的带情感类别标签的虚拟样本集;

D^{U}——目标领域中的无情感类别标签的文本集;

D^{N}——目标领域中,用于训练分类器的数据集;

D^{N}_{new}——D^{U} 中被分类器标记了高隶属度类别标签的文本集.

设 x' 是目标领域中的一个无标签文本,f 是一个支持向量机(support vector machine, SVM)分类器,$p(c|x')$ 为 x' 被 f 归属于类别 c(正面或者负面)的概率,即 x' 属于类别 c 的隶属度,记作 $membership(x',c)$.

设 x 是目标领域中的一个带标签文本,k 是一个正整数,称离 x 最近的 k 个同类样本构成的集合为 x 的同类 k 邻域,记为 $\delta_k(x)$,称 $\theta_k(x)=\max\limits_{x_i\in\delta_k(x)} dis(x,x_i)$ 为 x 的同类 k 邻域半径,这里 $dis(x,x_i)$ 为样本 x_i 与 x 的欧氏距离.

SMOTE 是 Chawla 提出的一种向上采样生成虚拟样本的方法. 其思想是:对一个样本,在它与它的若干个邻居样本间进行随机线性插值,将这个插值点作为一个虚拟数据.

本文中,SMOTE 被用来生成目标领域中带类别标签的虚拟样本,以扩展用于分类器的初始训练数据集规模.

具体地,设 $x\in D^{t}$,k 是一个正整数,对 $\forall x_i\in\delta_k(x)$,令 $x^{i}_{new}=x+\sigma(x_i-x)$,并赋予 x^{i}_{new} 与 x 相同的类别标签,这里 $\sigma\in(0,1)$ 是一个随机数.

将 SMOTE 作用于目标领域人工标注的小规模文本集 D^{t},这时可产生一个虚拟的带标签数据集 D^{S},将 D^{t} 与 D^{S} 合并,用来训练目标领域分类器,增大了训练样本的规模,一定程度上可以提高所训练的分类器的性能.

IDLA 算法的主要思想是利用了 BootStrapping 方法. 首先,在 D^{t} 上利用 SMOTE 获得一个带标签的虚拟样本集 D^{S},将 D^{t} 与 D^{S} 合并作为训练目标领域初始分类器的种子训练数据集;其次,利用初始分类器对目标领域的无标签样本数据集 D^{U} 中的文本进行分类,并从中挑选标签可信度高(即隶属度大)的样本 D^{C} 合并到训练数据集 D^{N} 中;然后,再利用新的训练数据集重新训练一个分类器,并用新的分类器对 D^{U} 中上一轮被判断为非高可信度的样本重新分类,并对 D^{N} 更新. 循环过程直到 D^{N} 达到一定的规模停止.

为直观起见,IDLA 算法示意图如图 1 所示:

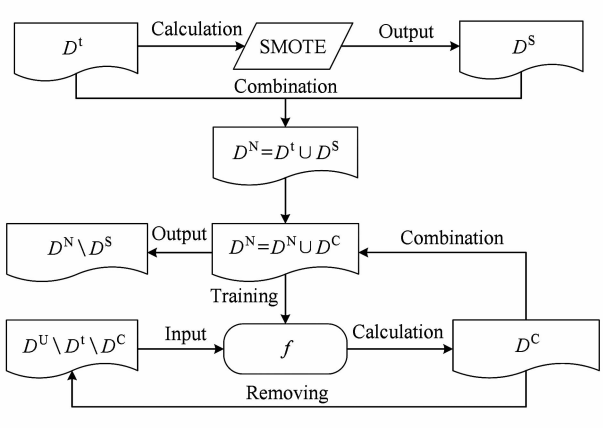


Fig. 1 Schematic diagram of IDLA algorithm.

图 1 IDLA 算法示意图

算法 1. IDLA.

输入:D^{U}, D^{t}, k, ε, p, $D^{N}_{new}=\varnothing$, $D^{S}=\varnothing$, $D^{C}=\varnothing$;

输出:D^{N}.

① 对每个 $x\in D^{t}$,计算 $\delta_k(x)=\{x_i\}_{i=1}^{k}$;对每个 $x_i\in\delta_k(x)$,生成一个随机数 $\sigma\in(0,1)$,并计算 x^{i}_{new},$D^{S}=D^{S}\cup\{x^{i}_{new}\}$;

② $D^{N}=D^{t}\cup D^{S}$;

③ 在 D^{N} 上训练一个 SVM 分类器 f;$D^{C}=\varnothing$;

④ 对每个 $x\in D^{U}\setminus D^{N}$,用 f 对 x 分类. 若对某个类别 c,有 $membership(x,c)\geqslant\varepsilon$,则 $D^{N}_{new}=D^{N}_{new}\cup\{x\}$,$D^{C}=D^{C}\cup\{x\}$;如果 $|D^{N}_{new}|>p$,则转步骤⑥;

⑤ $D^{N}=D^{N}\cup D^{C}$,转步骤③;

⑥ $D^{N}=D^{N}\setminus D^{S}$;输出 D^{N}.

IDLA 算法中的符号说明:

k——正整数,样本的同类邻域参数,用来控制领域的大小;

σ——随机数,$\sigma\in(0,1)$,用来控制生成的样本的范围;

ε——隶属度阈值,$\varepsilon\in(0,1)$,用于控制新标记样本的可信度;

f——表示一个 SVM 分类器;

p——预期得到的具有高可信度标签的新样本的个数.

3 分组提升集成学习算法

3.1 GAdaBEL 算法

针对引言中提到的实例迁移[24]用于跨领域情感分类时遇到的第 2 个问题,即如何组织和利用源领域和目标领域中的带标签数据问题,本文依据文献[25]中数据分组的思想和 AdaBoost 方法[26],设

计了 GAdaBEL 算法.

设源领域带标签文本集为 D^{L},目标领域带标签文本集为 D^{N}.

基于数据分组提升集成的学习算法(GAdaBEL)的主要思想如下:首先将 D^{L} 随机等量划分为 m 个子样本集 $D_1^{\mathrm{L}},D_2^{\mathrm{L}},\cdots,D_m^{\mathrm{L}}$,将 D_j^{L} 与 D^{N} 组合,得到 m 个组合块 $D_j^{\mathrm{L}}\cup D^{\mathrm{N}}(j=1,2,\cdots,m)$[25]. 然后,运用 AdaBoost 算法提升每个组合块 $D_j^{\mathrm{L}}\cup D^{\mathrm{N}}$ 上训练的 SVM 分类器,获得分类器 $L_j(j=1,2,\cdots,m)$. 最后,将 m 个分类器 $L_j(j=1,2,\cdots,m)$集成为一个分类器 L_{ensemble}. 在每个组合块上运用 AdaBoost 算法进行提升的流程如图 2 所示:

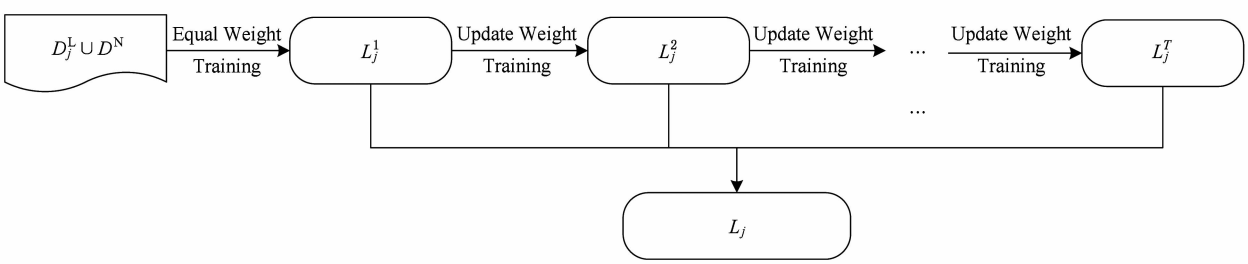


Fig. 2 Schematic diagram for AdaBoost on a data block.

图 2 在数据块上 AdaBoost 提升示意图

算法 2. GAdaBEL.

输入:$D^{\mathrm{N}},D^{\mathrm{L}},m$;

输出:集成分类器 $L_{\mathrm{ensemble}}(x)$.

① 将 D^{L} 随机等量划分成 m 个子样本集 D_1^{L},$D_2^{\mathrm{L}},\cdots,D_m^{\mathrm{L}}$,将 D_j^{L} 与 D^{N} 组合,得到 m 个组合块 $D_j^{\mathrm{L}}\cup D^{\mathrm{N}},j=1,2,\cdots,m$;

② for $D_j^{\mathrm{L}}\cup D^{\mathrm{N}},j=1,2,\cdots,m$

③ 对每个训练样本 $x_i\in D_j^{\mathrm{L}}\cup D^{\mathrm{N}}$,赋予初始权重 $H_j^1(x_i)=1/|D_j^{\mathrm{L}}\cup D^{\mathrm{N}}|$;

④ $M=|D_j^{\mathrm{L}}\cup D^{\mathrm{N}}|$;

⑤ for $(t=1;e_j^t\neq 0;T=t++)$ do

⑥ 在 $D_j^{\mathrm{L}}\cup D^{\mathrm{N}}$ 上,训练 SVM 分类器 $L_j^t(D_j^{\mathrm{L}}\cup D^{\mathrm{N}})$,获得分类函数 $f_j^t(x)$;

⑦ 计算分类器 $L_j^t(D_j^{\mathrm{L}}\cup D^{\mathrm{N}})$的分类错误率,

$$e_j^t=\frac{1}{2}\sum_{i=1}^{M}(H_j^t(x_i)\mid f_j^t(x_i)-y_i\mid);$$

⑧ 计算分类器 $L_j^t(D_j^{\mathrm{L}}\cup D^{\mathrm{N}})$的权重

$$c_j^t=\frac{1}{2}\ln((1-e_j^t)/e_j^t);$$

⑨ 调整每个样本 $x_i\in D_j^{\mathrm{L}}\cup D^{\mathrm{N}}$ 的权重 $H_j^{t+1}(x_i)$,并归一化,$H_j^{t+1}(x_i)=H_j^t(x_i)\exp\{-c_j^t y_i f_j^t(x_i)\}/Z(c_t)$;

其中 $Z(c_t)=\sum_{i=1}^{M}H_j^t(x_i)\exp\{-c_j^t y_i f_j^t(x_i)\}$;

⑩ end for

⑪ end for

⑫ $L_j(x)=\mathrm{sgn}[\sum_{t=1}^{T}c_j^t f_j^t(x)]$;

⑬ 集成分类器 $L_{\mathrm{ensemble}}(x)=\mathrm{sgn}[\sum_{j=1}^{m}L_j(x)]$.

若混合源领域全体带标签数据与目标领域带标签数据,再训练分类器,则由于目标领域带标签数据量相对于源领域太少,可能会导致目标领域数据被淹没,从而使分类器产生更倾向于表达源领域文本情感类别分布而非目标领域数据的类别分布的偏置. GAdaBEL 算法将源领域数据集 D^{L} 进行等量分割,并将划分的每一份与目标领域带标签数据集 D^{N} 进行组合作为训练集用于构建分类器 $L_j(x)$,在平衡源领域与目标领域数据量的同时,既利用了源领域的带标签数据,同时又充分利用了数量不足的目标领域带标签数据. 在每个数据块上使用 AdaBoost 方法是为了追求分类器 $L_j(x)$在部分数据上的正确率,而最后的集成分类器 $L_{\mathrm{ensemble}}(x)$是分组学习的自然要求,同时也是为了从整体上提高目标领域数据的分类正确率.

3.2 算法 GAdaBEL 收敛性分析

由文献[26]可知,组合数据块 $D_j^{\mathrm{L}}\cup D^{\mathrm{N}}$ 训练的分类器 $L_j(x)$的训练误差 $\varepsilon(L_j(x))$满足式(1):

$$\varepsilon(L_j(x))=\frac{1}{M}|\{i:L_j(x_i)\neq y_i\}|\leqslant\prod_{t=1}^{T}Z(c_t),\tag{1}$$

其中,T 是最大迭代次数.

$$Z(c_t)=\sum_{i=1}^{M}H_j^t(x_i)\exp\{-c_j^t y_i f_j^t(x_i)\}.\tag{2}$$

由式(1)(2)可知,减小 $L_j(x)$的训练误差可通过最小化 $Z(c_t)$来实现. 由 $Z(c_t)$是 c_j^t 的凸函数,$Z(c_t)$的最小值可通过求其极值点获得.

以 e_j^t 记第 t 次迭代所得到的分类器的分类错误率,它是被分类器错分对象的权重和,即 $e_j^t=$

$\sum_{i:y_i\neq f_t(x_i)} H_t(x_i)$. 注意到 $\sum_{i=1}^{M} H_t(x_i)=\sum_{i:y_i\neq f_t(x_i)} H_t(x_i)+\sum_{i:y_i=f_t(x_i)} H_t(x_i)=1$, 则有:

$$Z(c_t)=\sum_i H_t(x_i)\exp(-c_j^t y_i f_t(x_i))=\sum_{i:y_i\neq f_t(x_i)} H_t(x_i)\exp(-c_j^t y_i f_t(x_i))+\sum_{i:y_i=f_t(x_i)} H_t(x_i)\exp(-c_j^t y_i f_t(x_i))=e_j^t\exp(c_j^t)+(1-e_j^t)\exp(-c_j^t). \tag{3}$$

将 $Z(c_t)$ 对 c_j^t 求导,并令导数为 0,即 $Z'(c_t)=e_j^t\exp(c_j^t)-(1-e_j^t)\exp(-c_j^t)=0$,可解得:

$$c_j^t=\frac{1}{2}\ln\frac{1-e_j^t}{e_j^t}, \tag{4}$$

将式(4)代入式(3),得到:

$$Z(c_t)=2\sqrt{e_j^t(1-e_j^t)}, \tag{5}$$

将式(5)代入式(1)中的 $\prod_{t=1}^{T} Z(c_t)$, 得到:

$$\prod_{t=1}^{T} Z(c_t)=\prod_{t=1}^{T} 2\sqrt{e_j^t(1-e_j^t)}=\prod_{t=1}^{T}\sqrt{4e_j^t-4(e_j^t)^2}, \tag{6}$$

令 $\gamma_j^t=(\frac{1}{2}-e_j^t)$, 则 $\prod_{t=1}^{T} Z(c_t)=\prod_{t=1}^{T}\sqrt{1-4(\gamma_j^t)^2}$.

易知,当 $0\leqslant x\leqslant 1$ 时,$\exp(-x)-1\geqslant -x$. 由 $0\leqslant\gamma_j^t\leqslant\frac{1}{2}$,可知 $0\leqslant 4(\gamma_j^t)^2\leqslant 1$. 故有 $\exp(-4(\gamma_j^t)^2)-1\geqslant -4(\gamma_j^t)^2$,即 $\exp(-2(\gamma_j^t)^2)\geqslant\sqrt{1-4(\gamma_j^t)^2}$, 所以 $\prod_{t=1}^{T} Z(c_t)=\prod_{t=1}^{T}\sqrt{1-4(\gamma_j^t)^2}\leqslant\prod_{t=1}^{T}\exp(-2(\gamma_j^t)^2)=\exp(-2\sum_{t=1}^{T}(\gamma_j^t)^2)$.

由式(1)可知,$L_j(x)$ 的训练误差 $\varepsilon(L_j(x))=\frac{1}{M}\times|\{i:L_j(x_i)\neq y_i\}|\leqslant\exp(-2\sum_{t=1}^{T}(\gamma_j^t)^2)$. 这表明:若 $f_t(x_i)$ 的结果略好于随机猜测,即存在一个 $\gamma>0$ 使得 $\gamma_j^t>\gamma$,则 $\varepsilon(L_j(x))\leqslant\exp(-2\sum_{t=1}^{T}\gamma^2)=\exp(-2T\gamma^2)$. 容易推导,对任意的 $\varepsilon>0$,存在一个与 ε 有关的最大迭代次数 $T\leqslant\frac{1}{2\gamma^2}\ln\frac{1}{\varepsilon}$,当迭代次数大于 T 时,即可保证使 $\varepsilon(L_j(x))\leqslant\varepsilon$. 所以,对任意的 $j=1,2,\cdots,m$, $L_j(x)$ 收敛.

对构造的集成分类器 $L_{\text{ensemble}}(x)=\text{sgn}[\sum_{j=1}^{m} L_j(x)]=\text{sgn}[\sum_{j=1}^{m}\text{sgn}[\sum_{t=1}^{T_j} c_j^t f_j^t(x)]]$, 由上述推导可知,其训练误差:

$$\varepsilon(L_{\text{ensemble}}(x))=\sum_{j=1}^{m}\varepsilon(L_j(x))\leqslant\sum_{j=1}^{m}\exp(-2\sum_{t=1}^{T_j}(\gamma_j^t)^2). \tag{7}$$

式(7)表明:若每个 $f_j^t(x)$ 略好于随机猜测,即存在一个 $\gamma>0$,使得对任意的 j 有 $\gamma_j^t>\gamma$. 则 $\varepsilon(L_{\text{ensemble}}(x))\leqslant\sum_{j=1}^{m}\exp(-2\sum_{t=1}^{T_j}(\gamma_j^t)^2)\leqslant m(\max_j(\exp(-2T^j\gamma^2)))$. 容易推导,对任意的 $\varepsilon>0$,存在一个与 ε 和 m 有关的最大迭代次数 $T_{\max}\leqslant\frac{1}{2\gamma^2}\frac{1}{\ln\varepsilon-\ln m}$,当每个机器的迭代次数大于 $T_{\max}$ 时,即可保证使 $\varepsilon(L_{\text{ensemble}}(x))\leqslant\varepsilon$. 所以,集成分类器 $L_{\text{ensemble}}(x)$ 收敛.

4 实验结果及分析

4.1 实验数据与预处理

本文实验采用 4 个领域的英文语料,分别是 DVD 评论(digital versatile disc, DVD)、书籍评论(Book)、电子评论(Electronics)和厨房以及家庭用品评论(Kitchen),所有的评论文本来自 Amazon 购物网站上的英文评论,并由专家手工标注每篇文本的情感倾向性,情感倾向性类别分为正面(+1)和负面(-1). 每个领域的正面和负面文本各有 1 000 篇. 每篇文本的特征采用一元和二元混合语法形式,如"even_enjoy","ruined","could_not","you_like"等. 特征选择采用文献[27]提出的基于 Fisher 判别准则的特征选择方法. 在实验中,我们发现选取 Fisher 值最大的前 1 000 个有效特征时迁移效果较好. Pang 等人[5]认为在情感分类中情感倾向往往不是通过重复某个词来表达的,他们在实验中发现"使用出现与否作为词的权重比使用词出现的频率作为其权重效果要好",我们通过实验也验证了这个结论. 因此,本文采用布尔值(Boolean value)作为特征权重,即若一个特征在一篇文本中出现,其权重为 1,否则为 0.

4.2 实验结果与分析

本文实验中使用支持向量机(SVM)作为分类器(台湾大学林智仁等人[28]开发设计的 LibSVM 工具包,所有参数均采用默认值). 我们设计了 3 个实

验,并采用正确率(*Accuracy*)作为评价指标,最终结果采用5倍交叉验证取平均值.

1) 直接用源领域数据训练分类器

训练集为正面文本和负面文本各800篇,共1 600篇,测试集为正面文本和负面文本各200篇,共400篇.实验结果如表1所示:

Table 1 Accuracy of Source Domain Transferring to Target Domains Directly

表1 源领域迁移到目标领域的分类精度

Source Domain	Target Domain			
	DVD	Book	Electronics	Kitchen
DVD	**0.835**	0.766	0.734	0.746
Book	0.791	**0.818**	0.738	0.745
Electronics	0.740	0.730	**0.846**	0.749
Kitchen	0.704	0.707	0.804	**0.881**

从表1可以看出:

① 对DVD,Book,Electronics,Kitchen四个领域,本领域分类结果明显优于跨领域分类结果,它们的本领域分类精度分别可达到0.835,0.818,0.846,0.881.这表明文本情感分类是一个领域相关问题.

② 在跨领域测试中,Book→DVD的结果优于Book→Electronics和Book→Kitchen;DVD→Book的结果优于DVD→Electronics和DVD→Kitchen;Electronics→Kitchen的结果优于Electronics→Book和Electronics→DVD;Kitchen→Electronics的结果要优于Kitchen→Book和Kitchen→DVD.这表明Book与DVD,Electronics与Kitchen的相关性比较大.

2) 参数k和σ对GAdaBEL算法的影响

SMOTE中的参数$k=\{1,2,\cdots,6\}$是同类别最近邻样本控制数,它决定了在多大范围内产生多少个虚拟样本;参数σ用来控制一个虚拟样本邻近目标样本的程度,尽管在SMOTE中它是一个随机数,但本实验中令其取$\sigma=\{0.1,0.2,\cdots,0.9\}$九个不同的值,以考察其影响.参数$k$和$\sigma$对于IDLA算法产生的所谓高可信度样本具有一定影响,进而对GAdaBEL算法的结果产生影响.

本实验中,选取了2 000篇目标领域已标记文本,其中的正面和负面文本各50篇作为最初的"人工标记的少量文本".从剩下的1 900篇文本中抽取正面和负面文本各500篇作为测试样本.选取源领域已标记文本2 000篇,将源领域的数据分成8组,即$m=8$.在IDLA算法中,参数$\varepsilon=0.72$,$p=150$.实验中,在4个领域中指定其中1个为目标领域,其余3个领域作为源领域,图3~6分别给出了源领域向目标领域迁移的平均实验结果.

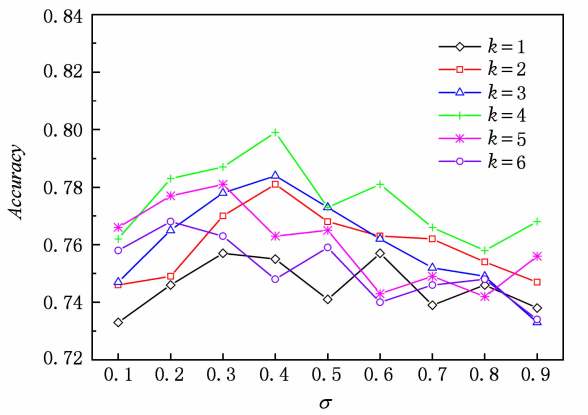


Fig. 3 Average accuracy of transferring to DVD in different values of k and σ.

图3 不同k和σ值下迁移到DVD正确率

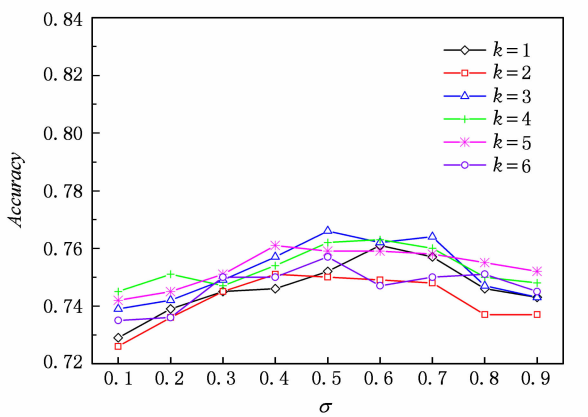


Fig. 4 Average accuracy of transferring to Book in different value of k and σ.

图4 不同k和σ值下迁移到Book正确率

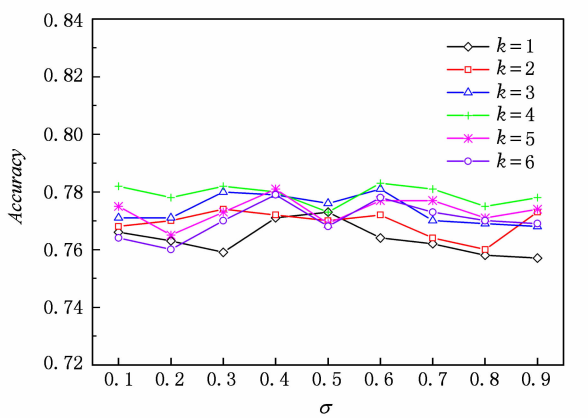


Fig. 5 Average accuracy of transferring to Electronics in different value of k and σ.

图5 不同k和σ值下迁移到Electronics正确率

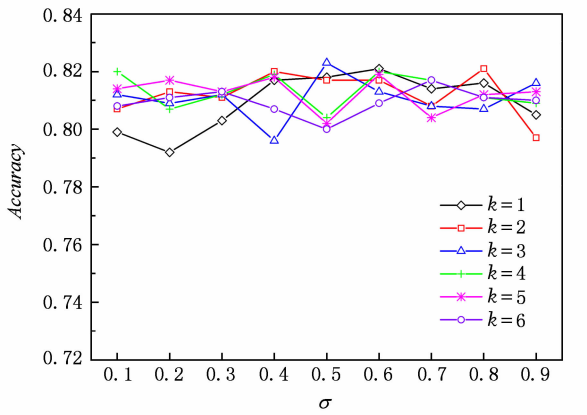


Fig. 6 Average accuracy of transferring to Kitchen in different value of k and σ.

图 6 不同 k 和 σ 值下迁移到 Kitchen 正确率

由图 3～6 可以看出:

① 总体上,平均分类精度随着 k 值的增大,重现由低到高再到低的过程. 无论哪种目标领域,当 k 值取 3 或 4 时,平均分类精度最大. 因此,SMOTE 通过生成一定量的虚拟样本,可在一定程度上缓解目标领域标记样本不足的问题.

② 随着 σ 值增大,尽管对 Electronics 和 Kitchen 两个目标领域平均精度比较平稳,但总体上呈现先提高后下降的趋势,且在 σ 取值为 0.4,0.5 或 0.6 时,效果更好. 这表明产生的虚拟样本太靠近 2 个真实的参照样本时,效果不好.

由实验 2,得到针对各目标领域的最佳 k 值和 σ 值组合,如表 2 所示. 后续实验中,针对不同的目标领域,采用表 2 中对应的 k 值和 σ 值组合.

3) 不同跨领域文本情感分类方法的比较实验

实验数据:源领域训练数据集为正面文本和负面文本各 800 篇,共 1 600 篇,目标领域测试数据集为正面文本和负面文本各 200 篇,共 400 篇. 在涉及 IDLA 算法的方法中,在 400 篇目标领域测试数据以外,选用了正面、负面文本各 50 篇作为少量的人工标注数据,即 IDLA 算法中的 D^t.

Table 2 Best Combined Values of k and σ for 4 Target Domains

表 2 4 个目标领域的最优 k 值和 σ 值

Target Domain	k	σ	*Accuracy*
DVD	4	0.4	0.799
Book	3	0.5	0.766
Electronics	4	0.6	0.783
Kitchen	3	0.5	0.823

本实验中采用的方法名称及含义:

1) LibSVM[28]. 在源领域数据集上训练 SVM 分类器,直接在目标领域数据集上进行测试.

2) LP-based[23]. 通过基于图排序的方法(LP-based)实现源领域带标签数据向目标领域无标签数据的情感标签传播.

3) IDLA. 将 IDLA 得到的目标领域带标签数据集 D^N 与源领域带标签数据合并,并用来训练 SVM 分类器,然后在目标领域数据集上进行测试.

4) GAdaBEL($-D^S$). 在 GAdaBEL 算法中令 $D^N=D^t$,将源领域带标签数据分组,并直接与目标领域少量人工标记数据组合,即不在目标领域中执行 IDLA 算法.

5) GAdaBEL. 本文所提出的基于分组提升集成的跨领域文本情感分类方法.

5 种方法的比较实验结果如表 3 所示:

Table 3 Accuracy of Cross-domain Classification of 5 Transferring Algorithms

表 3 5 种迁移方法跨领域分类正确率

Task	LibSVM	LP-based	IDLA	GAdaBEL($-D^S$)	GAdaBEL
Book→DVD	0.791	0.798	0.809	0.788	**0.832**
Book→Electronics	**0.738**	0.720	0.731	0.724	0.737
Book→Kitchen	0.745	0.770	0.727	0.744	**0.779**
DVD→Book	0.766	0.780	0.780	0.770	**0.792**
DVD→Electronics	0.734	0.768	**0.790**	0.731	0.785
DVD→Kitchen	0.746	0.743	0.755	0.762	**0.813**
Electronics→Book	0.740	0.715	0.748	0.737	**0.771**
Electronics→DVD	0.730	0.740	0.769	0.729	**0.784**
Electronics→Kitchen	0.749	0.835	0.770	0.833	**0.877**
Kitchen→Book	0.704	**0.735**	0.714	0.709	0.734
Kitchen→DVD	0.707	0.733	0.772	**0.783**	0.781
Kitchen→Electronics	0.804	0.815	0.820	0.811	**0.826**
Average	0.746	0.763	0.765	0.760	**0.793**

由表 3 可以看出:

① 对几乎所有的领域间迁移,其他方法均优于 LibSVM,这表明对基于迁移策略的跨领域文本分类,在利用源领域数据条件下,融合目标领域数据信息,有助于提高分类精度.

② GAdaBEL 与 IDLA 相比,平均精度提高了 0.028,这表明采用数据分组集成提升学习策略比采用不进行提升的单个分类器效果更好.

③ GAdaBEL 与 GAdaBEL($-D^S$)相比,平均精度提高了 0.033,这表明通过在目标领域使用 IDLA 算法扩充一部分 IDLA 算法认为的高可信度样本,有利于基于迁移策略的跨领域文本分类.

④ GAdaBEL 与 LP-based 相比,对所有领域间的迁移,精度均有所提高,且平均精度提高了 0.03.

⑤ 总体上,与其他方法相比,GAdaBEL 的表现是最好的,Electronics→Kitchen 的精度达到了 0.877,平均精度达到了 0.793.

5 结束语

对于互联网环境下的文本数据,其领域多样性是普遍存在的.针对跨领域文本情感分类任务目标领域中带标记样本偏少,采用样本迁移策略,本文提出了一种基于分组提升集成的半监督学习方法(GAdaBEL).该方法分为 2 个阶段:1)为了解决目标领域带标记数据较少这一问题,基于半监督学习思想,提出了一种综合运用 SMOTE 和 BootStrapping 技术的初始数据标签算法(IDLA),它以少量的目标领域人工标注数据为半监督信息,以 SMOTE 产生一定数量的虚拟样本,在人工标记的少量数据、虚拟样本以及分类器认定的高可信度样本所组成的混合数据集上,利用 BootStrapping 技术迭代地提升分类器,以获得较多的目标领域高可信度标注数据;2)源领域的带标签数据进行等量分割,并将其中的每一份与目标领域高可信数据进行组合,在每一个组合块上运用 AdaBoost 的方法提升地训练分类器,然后将每个组合块上的分类器进行集成,得到最终的分类器.在来自于亚马逊网站的 DVD,Book,Electronics,Kitchen 四个领域的英文语料上进行了实验,验证了本文提出的方法在跨领域文本情感分类中的有效性.

本项工作的特点是综合运用了多种机器学习策略与技术,包括半监督学习策略和数据分组集成策略、SMOTE、AdaBoost 技术、集成学习技术等.其中,半监督学习策略、SMOTE、BootStrapping 技术主要解决目标领域带标签数据量不足的问题;数据分组集成策略、AdaBoost 技术和集成学习技术,主要用来解决由于迁移过程中源领域与目标领域数据量失衡所引起的分类器偏置问题.

本文仅从实例迁移的角度对跨领域的情感分类的问题进行了分析,未来的研究工作将同时考虑实例迁移和特征迁移问题,以进一步提高跨领域情感分类精度.另外,带标签的数据往往可能来自于多个源领域,如何利用多个源领域的数据来辅助单个目标领域的情感分类也是一个值得研究的问题.

参 考 文 献

[1] Du Weifu, Tan Songbo, Cheng Xueqi, et al. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2010: 111-120

[2] Wang Suge, Li Deyu, Wei Yingjie. A method of text sentiment classification based on weighted rough membership [J]. Journal of Computer Research and Development, 2011, 48(5): 855-861 (in Chinese)
(王素格,李德玉,魏英杰. 基于赋权粗糙隶属度的文本情感分类方法[J]. 计算机研究与发展, 2011, 48(5): 855-861)

[3] Bollegala D, Weir D, Carroll J. Cross-Domain sentiment classification using a sentiment sensitive thesaurus [J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(8): 1719-1731

[4] Lin Zheng, Tan Songbo, Cheng Xueqi. Sentiment classification analysis based on extraction of sentiment key sentence [J]. Journal of Computer Research and Development, 2012, 49(11): 2376-2382 (in Chinese)
(林政,谭松波,程学旗. 基于情感关键句抽取的情感分类研究[J]. 计算机研究与发展, 2012, 49 (11): 2376-2382)

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [C] //Proc of the Association of Computational Linguistics Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2002: 79-86

[6] Yu L C, Wu J L, Chang P C, et al. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news [J]. Knowledge-Based Systems, 2013, 41: 89-97

[7] Zhu Zhu, Dai Daming, Ding Yaxing, et al. Employing emotion keywords to improve cross-domain sentiment classification [G] //LNCS 7717: Chinese Lexical Semantics. Berlin: Springer, 2013: 64-71

[8] Kaya M, Fidan G, Toroslu I H. Transfer learning using twitter data for improving sentiment classification of turkish political news [G] //LNEE 264: Information Sciences and Systems 2013. Berlin: Springer, 2013: 139-148

[9] Jambhulkar P, Nirkhi S. A survey paper on cross-domain sentiment analysis [J]. International Journal of Advanced Research in Computer and Communication Engineering, 2014, 3(1): 5241-5245

[10] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357

[11] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification [C] //Proc of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg, PA: ACL, 2007: 440-447

[12] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment [C] //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 751-760

[13] Xia Rui, Zong Chengqing. A POS-based ensemble model for cross-domain sentiment classification [C] //Proc of the 5th Int Joint Conf on Natural Language Processing. Berlin: Springer, 2011: 614-622

[14] Tan Songbo, Wu Gaowei, Tang Huifeng, et al. A novel scheme for domain-transfer problem in the context of sentiment analysis [C] //Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 979-982

[15] Wu Qiong, Tan Songbo, Xu Hongbo, et al. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon [J]. Journal of Computer Research and Development, 2010, 47(12): 2123-2131 (in Chinese)
(吴琼, 谭松波, 徐洪波, 等. 基于随机游走模型的跨领域倾向性分析研究[J]. 计算机研究与发展, 2010, 47(12): 2123-2131)

[16] Wu Qiong, Tan Songbo, Zhang Gang, et al. Research on cross-domain opinion analysis [J]. Journal of Chinese Information Processing, 2010, 24(1): 77-83 (in Chinese)
(吴琼, 谭松波, 张刚, 等. 跨领域倾向性分析相关技术研究[J]. 中文信息学报, 2010, 24(1): 77-83)

[17] Wu Qiong, Tan Songbo. A two-stage framework for cross-domain sentiment classification [J]. Expert Systems with Applications, 2011, 38(11): 14269-14275

[18] Wu Qiong, Liu Yue, Shen Huawei, et al. A unified framework for cross-domain sentiment classification [J]. Journal of Computer Research and Development, 2013, 50(8): 1683-1689 (in Chinese)
(吴琼, 刘悦, 沈华伟, 等. 面向跨领域情感分类的统一框架[J]. 计算机研究与发展, 2013, 50(8): 1683-1689)

[19] Liao Xuejun, Xue Ya, Carin L. Logistic regression with an auxiliary data source [C] //Proc of the 22nd Int Conf on Machine Learning. New York: ACM, 2005: 505-512

[20] Jiang Jing, Zhai Chengxiang. Instance weighting for domain adaptation in NLP [C] //Proc of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg, PA: ACL, 2007: 264-271

[21] Aue A, Gamon M. Customizing sentiment classifiers to new domains: A case study [C/OL] //Proc of Recent Advances in Natural Language Processing (RANLP). Amsterdam: John Benjamins, 2005 [2014-02-20]. http://research.microsoft.com/pubs/65430/new_domain_sentiment.pdf

[22] Xia Rui, Hu Xuelei, Lu Jianfeng, et al. Instance selection and instance weighting for cross-domain sentiment classification via PU learning [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2176-2182

[23] Li Shoushan, Xue Yunxia, Wang Zhongqing, et al. Active learning for cross-domain sentiment classification [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2127-2133

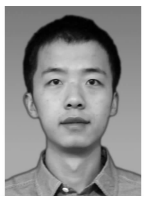
[24] Pan S J, Yang Qiang. A survey on transfer learning [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(10): 1345-1359

[25] Li Cen. Classifying imbalanced data using a bagging ensemble variation (BEV) [C] //Proc of the 45th Annual Southeast Regional Conference. New York: ACM, 2007: 203-208

[26] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297-336

[27] Wang Suge, Li Deyu, Song Xiaolei, et al. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification [J]. Expert Systems with Applications, 2011, 38(7): 8696-8702

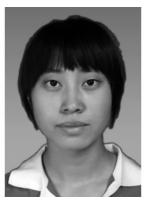
[28] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-39



Zhao Chuanjun, born in 1986. PhD candidate in Shanxi University. Student member of China Computer Federation. His main research interests include intelligent retrieval and natural language processing (zhaochuanjun@foxmail.com).

Wang Suge, born in 1964. Professor and PhD supervisor in Shanxi University. Member of China Computer Federation. Her main research interests include natural language processing and intelligent retrieval.

Li Deyu, born in 1965. Professor and PhD supervisor in Shanxi University. Senior member of China Computer Federation. His main research interests include data mining, intelligent retrieval, social network (lidy@sxu.edu.cn).



Li Xin, born in 1989. Master candidate in Shanxi University. Her main research interests include intelligent retrieval (lixin_it@163.com).