计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170496
55(11):2439-2451,2018

# 多视角特征共享的空间对齐跨领域情感分类

贾熹滨[1,2] 靳 亚[1,2] 陈军成[1]
[1](北京工业大学信息学部 北京 100124)
[2](多媒体与智能软件技术北京市重点实验室(北京工业大学) 北京 100124)
(jiaxibin@bjut.edu.cn)

# Domain Alignment Based on Multi-Viewpoint Domain-Shared Feature for Cross-Domain Sentiment Classification

Jia Xibin[1,2], Jin Ya[1,2], and Chen Juncheng[1]
[1](*Faculty of Information Technology, Beijing University of Technology, Beijing* 100124)
[2](*Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology (Beijing University of Technology), Beijing* 100124)

**Abstract** Plenty and well labeled training samples are significant foundation to make sure the good performance of supervising learning, whereas there is a problem of high labor-cost and time-consuming in the samples. Furthermore, it is not always feasible to get the plenty of well-labeled sample data in every application to support the classification training. Meanwhile, directly employing the trained model from the source domain to the target domain normally causes the problem of accuracy degradation, due to the information distribution discrepancy between the source domain and the target domain. Aiming to solve the above problems, we propose an algorithm named domain alignment based on multi-viewpoint domain-shared feature for cross-domain sentiment classification (DAMF). Firstly, we fuse three sentiment lexicons to eliminate the polarity divergence of domain-shared feature words that are chosen by mutual information value. On this basis, we extract the word pairs that have the same sentiment polarity in the same domain by utilizing four syntax rules and the word pairs that have strong association relation in the same domain by utilizing association rules algorithm. Then, we use the domain-shared words that have no polarity divergence as a bridge to establish an indirect mapping relationship between domain-specific words in different domains. By constructing the unified feature representation space of different domains, the domain alignment is achieved. Meanwhile, the experiments on four public data sets from Amazon product reviews corpora show the effectiveness of our proposed algorithm on cross-domain sentiment classification.

**Key words** sentiment classification; cross-domain; polarity divergence; association rules; unified feature representation space; domain space alignment

**收稿日期**:2017-06-21;**修回日期**:2017-12-29

**基金项目**:国家重点研发计划项目(2017YFC0803300);国家自然科学基金项目(91546111,91646201,61672071);北京市教委重点项目(KZ201610005009)
This work was supported by the National Key Research and Development Program of China (2017YFC0803300), the National Natural Science Foundation of China (91546111, 91646201, 61672071), and the Key Projects of Beijing Municipal Education Commission (KZ201610005009).

**通信作者**:陈军成(juncheng@bjut.edu.cn)

**摘 要** 大量有效样本标注是有监督学习性能的重要保证,但又存在耗时且人力成本高的问题.加之,在实际应用环境,很难在每个应用领域都有足够的标定样本数据支持分类器的训练.而将源领域所获的训练模型直接用于目标领域,又由于目标领域和源领域信息分布差异,会导致跨领域分类器应用准确率降低的问题.针对以上问题,提出一种基于多视角共享特征的领域空间对齐的跨领域情感分类(domain alignment based on multi-viewpoint domain-shared feature for cross-domain sentiment classification, DAMF)算法.该算法首先通过融合多个情感词典,消除通过互信息值所选择的领域共享特征中情感词的极性分歧问题.在此基础上,以领域间无歧义共享特征为桥梁,结合通过语法规则提取的各领域中有相同极性的情感词对和通过关联规则学习的各领域中有强关联关系的特征词对,进行领域间相同极性的专有情感词对和强关联关系的特征词对的提取,构建目标领域和源领域数据的统一特征表示空间,减小了领域间因极性分歧和特征分布不同造成的差异,实现不同领域空间对齐.同时在公共数据集上的跨领域实验表明,基于多视角共享特征的领域空间对齐跨领域倾向性分析算法一定程度上提高了跨领域情感分类的准确率.

**关键词** 情感分类;跨领域;极性分歧;关联规则;统一特征表示空间;领域空间对齐

**中图法分类号** TP391

随着互联网的快速发展,网络上出现了大量由用户发起的评论信息,包括对电影、产品、社会热点事件等的评论,这些评论信息中通常包含了丰富的情感信息,表达了用户对商品、电影等话题的观点和看法.如果能自动地将这些信息加以处理、分析和总结可以为用户和公司等提供决策帮助[1],同时也方便政府了解群众对于社会热点事件的观点和看法[2].例如,用户在网络上购买商品时可以参考该商品其他用户的评价意见,为自己提供决策帮助;公司可以通过收集用户对商品的评价信息并分析出商品在各个方面的优点和不足,为公司改善商品质量、对用户进行个性化推荐和增加商品销售量提供帮助等.因此,情感分类技术(又称意见挖掘技术)因能及时地对网络中带有主观情绪色彩的文本进行分析并带来巨大的经济和社会价值而引起了较广泛的关注,成为了近年来的研究热点[3].

传统的机器学习算法,尤其是有监督学习算法已经被广泛应用在电影影评、产品评论和微博短文本等带有主观情感色彩的文本情感分类中[3-6],并且作为情感分类问题主流的算法,也取得了很好的研究成果.但是,有监督学习算法通常需要大量的标定数据来训练情感分类器,并且要求训练样本和测试样本应同分布以便共享信息[7].而网络上评论信息涉及的领域范围非常广泛,为每一个领域手工标定足够的训练样本是非常耗时耗力的[8].同时,不同的用户在面对不同的评价主体时,评价角度和表达方式通常存在很大的差异,导致领域间信息非同分布.比如,“分辨率”、“电池”、“durable”等词语经常会出现在电子产品的评论语料中,而极少出现在电影、书籍类产品的评论语料中;同样,“纸质”、“印刷”、“misspelling”等词语会频繁出现在书籍类产品中,却很少出现在电影、电子产品等的评论语料中.最后是情感词的极性分歧问题,即同一情感词在不同领域的语料中可能有不同的情感倾向.比如,“long”在厨房用具的评论语料中可能表示使用寿命长,是一个正极性的情感词,而在书籍的评论语料中可能表示段落冗长,是一个负极性的情感词.所以基于以上问题,很难将一个在源领域训练好的分类器直接应用到一个全新的目标领域[9-10].

近年来,为了解决领域间差异造成的情感分类器准确率降低的问题.跨领域情感分类技术的研究得到了快速的发展,目前的解决方法主要从样本、特征和主题3个方面的迁移进行研究.就特征迁移而言,主要通过一些策略寻找源领域和目标领域间的共享特征,构建跨领域数据的统一特征表示空间来消除领域间的差异[8-9,11-13].

为解决领域间差异造成的情感分类器准确率降低的问题,本文提出了一种基于多视角共享特征的领域空间对齐跨领域情感分类(domain alignment based on multi-viewpoint domain-shared feature for cross-domain sentiment classification, DAMF)算法.本文中的特征词是指包含在各领域语料中的词汇,通常分为有情感极性的特征词(也叫情感词)和其他特征词(指描述对象等无极性词汇).算法借助已有的情感词典和改进的互信息(mutual information, MI)[8]技术,建立领域间无歧义共享特征集合,并通

过句法分析和关联规则算法进行领域间专有特征词对的提取,实现领域词典的扩展和领域间信息分布空间的对齐.同时,在 Amazon 产品评论数据集[11]上和已有的相关算法进行比较实验,表明本文提出的算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

如引言所述,目前解决跨领域的情感分类问题的方法主要有 3 种:基于样本加权重采样的方法、基于特征对齐的方法和基于主题模型的方法.

基于样本加权重采样解决跨领域情感分类问题的关键技术在于为原始领域的标定样本采用加权策略,使训练数据和测试数据有相似的分布,适用于源领域和目标领域的样本分布差距较小的情况.主要的研究成果有:Dai 等人[14]提出的 TrAdaBoost 的半监督算法,在训练过程中通过加入目标领域少量的标定样本,在优化损失函数的过程中,加强原始领域训练样本中与目标领域有相似分布的样本权重值,减少与目标领域不相似的样本权重值,使训练过程更倾向于目标领域的分布,从而建立目标领域的情感分类器;Hu 等人[15]提出了基于类分布的多领域自适应算法(muti-domain adaptation algorithm based on the class distribution, MACD),算法通过多个源领域的标定样本训练多个基础分类器,并根据源领域和目标领域的类别分布距离来动态调整和选择高置信度的标定数据加入训练样本集,使每一个原始领域都更好地适应目标领域,建立应用于目标领域的情感分类器的集成分类器.Li 等人[16]首先通过主动学习的策略选取少量目标领域带标签的数据,然后用源领域和目标领域中带标签的数据训练 2 个独立的分类器,采用委员会投票算法根据 2 个分类器的结果作出最后的决策.

基于特征对齐解决跨领域情感分类问题的关键技术在于学习 2 个领域信息的统一特征表示空间,减少领域信息分布的差异,适用于因源领域和目标领域的样本分布差距较大、很难在样本层面找到 2 个领域间交集的情况.主要的研究成果有:Blitzer 等人[9]提出了结构对应学习(structural correspondence learning, SCL)的算法,通过选择原始领域和目标领域都频繁出现的“枢纽”特征集合,建立学习“枢纽”特征和其他特征间的关联关系模型,实现源领域和目标领域特征层面的对齐.Pan 等人[8]提出了光谱对齐(spectral feature alignment, SFA)算法,通过改进的互信息来选取领域专有特征和领域通用特征,并通过在通用特征和专有特征建立的二部图中进行图谱聚类操作学习到新的特征表示,以领域通用特征为桥梁,实现领域专有特征的对齐,减少领域间的差距.吴琼等人[7]提出了基于图的随机游走模型,通过利用源领域和目标领域的文本和词之间的关联关系来实现知识在领域间的迁移,借助图迭代计算的思想对待标注文本计算情感分层,来判断文本的情感倾向性.Glorot 等人[17]提出了基于堆叠去噪自动编码机(stacked denoising auto-encoders, SDA)的跨领域情感分类算法,通过深度神经网络的隐层节点学习不同领域间通用的特征表示,通过通用特征构建新的特征空间,减少不同领域间特征分布的差异,实现源领域和目标领域特征的对齐.

基于主题的跨领域情感分类技术主要通过提取能代表不同领域文本的共有潜在特征(包括潜在主题、主要组成元素等)来减少领域间信息分布的差异.主要的研究成果有:Li 等人[12]提出了主题关联分析(topic correlation analysis, TCA)算法,通过提取领域间共享主题和各个领域的特定主题,计算各个领域特定主题间的相关性,利用相关性将各个领域的特征映射到新的特征空间,在新的特征空间训练情感分类器,用于目标领域的情感分类.

这 3 种方法都是通过一定技术学习目标领域与源领域之间具有相同分布的样本或者潜在的共享特征,并以具有相同分布的源领域样本或者共享特征为桥梁,实现目标领域与源领域在样本层面的对齐,来获得领域间的统一特征表示空间.但是,当 2 个领域的数据分布差异非常大或者选取的共享特征存在极性分歧时,都将会导致跨领域情感分类器的准确率降低,甚至会出现负迁移[18].

## 2 一种空间对齐跨领域情感分类算法

### 2.1 总体框架

基于多视角共享特征的领域空间对齐的跨领域情感分类算法的总体框架如图 1 所示.算法首先利用已有的情感词典,建立无极性分歧的情感词集合,并结合改进的 MI[8]技术来选择预处理后的源领域和目标领域语料中共享的无极性分歧的特征,构成共享特征集合.然后,通过句法分析和关联规则算法,分别迭代地获取各领域中具有相同极性的特征

词对和具有强关联关系的特征词对. 在此基础上,以领域间无歧义共享特征集合为桥梁,进行领域间专有特征词对的提取,实现领域词典的扩展和领域间信息分布空间的对齐. 最后根据源领域对齐后的标定样本训练分类器,即可得到适用于目标领域的情感分类模型.

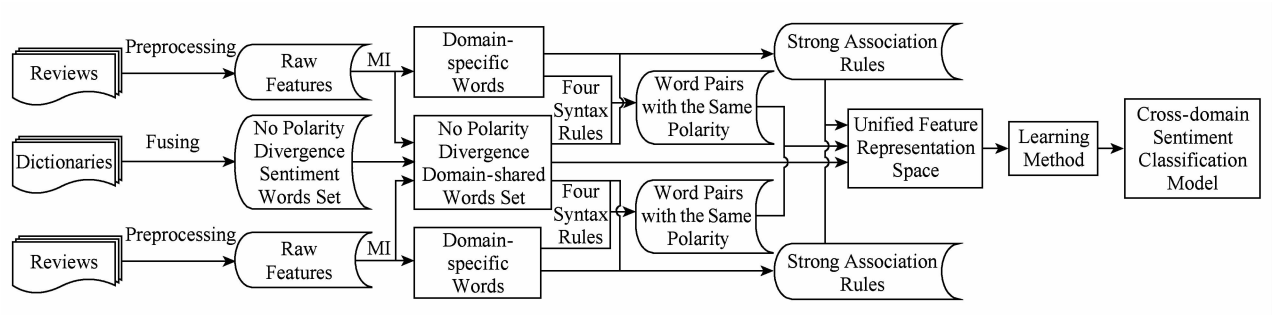



Fig. 1 Overview of our proposed algorithm for cross-domain sentiment classification

图 1 基于多视角共享特征的领域空间对齐跨领域情感分类算法的总体框架

## 2.2 领域间无歧义共享特征集合的构建

本节我们将介绍如何从多视角构建领域间无歧义共享特征集合,为消除不同领域中情感词极性分歧和对齐领域信息分布空间,并建立跨领域情感分类模型做基础.

### 2.2.1 无极性分歧的情感词集合的构建

情感词典通常为待分析文本的关键情感词提供极性参考,所以在情感分析系统中扮演着重要的角色. 常见的英文情感词典有:SentiWordNet[19],Bing Liu's Sentiment Lexicon[20](下文简写为 BLSentiLex),MPQA[21]等. 这些情感词典通常都是基于一定的语料库进行统计手工标注或者利用算法学习获得,包含情感词的集合及其对应的极性等属性. 因语料库的差异,不同的情感词典包含的情感词和其对应的极性不尽相同,如情感词"defeat"在 BLSentiLex[20]词典中的情感极性为正极性,但是在 MPQA[21]词典中的情感极性为负极性. 同时也有"good","wonderful","bad","worst"等情感词在各词典中的极性完全一致. 所以为了消除情感词的极性分歧,我们将常用的情感词典进行融合,构建无极性分歧的情感词集合.

3 种英文情感词典的详细描述如表 1 所示. 在对 3 种情感词典中的情感词进行清洗整理时,去除情感词极性强弱、词性等属性信息,仅保留正负极性信息. 在 SentiWordNet 词典中,分别给出了情感词的正负极性的分数值,我们通过计算情感词正负极性分数值的差值来标注情感极性,即当差值大于 0 时,该词的情感极性被标注为正极性;当差值小于 0 时,该词的情感极性被标注为负极性;当差值等于 0 时,该词的情感极性被标注为中性极性. 在融合 3 个情感词典时,通过选择在 3 个情感词典中均有出现的且具有相同的情感极性的词,构成无极性分歧的情感词集合来消除情感词在不同语料中的极性分歧问题.

**Table 1 Detailed Description of the Four Sentiment Dictionaries**

**表 1 词典详情描述**

| Dictionaries | Number of All Kinds of Sentiment Polarity Words | | |
|---|---|---|---|
| | Positive Polarity | Negative Polarity | Neutral Polarity |
| BLSentiLex[20] | 2 006 | 4 783 | 0 |
| SentiWordNet[19] | 23 147 | 26 440 | 157 354 |
| MPQA[21] | 2 304 | 4 152 | 430 |
| Fused Sentiment Word Set | 1 151 | 2 292 | 0 |

通过融合 3 个情感词典,构建包含 3 443 个无极性分歧的情感词集合 $W_{\mathrm{sh}}^{\mathrm{dict}}$. 表 2 中列举了部分融合后的情感词.

**Table 2 An Example of the Fused Sentiment Word Set**

**表 2 融合后的情感词集合举例**

| Positive Polarity Words | Negative Polarity Words |
|---|---|
| stable, worth, grateful, good, well, excellent, smile, encouragingly, enjoy, outsmart, fine, successful … | poorly, unfeeling, disobedient, incompatible, blur, undesirable, disappoint, dissatisfactory, bad… |

### 2.2.2 领域间无歧义共享特征集合的构建

除了消除各情感词典因语料不同而造成的情感词差异和极性分歧,我们还需确定在源领域和目标领域间共享的特征集合,以及各领域的专有特征集合. 通过构建领域间无歧义共享特征集合,为实现领域间专有特征词对的提取和领域间信息分布空间的对齐建立基础.

在信息论中,MI 技术通常是用来描述 2 个变量之间的关联关系.在文献[8]中,同样采用 MI 的方法来衡量 2 个领域中的特征词与领域间的关联关系.如果 1 个特征词与领域有较高的 MI 值,则认为该词是领域的专有特征词,否则认为该词是领域共享特征词.所以本文也采用同样的方法进行领域间共享特征词和专有特征词的选取.

$$\psi(\omega;D)=\sum_{d\in D}p(\omega,d)\ \mathrm{lb}\left(\frac{p(\omega,d)}{p(\omega)p(d)}\right),\quad(1)$$

其中,$D$ 是领域的集合,包含源领域和目标领域;$\omega$ 是在 2 个领域中均有出现的待评估特征词;$p(\omega,d)$ 是特征词 $\omega$ 和领域 $d$ 的联合概率;$p(\omega)$是领域 $d$ 中包含特征词 $\omega$ 的文档个数与领域中的总文档个数的比例;$p(d)$是领域 $d$ 在领域集合 $D$ 中的概率.$\psi(\omega;D)$的值越小,特征词 $\omega$ 越有可能是领域共享特征词.所以,可以通过设置经验参数 $l$,选取前 $l$ 个按 MI 值升序排列的特征词作为领域间共享特征集合 $W_{\mathrm{sh}}^{\mathrm{MI}}$.

利用 MI 进行领域间共享特征集合构建时,仅考虑到特征词在各个领域中和各领域间的出现频率,所以会导致所选择的共享特征集合中包含有极性分歧的情感词.比如,在书籍和电子产品领域,用 MI 的方法,情感词“easy”会被选为共享特征.但是“easy”在 2 个领域中存在极性分歧:在书籍领域中更多地表达了书籍过于简单的消极情感,是负极性;在电子产品领域中更倾向于表达使用便捷、操作简单的积极情感,是正极性.为了消除通过 MI 提取的共享特征集合中情感词的极性分歧,本文将结合 2.2.1 节中构建的无极性分歧的情感词集合,完成领域间无歧义共享特征集合的构建,确保所选择作为桥梁进行领域专有特征词对提取的特征极性的唯一性.

通常在语料中能反映文本情感倾向的特征词被称为情感词,情感词的词性大多数为形容词,所以为了确保领域间情感词的极性唯一,构建领域间无歧义共享特征集合 $W_{\mathrm{sh}}$的基本准则为:如果特征词$\omega\in W_{\mathrm{sh}}^{\mathrm{MI}}$,则词性标注为形容词(JJ),并且 $\omega\notin W_{\mathrm{sh}}^{\mathrm{dict}}$,该特征词会被剔除;否则特征词 $\omega$ 会被保留.

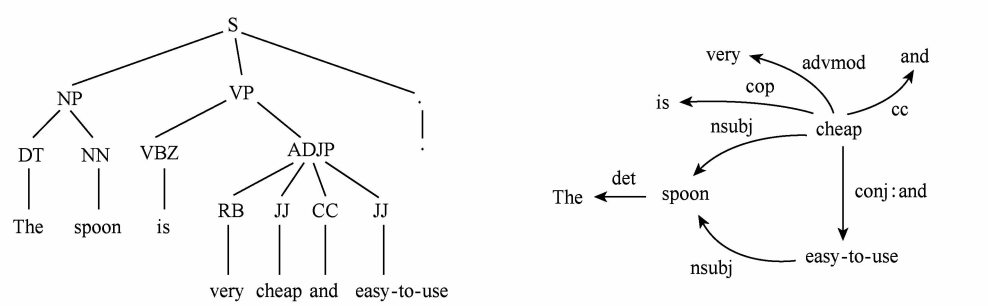
### 2.3 领域间专有特征词对的提取

2.2 节通过情感词典和 MI 的方法构建领域间无歧义共享特征集合,并提取领域专有特征.本节我们将通过 2 种方法以 2.2 节构建的领域间无歧义共享特征为桥梁进行专有特征词对的提取,实现领域词典的扩展和领域间统一特征空间构建.

#### 2.3.1 基于语法规则进行情感词对提取

1) 相同极性的情感词对挖掘

受文献[22-24]的启发,通常可通过 4 条规则来挖掘未标定样本中情感词的极性关系:①情感词间用连词“and”,“or”,“as well as”相连,并且没有否定词修饰时,可以推断 2 个情感词可能具有相同的情感极性.比如:句子“The spoon is very cheap and easy-to-use.”中的情感词“cheap”和“easy-to-use”在修饰“spoon”时可以推断它们具有相同的情感极性.②情感词在没有否定词修饰和连词相连的情况下并列出现来描述同一对象时,可以推断它们可能具有相同的情感极性.比如:句子“It is a beautiful,durable,convenient table lamp.”中的情感词“beautiful”,“durable”,“convenient”通常具有相同的情感极性.③情感词用连词“but”,“however”相连并且没有否定词修饰时,可以推断 2 个情感词可能具有相反的情感极性.比如:句子“This book is very beautiful but too easy for me.”中的情感词“beautiful”和“easy”的描述对象都是“book”,但是用转折词“but”相连,它们可能表达了相反的情感极性.④情感词并列出现或者用“and”,“or”,“as well as”连词相连来描述同一对象,但是有否定词修饰时,可以推断它们可能




(a) Syntactic parsing tree (b) Dependency parsing and sentiment polarity extraction

Fig. 2 An illustrative example of extracting sentiment polarity relations based on syntactic parsing and dependency parsing

图 2 基于句法和依存关系解析的情感词极性关系提取的示意图

具有相反的情感极性. 比如:句子“The battery of this camera is small and not durable.”中的情感词“small”和“durable”用连词“and”相连但是“durable”用否定词“not”作修饰,所以它们可能具有相反的情感极性. 图 2 中用句法解析树和依存关系分析,具体说明了利用上面的①～④条规则从评论语句中提取情感词极性关系的方法.

仅通过一条符合 4 条规则的评论语句进行情感词间极性关系判别时,会出现误判的可能. 比如,在电子产品领域中有一条评论语句为“This product has good and bad points.”根据判别规则,由于情感词“good”和“bad”用连词“and”相连,同时没有否定词修饰,可以判断它们具有相同的情感极性,但是这明显是一种误判的情况. 所以,本文将通过结合 2 个情感词在领域的整个语料中,基于 4 条规则所提取到的情感极性关系来减少误判的可能性. 具体的 2 个特征词的情感极性关系判别如式(2):

$$PR(\omega_i,\omega_j)=\frac{n_s-n_o}{n_s+n_o}, \tag{2}$$

其中,$n_s$ 和 $n_o$ 分别代表特征词 $\omega_i$ 和 $\omega_j$ 在语料中基于语法规则所提取的相同和相反情感极性关系的频率. 当 $PR(\omega_i,\omega_j)>0$ 时,说明特征词 $\omega_i$ 和 $\omega_j$ 有相同的情感极性;否则,特征词间有相反的情感极性. 本节需要提取所有具有相同情感极性关系的情感词对,所以暂不考虑所有 $PR(\omega_i,\omega_j)<0$ 的词对.

2) 领域间相同极性的专有情感词对提取

根据从目标领域和源领域中挖掘到的具有相同极性的情感词对,并以 2.2 节提取的领域间无歧义共享特征集合为桥梁,进行领域间专有情感词对提取的描述如算法 1.

利用 4 条规则在语法解析的基础上进行领域中相同极性的情感词对的挖掘和领域间相同极性的专有情感词对的提取,其结果通常有较高的准确率. 但是,由于语料中能满足 4 条规则可以进行极性关系挖掘的评论语句所占的比例非常小,比如在句子“What an amazing product for such a small price.”中,虽然可以推断出情感词“amazing”和“small price”有相同的极性,但是并不能通过 4 条规则进行极性关系的挖掘. 所以该方法不能挖掘到所有包含在领域语料中具有相同极性的情感词,也不能对领域间中性特征词的关系进行挖掘,故仅通过语法解析进行领域间情感词对的提取无法构建领域间统一特征空间. 为了弥补这种不足,我们提出了第 2 种基于关联规则提取领域间特征词对的方法.

### 2.3.2 基于关联规则进行特征词对提取

1) 强关联规则集的挖掘

关联规则算法的主要思想是通过统计分析,挖掘事物之间的联系. 较常用的是 Apriori 算法,通过最小支持度来进行频繁项集的挖掘,并利用频繁项集和最小置信度来挖掘强关联规则. 本节将通过该算法挖掘领域中特征词间的强关联关系,并以领域间无歧义的共享特征集合为桥梁,提取领域间有强关联关系的专有特征词对,实现领域间统一特征空间的构建.

首先,记 $D=\{D_s,D_t\}$ 为领域集合,包括源领域 $D_s$ 和目标领域 $D_t$,$W$ 为源领域和目标领域的特征词集合,即

$$\begin{aligned}W=W_{sp}\cup W_{sh}=W_{sp^t}\cup W_{sp^s}\cup W_{sh}=\\ \{\omega_{sp^t}^1,\omega_{sp^t}^2,\cdots,\omega_{sp^t}^n,\omega_{sp^s}^1,\omega_{sp^s}^2,\cdots,\\ \omega_{sp^s}^m,\omega_{sh}^1,\omega_{sh}^2,\cdots,\omega_{sh}^l\},\end{aligned} \tag{3}$$

其中,专有情感特征集合 $W_{sp}$ 由源领域的专有特征词集合 $W_{sp^t}$ 和目标领域的专有特征词集合 $W_{sp^s}$ 组成;$W_{sh}$ 为 2 个领域共享的无歧义特征词集合;$n$,$m$,$l$ 分别表示目标领域、源领域专有特征词数量及领域间无歧义共享特征词的数量.

**算法 1.** 领域间相同极性的专有情感词对挖掘算法.

输入:目标领域中相同极性的情感词对集合 $Pair_t=\{(\omega_i,\omega_j)_k\}_{k=1}^{n_t}$、源领域中相同极性的情感词对集合 $Pair_s=\{(\omega_i,\omega_j)_k\}_{k=1}^{n_s}$、领域间无歧义共享特征集合 $W_{sh}$、目标领域专有特征集合 $W_{sp^t}$、源领域专有特征集合 $W_{sp^s}$;

输出:专有情感特征词对 $Couple^1=\{(\omega_i,\omega_j)_k\}_{k=1}^{m}$,$\omega_i\in W_{sp^t}$,$\omega_j\in W_{sp^s}$ 或 $\omega_j\in W_{sp^t}$,$\omega_i\in W_{sp^s}$.

① for each $(\omega_i,\omega_j)$ in $Pair_t$
② if $\omega_i\in W_{sh}$ and $\omega_j\in W_{sp^t}$
$SET1.add((\omega_i,\omega_j))$;
else $\omega_j\in W_{sh}$ and $\omega_i\in W_{sp^t}$
$SET1.add((\omega_i,\omega_j))$;
③ end if
④ end for
⑤ for each $(\omega_i,\omega_j)$ in $Pair_s$
⑥ if $\omega_i\in W_{sh}$ and $\omega_j\in W_{sp^s}$
$SET2.add((\omega_i,\omega_j))$;
⑦ else $\omega_j\in W_{sh}$ and $\omega_i\in W_{sp^s}$
$SET2.add((\omega_i,\omega_j))$;
⑧ end if
⑨ end for

⑩ for each $(\omega_i,\omega_k)$ in $SET1$ and $(\omega_k,\omega_j)$ in $SET2$

⑪ if $\omega_k\in W_{\text{sh}}$

$Couple^1.add((\omega_i,\omega_j))$;

⑫ end if

⑬ end for

⑭ return $Couple^1$.

Apriori 算法的主要思想是通过 $k$ 项频繁集的先验知识和最小支持度 $min\_s$ 来生成 $k+1$ 项频繁集,并根据最小置信度 $min\_c$ 完成强关联关系的挖掘.本文将通过 Apriori 算法进行各领域中特征词间强关联关系的挖掘.其中,用 $item^1$ 表示生成的 1 项频繁集,$item^2$ 表示生成的 2 项频繁集.其中 $item^1$ 和 $item^2$ 表示为

$$item^1=\{\omega^i\},\omega^i\in W;$$

$$item^2=\{(\omega^i_{\text{sp}},\omega^j_{\text{sh}})\},\omega^i_{\text{sp}},\omega^j_{\text{sh}}\in W.$$

频繁集中的任意元素 $it$ 的支持度都大于最小支持度 $min\_s$,支持度计算为

$$support(it)=P(it), \tag{4}$$

其中,$it\in item^1$ 或 $it\in item^2$,$P(it)$表示 $it$ 在样本集中出现的概率.

在 2 项频繁集中找到满足最小置信度 $min\_c$ 并且由一个领域共享特征词和一个领域专有特征词构成的强关联规则,强关联规则 $r_k$ 的挖掘和置信度计算为

$$r_k=\omega^i_{\text{sh}}\Rightarrow\omega^j_{\text{sp}}; \tag{5}$$

$$confidence(r_k)=confidence(\omega^i_{\text{sh}}\Rightarrow\omega^j_{\text{sp}})=$$

$$P(\omega^j_{\text{sp}}\mid\omega^i_{\text{sh}})=\frac{P(\omega^j_{\text{sp}},\omega^i_{\text{sh}})}{P(\omega^i_{\text{sh}})}\geqslant min\_c, \tag{6}$$

其中,$r_k\in RS$,$RS$ 是强规则集;规则的置信度为一个领域共享特征词 $\omega^i_{\text{sh}}$ 和领域专有特征词 $\omega^j_{\text{sp}}$ 的条件概率值.具体的强关联规则集的挖掘算法描述如算法 2.

**算法 2.** 强关联规则集挖掘算法.

输入:最小支持度 $min\_s$、最小置信度 $min\_c$、领域间无歧义共享特征词集合 $W_{\text{sh}}$、单个领域的评论集合 $Reviews=\{(\{\omega^i\}_{i=1}^{n_{\text{sent}}})_j\}_{j=1}^{n_{\text{dom}}}$,单个领域专有特征词集合 $W_{\text{sp}}$;

输出:强规则集 $RS=\{r_k\}_{k=1}^{n_{RS}}$.

① $L1=find\_frequent\_1\_itemsets(W),\omega^i\in W$;

② $L2\_candidate=apriori\_gen(L1,min\_s)$;

③ for $w$ in $Reviews$

$C_w=subset(w)$;

④ for $c$ in $C_w$

$c.count++$;

⑤ end for

⑥ end for

⑦ $L2=\{c\},c\in L2\_candidate$ and $c.count/\sum c.count\geqslant min\_s$;

⑧ for $r$ in $L2$

⑨ if $(support\_count(r)/support\_count(r.\omega_{\text{sh}}))\geqslant min\_c$

⑩ $RS.add(r)$;

⑪ end if

⑫ end for

⑬ return $RS$.

2) 领域间强关联关系的专有特征词对提取

根据从源领域和目标领域中挖掘到的各领域中的专有特征词和领域间共享特征词的强关联关系,来进行 2 个领域间具有强关联关系的专有特征词对的提取.具体操作可通过构造有向图 $G=\{W_{\text{sp}}\cup W_{\text{sh}},E\}$来进行,如图 3 所示.在图 3 中,每一个顶点对应着一个共享特征词(白色圆圈)或领域专有特征词(灰色圆圈),顶点间的有向箭头表示共享特征词和领域专有特征词间通过 Apriori 算法挖掘到了强关联关系.同时,用灰白顶点间的垂直距离来可视化特征词间关联关系的强弱,距离越大表示关联关系越强,否则,关联关系越弱.比如,在图 3 中顶点 $\omega^1_{\text{sh}}$ 和顶点 $\omega^1_{\text{sp}^s}$ 间的垂直距离 $d^{11}$ 比顶点 $\omega^1_{\text{sh}}$ 和顶点 $\omega^2_{\text{sp}^s}$ 间的垂直距离 $d^{12}$ 小,说明 $\omega^1_{\text{sh}}$ 和 $\omega^1_{\text{sp}^s}$ 间的关联关系比 $\omega^1_{\text{sh}}$ 和 $\omega^2_{\text{sp}^s}$ 间的关联关系弱.在本文中,特征词对的关联程度用置信度的值来衡量,具体的计算如式(8).

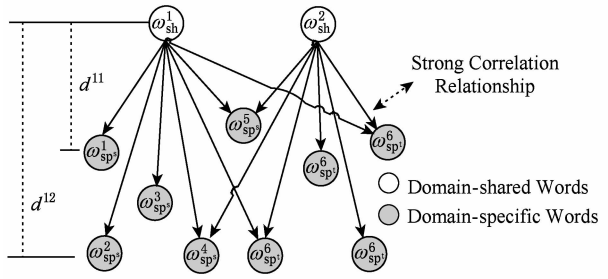



Fig. 3 Directed graph G: the description of strong correlation relationship between domain-shared words and domain-specific words

图 3 有向图 $G$:描述领域共享词和专有词的强关联关系

在图谱理论中,如果 2 个顶点与多个公共顶点相连,则这 2 个顶点相似或者存在较强的关联关系.即如果 2 个领域专有特征词 $\omega^i_{\text{sp}}$ 和 $\omega^j_{\text{sp}}$ 与多个领域

共享特征词相连,并且它们与同一共享特征的关联程度相似,那么这 2 个领域专有特征词也存在某种关联关系. 比如在图 3 中源领域专有特征词 $\omega_{\mathrm{sp^s}}^{5}$ 和目标领域专有特征词 $\omega_{\mathrm{sp^t}}^{9}$ 均与领域共享特征词 $\omega_{\mathrm{sh}}^{1}$ 和 $\omega_{\mathrm{sh}}^{2}$ 有强关联关系,且它们的关联程度比较相近,即:

$$confidence(\omega_{\mathrm{sh}}^{1}\Rightarrow\omega_{\mathrm{sp^s}}^{5})\approx confidence(\omega_{\mathrm{sh}}^{1}\Rightarrow\omega_{\mathrm{sp^t}}^{9}),\quad(7)$$

$$confidence(\omega_{\mathrm{sh}}^{2}\Rightarrow\omega_{\mathrm{sp^s}}^{5})\approx confidence(\omega_{\mathrm{sh}}^{2}\Rightarrow\omega_{\mathrm{sp^t}}^{9}),\quad(8)$$

所以 2 个领域专有特征词 $\omega_{\mathrm{sp^s}}^{5}$ 和 $\omega_{\mathrm{sp^t}}^{9}$ 间也存在关联关系.

由于在实际应用中,挖掘到的强关联规则的关联程度不可能完全相同,所以本文采用近似相等的策略,故引入误差阈值参数 $\varepsilon$. 当 2 个强关联规则的信度的差值小于给定阈值 $\varepsilon$ 时,则认为 2 条关联规则的关联程度是相似的. 即当式(11)满足时,可提取 $(\omega_{\mathrm{sp^s}}^{a},\omega_{\mathrm{sp^t}}^{b})$ 为领域专有特征词对.

$$|confidence(\omega_{\mathrm{sh}}^{i}\Rightarrow\omega_{\mathrm{sp^s}}^{a})-confidence(\omega_{\mathrm{sh}}^{i}\Rightarrow\omega_{\mathrm{sp^t}}^{b})|\leqslant\varepsilon,\ i\in(0,n_{\mathrm{c}}),\quad(9)$$

其中,$n_{\mathrm{c}}$ 表示领域专有特征 $\omega_{\mathrm{sp^s}}^{a}$ 和 $\omega_{\mathrm{sp^t}}^{b}$ 共同相连的领域共享特征词 $\omega_{\mathrm{sh}}$ 的个数. 同时得到用该方法提取的领域间有强关联关系的专有特征词对用 $Couple^{2}$ 表示.

### 2.4 领域间统一特征空间的构建和分类模型的训练

本节将结合在 2.3.1 节和 2.3.2 节中提取的领域间相同极性的专有情感特征词对和领域间强关联关系的专有特征词对进行领域间统一特征空间的构建,具体构建过程如图 4 前 4 层所示. 并利用源领域标定样本的统一特征表示来训练跨领域情感分类模型,如图 4 中的层 5、层 6.

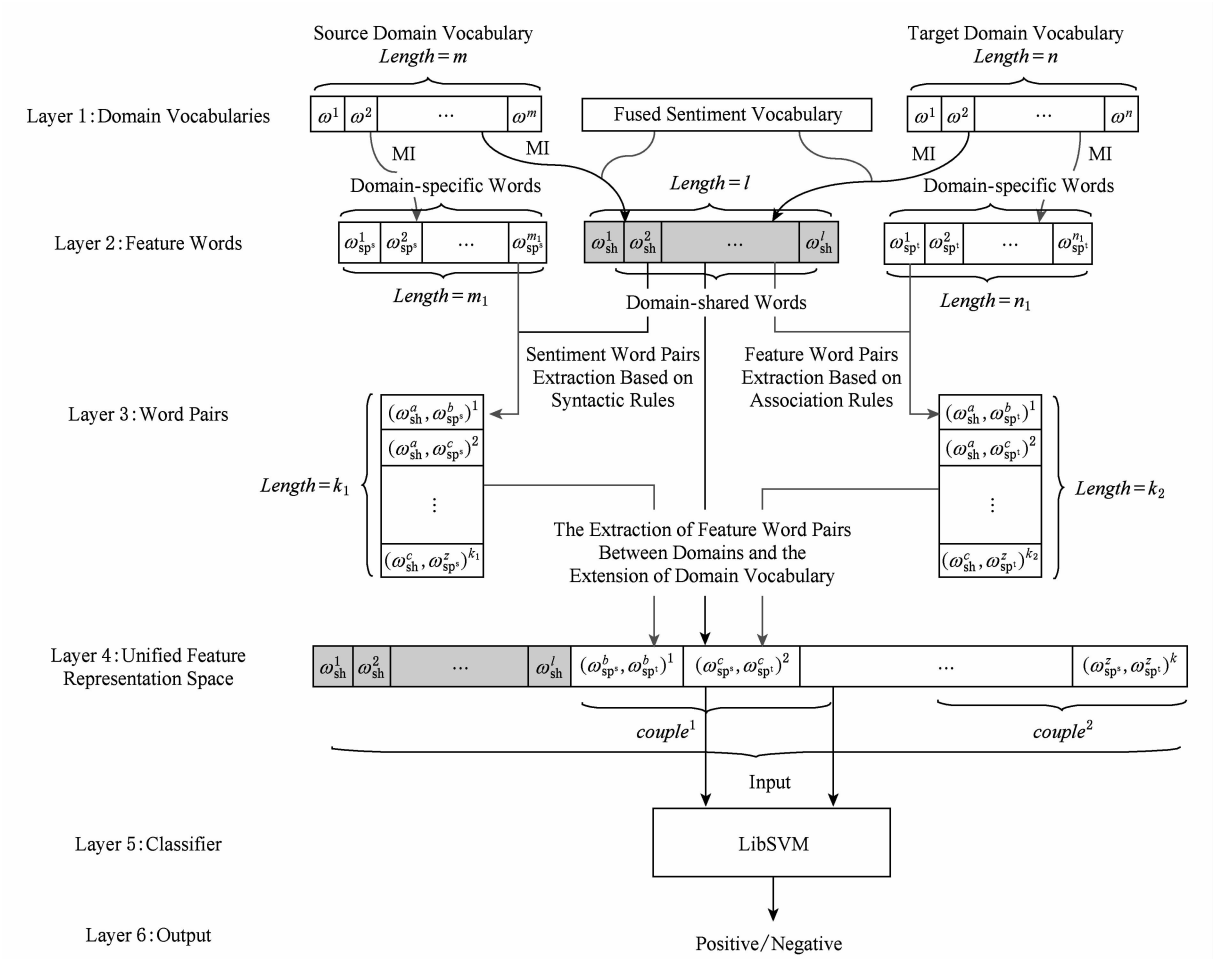



Fig. 4 An illustrative example of training the cross-domain sentiment classifier

图 4 训练跨领域情感分类器的示例图

在扩展领域词典,构建领域间统一特征空间后,目标领域和源领域每条评论语句 $X_{\mathrm{sent}}^{j}=\{\omega^{i}\}_{i=1}^{n_{\mathrm{sent}}}$ 都可被映射到由领域共享无歧义情感特征和领域间专有特征词对构成的特征空间中. 对于语料中的每条评论语句具体可表示为

$$\widetilde{X}_{\mathrm{sent}}^{j}=(\omega_{\mathrm{sh}}^{1},\cdots,\omega_{\mathrm{sh}}^{i},\cdots,\omega_{\mathrm{sh}}^{l},\cdots,pair^{1},\cdots,pair^{j},\cdots,pair^{k}),\quad(10)$$

其中,$\omega_{\mathrm{sh}}^{i}\in W_{\mathrm{sh}}$,$l$ 为领域间无歧义共享特征的数量,

$pair^j \in Couple^1 \cup Couple^2$,$k$ 为专有特征词对的数量,新的特征维度为 $l+k$. 即可通过领域间无歧义共享特征词和专有特征词对的提取,实现领域词典的扩展以及源领域和目标领域统一特征空间的构建. 在一定程度上消除了领域间情感词的极性分歧和信息分布空间差异的问题.

在此基础上,利用源领域中的标定样本在领域间统一特征空间的映射,完成跨领域情感分类器的训练. 具体分类器的选择,在第 3 节中选择 LibSVM[25] 作为跨领域情感分类器,其中参数均为默认参数.

## 3 实验设计与结果分析

本节使用本文提出构建领域间统一特征空间的方法,消除领域间情感词的极性分歧和信息空间分布的差异,完成跨领域情感分类器的训练,并且在 Amazon 产品评论数据集[9] 测试了我们的方法.

### 3.1 数据集概述和预处理

在实验中,所采用的数据集是由 Blitzer[9] 收集的 Amazon 产品评论数据集,是被广泛应用在跨领域情感分类的基准数据库. 数据库中包含了 4 个领域的英文评论数据,分别是 B(Book),D(Dvds),E(Electronics)和 K(Kitchen)领域. 每个领域中均有 2 000 条标定评论,其中 1 000 条是积极评论,1 000 条是消极评论和若干条未标定评论. 积极评论的情感标签为+1,消极情感标签为−1. 表 3 是对 Amazon 产品评论数据集的详细描述.

**Table 3 Detailed Description of Amazon Data Sets Used for Experiments**

**表 3 Amazon 产品评论数据集的详细描述**

| Domains | Positive Reviews | Negative Reviews | Unlabeled Reviews |
|---|---|---|---|
| B | 1 000 | 1 000 | 973 194 |
| D | 1 000 | 1 000 | 122 438 |
| E | 1 000 | 1 000 | 21 009 |
| K | 1 000 | 1 000 | 17 856 |

各领域的特征词集合都是由每条评论语句在去除停用词、词干提取、小写转换后的一元和二元混合语法形式组成. 如“i_love”,“worth”,“right”,“a_great”等. 同时,Pang 等人[3] 用实验证明了采用所有词作为特征,并且用一个词出现与否作为权重,比使用词出现的频率作为权重,可以达到更好的情感分类效果. 因此,在本文中也采用布尔值作为特征权重,即如果特征在评论语句中出现则权重为 1,否则为 0. 同时,对于用于构建统一特征空间中的特征词对,只要其中一个在评论语句中出现,则权重为 1,否则为 0.

### 3.2 实验设计和结果分析

为了验证本文所提出的基于多视角共享特征的领域空间对齐模型对跨领域情感分类的有效性,本文将 4 个领域的产品评论语料组成了 12 个跨领域任务:D→B,D→E,D→K,B→D,B→E,B→K,E→B,E→D,E→K,K→B,K→D,K→E,其中箭头左侧表示源领域,箭头右侧表示目标领域. 在领域空间对齐阶段采用的是 2 个领域所有的样本;在分类器训练阶段也就是图 4 中 5,6 层,我们使用 LibSVM[25] 作为跨领域情感分类器,其中参数均为默认参数. 源领域中消极评论和积极评论各 800 条构为训练数据,用目标领域消极评论和积极评论各 200 条进行测试. 实验涉及的超参数依次设置为:领域间无歧义共享特征个数 $l=600$,最小支持度 $min\_s=0.014$,最小置信度 $min\_c=0.08$,关联度阈值 $\varepsilon=0.005$. 为了避免实验结果的偶然性,我们对每个实验独立重复进行 5 次,并取平均值作为最终的跨领域情感分类的准确率. 选择以下 6 种算法进行对比实验.

1) NoTransf. 不进行领域空间对齐,在源领域数据集上训练 LibSVM[25] 分类器,直接在目标领域数据集上测试.

2) SCL[9]. 由 Blitzer 提出的结构对应学习算法进行跨领域情感分类.

3) SFA[8]. 由 Pan 等人提出的光谱对齐算法进行跨领域情感分类.

4) LP-based[16]. 由 Li 等人提出的基于图排序的算法,实现情感标签从源领域到目标领域的传播,实现跨领域情感分类.

5) DAMF(Single). 在 DAMF 算法中仅通过互信息进行领域间共享特征的选择,并仅通过关联规则(Apriori)算法实现领域空间的对齐.

6) DAMF. 本文提出的基于多视角共享特征的领域空间对齐的跨领域情感分类模型.

6 种算法的实验结果比较如表 4 所示.

在表 4 中我们可以看出:

1) 无论哪一种方法,任务 E→K 和 K→E 的结果均优于其他 10 项任务,这表明 Electronics 领域与 Kitchen 领域相较于其他领域的相关性较大.

2) 5种跨领域情感分类的算法几乎在所有子任务中均优于NoTransf,这表明在跨领域情感分类任务中,充分利用源领域和目标领域的样本来实现样本层面和特征层面的对齐,有助于提高分类的准确率.

3) DAMF与DAMF(Single)相比平均准确率提高0.42%,说明从多视角提取领域间共享特征,有助于消除共享特征词的极性分歧,并以共享特征为桥梁,通过2种方式提取领域中相同极性的情感词对和强关联关系词对,更有助于消除领域间信息分布的差异,实现领域空间的对齐,更有利于跨领域情感分类.

4) 在任务B→E,D→K,E→K中,本文所提算法DAMF的准确率略低于SCL和SFA,说明在一些情况下,基于语法规则和关联规则,不能提取到潜在的强关联关系,无法实现领域空间的对齐,使跨领域情感分类的准确率得到提升.

5) 总体上,DAMF与SCL,SFA,LP-based跨领域情感分类算法相比,在9个任务上的准确率均有提高,平均准确率达到了78.7%,说明通过以无歧义共享特征为桥梁挖掘领域专有特征间的关联关系,有助于消除领域信息分布的差异,实现跨领域情感分类.

**Table 4 Accuracy on 12 Subtasks of 6 Cross-Domain Sentiment Classification Algorithms**

**表4 6种跨领域算法在12个跨领域任务上的准确率**

| Algorithms | B→D | B→E | B→K | D→B | D→E | D→K | E→B | E→D | E→K | K→B | K→D | K→E | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NoTransf | 76.8 | 71.28 | 74.45 | 73.35 | 72.69 | 75.02 | 72.42 | 71.25 | 79.02 | 71.8 | 73.05 | 80.05 | 74.27 |
| SCL | 78.5 | **75.22** | 77.08 | 78.26 | 74.2 | **78.94** | 75.02 | 75.25 | 85.06 | 72.78 | 76.65 | 85.04 | 77.67 |
| SFA | 80.54 | 72.1 | 78.02 | 77.54 | 76.04 | 79.5 | 75.4 | 75.5 | **85.95** | 74.2 | 76.7 | 85.02 | 78.04 |
| LP-based | 79.8 | 72.0 | 77.0 | 78.0 | 76.8 | 74.3 | 71.5 | 74.0 | 83.5 | 73.5 | 73.3 | 81.5 | 76.27 |
| DAMF(Single) | 81.25 | 73.5 | 76.9 | 79.6 | 77.6 | 77.85 | 75.3 | 75.8 | 85.3 | 74.3 | 76.9 | 85.05 | 78.28 |
| DAMF | **81.76** | 74.23 | **78.3** | **80.04** | **77.8** | 78.37 | **75.67** | **76.3** | 85.32 | **74.43** | **77.1** | **85.09** | **78.70** |

Notes: The bold value in each cross-domain subtask means the best value.

为了进一步验证本文所提算法的有效性,我们分别计算了各算法在知识传递过程中的传递损失,结果如图5所示.传递损失的计算公式为

$$t(D_s,D_t)=e(D_s,D_t)-e(D_t,D_t), \quad (13)$$

其中,$e(D_s,D_t)$表示采用领域空间对齐策略后,用源领域样本训练得到分类器,在目标领域测试时产生的误差;$e(D_t,D_t)$表示以目标领域的标定样本训练分类器,并以目标领域的样本进行测试所产生的误差.$t(D_s,D_t)$表示采用跨领域情感分类所变化的传递误差.

由图5可看出在12个子任务中,不进行领域知识传递的NoTransf方法的传递损失最大.同时,在其中7个子任务中,相较于其他跨领域算法,本文所提出的基于领域间无歧义共享特征词为桥梁,实现领域空间对齐的传递损失最小.在子任务K→E中,除不进行知识传递的NoTransf和LP-based算法,其他跨领域算法均出现传递损失为负的情况,说明电子产品领域的评论数据分布可能与厨房用品的评

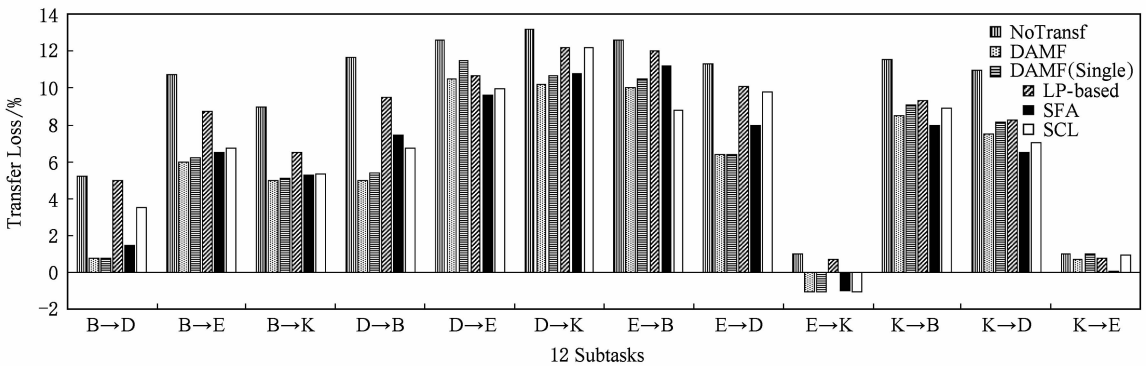


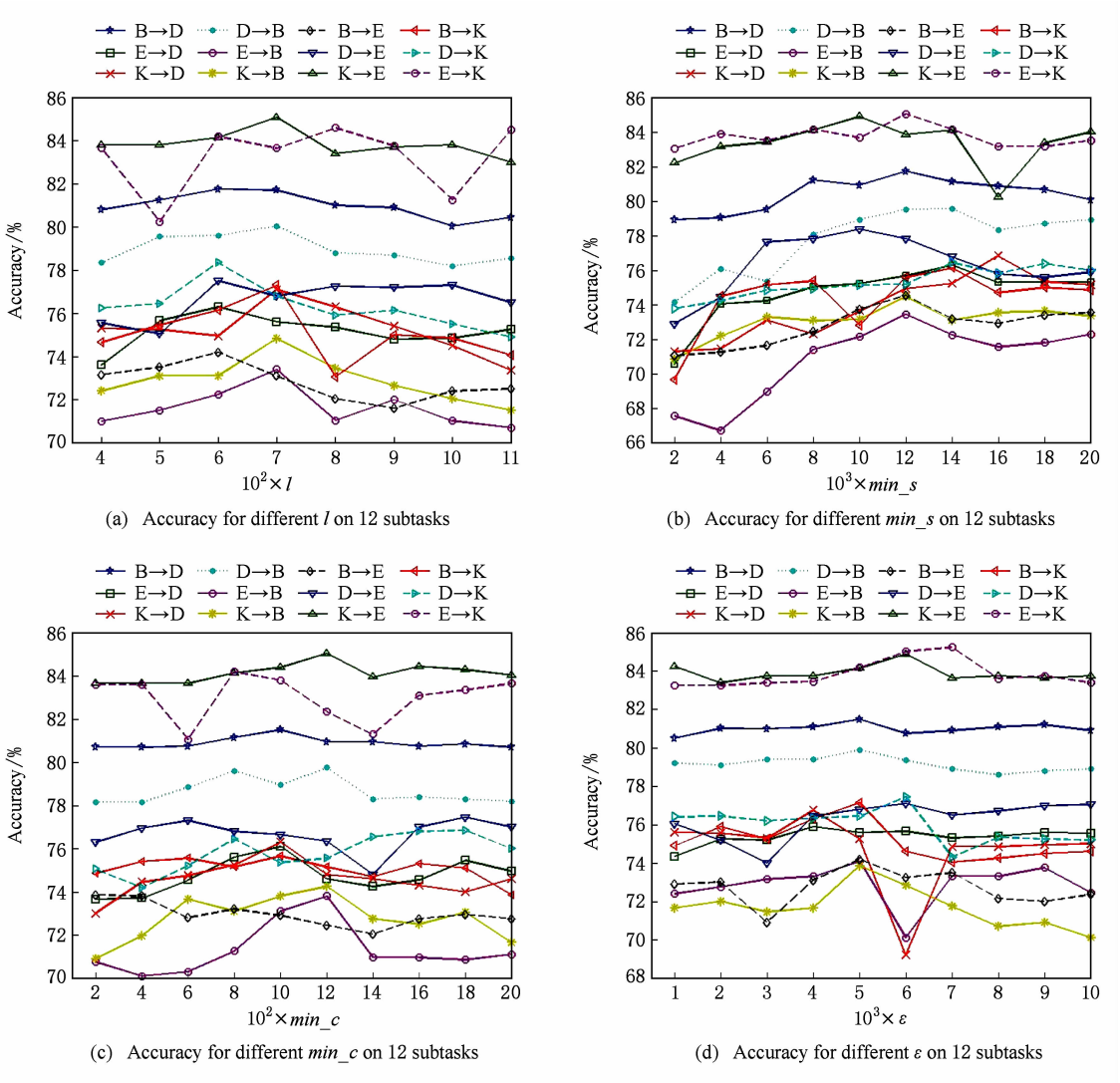

Fig. 5 The transfer loss of 6 cross-domain sentiment classification algorithms on 12 subtasks

图5 6种算法在12个跨领域任务上的传递损失率

论数据分布相似,但由于源领域的标定样本更丰富,所以导致预测的准确率更高,传递损失为负值.

### 3.3 参数分析

本节中,我们将分别进行实验来探索在 2.2 节和 2.3 节中所涉及的 4 个参数:$l$,$min\_s$,$min\_c$,$\varepsilon$,在不同取值情况下对 12 个跨领域分类任务准确率的影响.实验结果如图 6 所示,它们分别代表这 4 个参数在不同取值时对准确率的影响.




(a) Accuracy for different $l$ on 12 subtasks

(b) Accuracy for different $min\_s$ on 12 subtasks

(c) Accuracy for different $min\_c$ on 12 subtasks

(d) Accuracy for different $\varepsilon$ on 12 subtasks

Fig. 6 Effect of four parameters value on the accuracy of experiments

图 6 4 个参数的取值对实验准确率的影响

在图 6(a)中,领域间无歧义共享特征词个数 $l$ 的取值为 400～1 100,步长为 100;并观察到当 $l$ 的取值范围在 500～700 之间时,实验结果的准确相对较高.当 $l$ 取值较小时,部分可以作为领域间共享特征词被丢失,导致相应的关联关系丢失,所提取的词对数量减少,所以实验准确率相对较低;同时当 $l$ 取值较大时,部分与领域相关度较大的特征词会被误选为共享特征,导致无法消除领域间的差异,使实验的准确率降低.

在图 6(b)和图 6(c)中,通过设置最小支持度 $min\_s$ 和最小置信度 $min\_c$ 来进行关联规则的学习,发现最适合的参数取值使实验的准确率达到最高.从图 6(b)看出,设置 $min\_s$ 为 0.002～0.02,步长为 0.002 进行实验.当 $min\_s$ 的取值范围在 0.008～0.016 之间时,有利于进行频繁项集的挖掘,使分类的准确率较高;当 $min\_s<0.008$ 时,部分领域间共享特征和专有特征不会被选为频繁项集,则导致部分规则被丢失,使准确率降低.同时从图 6(c)可看出,通过设置 $min\_c$ 为 0.02～0.2,步长为 0.02 进行实验.当 $min\_c$ 取值范围从 0.06～0.12 之间时,分类器的准确率较高;当 $min\_c$ 取值较大时,由于较多的无关规则被保留,所以对于领域空间

对齐产生了干扰,准确率较低;当 $min\_c$ 取值较小时,部分有用规则会被丢弃所以使某些词对未能提取,领域空间不能对齐,也会使准确率降低.

在图 6(d)中,通过设置关联度阈值 $\varepsilon$ 为 0.001～0.01、步长为 0.001 来进行分类器准确率的分析.从图 6(d)中可以看出,虽然某些取值会使实验的准确率波动较大,但当 $\varepsilon$ 取值为 0.005 或 0.007 时,对各任务分类的准确率都相对较高.

## 4 结 论

本文针对跨领域情感词存在特征分布不一致性而导致的识别率低问题,提出了多视角共享特征提取和挖掘策略,在建立统一特征表示空间基础上,实现了跨领域情感分类,提升了分类准确率,降低了知识传递的损失。已完成的主要创新工作和下一步工作如下:

1) 本文提出的基于多视角共享特征的领域空间对齐的跨领域情感分类算法充分利用了现有的情感词典并结合特征词的互信息值进行领域间无歧义共享特征词的提取.以无歧义共享特征词为桥梁,利用语法规则提取到的相同极性情感词对和关联规则算法学习到的领域中有强关联关系的特征词对,建立领域间专有特征词的映射关系,构建领域数据的统一特征表示空间.实现了共享特征词中歧义情感词的消除和领域空间的对齐,提升了跨领域情感分类的准确性.

2) 本文工作中无论是通过特征互信息值来进行领域共享特征和专有特征的选择,还是利用语法规则和关联规则来进行相同情感词对和有强关联关系的特征词对的提取,均可在未标定样本集上进行,降低了对各个领域中标定样本的依赖,减少了标注样本所需的人力物力.扩大了算法在各个领域上的适用性,降低了对训练样本的依赖,提升了算法的泛化性能.

3) 本文所提算法仅以有共现关系的领域无歧义共享特征词为桥梁,完成领域间专有特征词的映射.所以当 2 个领域间信息分布差距较大、共现的特征词较少、挖掘到的领域间特征词对较少时,无法实现领域空间对齐,如任务 B→K 和 D→E.所以未来的研究工作将同时考虑如何利用多个源领域的语料来辅助单个目标领域的情感分类问题,以及如何充分利用各领域中的未标定数据基于数据驱动挖掘领域间潜在的关联关系,完成领域公共特征空间的学习.

## 参 考 文 献


[1] Balazs J A, Velasquez J D. Opinion mining and information fusion: A survey [J]. Information Fusion, 2016, 27(C): 95-110

[2] Wang Shaopeng, Peng Yan, Wang Jie. Research of the text clustering based on LDA using in network public opinion analysis [J]. Journal of Shandong University: Natural Science, 2014, 49(9): 129-134 (in Chinese)
(王少鹏, 彭岩, 王洁. 基于 LDA 的文本聚类在网络舆情分析中的应用研究[J]. 山东大学学报: 理学版, 2014, 49(9): 129-134)

[3] Pang Bo, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques [C] //Proc of the EMNLP'02. Stroudsburg, PA: ACL, 2002: 79-86

[4] Kushal D, Steve L, David M P. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C] //Proc of the 12th Int Conf on World Wide Web. New York: ACM, 2003: 519-528

[5] Go A, Bhayani R, Huang Lei. Twitter sentiment classification using distant supervision, CS224N Project Report [R]. Palo Alto: Stanford, 2009: 1-6

[6] Zhuang Li, Jing Fei, Zhu Xiaoyan. Movie review mining and summarization [C] //Proc of the 15th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2006: 43-50

[7] Wu Qiong, Tan Songbo, Xu Hongbo, et al. Cross-domain opinion analysis based on random-walk model [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)

[8] Pan Sinnojialin, Ni Xiaochuan, Sun Jiantao, 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

[9] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning [C] //Proc of the EMNLP'06. Stroudsburg, PA: ACL, 2006: 120-128

[10] Pang Bo, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2): 1-135

[11] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification [C] //Proc of the 2007 Annual Meeting Association for Computational Linguistics. Stroudsburg, PA: ACL, 2007: 187-205

[12] Li Lianghao, Jin Xiaoming, Long Mingsheng. Topic correlation analysis for cross-domain text classification [C] //Proc of the 26th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2012: 998-1004

[13] Liu Kang, Zhao Jun. Cross-domain sentiment classification using a two-stage method [C] //Proc of the 18th ACM Conf on Information and Knowledge Management. New York: ACM, 2009: 1717-1720

[14] Dai Wenyuan, Yang Qiang, Xue Guirong, et al. Boosting for transfer learning [C] //Proc of the 24th ICML. New York: ACM, 2007: 193-200

[15] Hu Kongbing, Zhang Yuhong, Hu Xuegang. A multi-domain adaptation for sentiment classification algorithm based on class distribution [C] //Proc of the 2012 IEEE Int Conf on Granular Computing. Piscataway, NJ: IEEE, 2012: 179-184

[16] 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

[17] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 513-520

[18] Pan Sinnojialin, Yang Qiang. A survey on transfer learning [J]. Journal IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359

[19] Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining [C] //Proc of the 5th Conf on Language Resources and Evaluation. Paris: European Language Resources Association (ELRA), 2006: 417-422

[20] Hu Minqing, Liu Bing. Mining and summarizing customer reviews [C] //Proc of the 10th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168-177

[21] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis [C] //Proc of the 2005 Conf on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2005: 347-354

[22] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives [C] //Proc of the 35th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 1997: 174-181

[23] Huang Sheng, Niu Zhendong, Shi Chongyang. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation [J]. Knowledge Based Systems, 2014, 56(C): 191-200

[24] Wu Fangzhao, Huang Yongfa. Sentiment domain adaptation with multiple sources [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 301-310

[25] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27:1-27:27

**Jia Xibin**, born in 1969. PhD and professor. Her main research interests include visual information cognition, text classification, and multi-information fusion, especially for facial expression recognition.

**Jin Ya**, born in 1994. Master in Beijing University of Technology. Her main research interests include text classification and machine learning (jinya@emails.bjut.edu.cn).

**Chen Juncheng**, born in 1980. Post doctor and lecturer. His main research interests include software testing and temporal-spatial database.