基于逐步优化分类模型的跨领域文本情感分类

张 军[1] 王素格[1,2]
(山西大学计算机与信息技术学院 太原 030006)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]

摘 要 跨领域文本情感分类已成为自然语言处理领域的一个研究热点。针对传统主动学习不能利用领域间的相关信息以及词袋模型不能过滤与情感分类无关的词语,提出了一种基于逐步优化分类模型的跨领域文本情感分类方法。首先选择源领域和目标领域的公共情感词作为特征,在源领域上训练分类模型,再对目标领域进行初始类别标注,选择高置信度的文本作为分类模型的初始种子样本。为了加快目标领域的分类模型的优化速度,在每次迭代时,选取低置信度的文本供专家标注,将标注的结果与高置信度文本共同加入训练集,再根据情感词典、评价词搭配抽取规则以及辅助特征词从训练集中动态抽取特征集。实验结果表明,该方法不仅有效地改善了跨领域情感分类效果,而且在一定程度上降低了人工标注样本的代价。

关键词 情感分类,跨领域,分类模型,特征抽取,置信度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.042

Cross-domain Sentiment Classification Based on Optimizing Classification Model Progressively

ZHANG Jun[1] WANG Su-ge[1,2]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]

Abstract Cross-domain sentiment classification has attracted more attention in natural language processing field. Given that tradition active learning can't make use of the public information between domains and the bag of words model can't filter these words not related with sentiment classification, a method of cross-domain sentiment classification based on optimizing classification model progressively was proposed. Firstly, this paper selected the public sentiment words as features to train classification model on the labeled source domain, then used the classification model to predict the initial category label for target domain and selected the texts with high confidence value as initial seed texts of the learning model. Secondly, we added the high confidence text and low confidence text to the training set at each iteration. Finally, the feature set was extracted to transform feature space based on the sentimental dictionary, evaluation collocation rules and assist feature words. The experimental results indicate that this method can not only improve the accuracy of cross domain sentiment classification effectively, but also reduce the manual annotation price to some extent.

Keywords Sentiment classification, Cross domain, Classification model, Feature extraction, Confidence

1 引言

随着 Web2.0 的迅速发展,网上出现了大量的多领域评论数据。文本情感分类是高效快速地挖掘多领域评论中有价值的信息的有效方法之一[1]。文本情感分类[2]即自动评判用户所表达的情感极性。当训练数据集和测试数据集属于不同领域时,使用单个领域的文本情感分类[3]方法往往不能获得理想的分类效果。随着评论文本的不断增加、新兴领域的不断出现,需要人工标注的训练集越来越多,这将会导致人工消耗巨大,由此跨领域的情感分类问题引起人们越来越多的关注。

主动学习[4]作为一种有效减少文本标注的学习方法,已成为机器学习、数据挖掘等领域的研究热点,并大量应用于信息抽取、句法分析、文本分类等领域,本文将动态变换特征空间融合到主动学习算法中,用于优化分类模型。陈霄[5]利用支持向量机模型将主动学习应用到组织机构名的识别中,并取得了一定效果;车万翔等[6]将主动学习应用到依存句法分析中,优先选择句法模型预测不准的实例交由人工标注,提升

到稿日期:2015-04-06 返修日期:2015-11-12 本文受国家自然科学基金资助项目(61175067,61272095,60875040),国家"八六三"高技术研究发展计划基金项目(2015AA015407),山西省科技攻关项目(20110321027-02),山西省回国留学人员科研项目(2013-014),山西省科技基础条件平台建设项目(2015091001-0102)资助。

张 军(1990-),男,硕士生,主要研究方向为情感分析;王素格(1964-),女,教授,博士生导师,主要研究方向为自然语言处理、情感分析等,E-mail:wsg@sxu.edu.cn(通信作者)。

了中文依存分析性能。Simon Tong 等[7]利用支持向量机模型将主动学习应用到文本分类上，从整个训练集中挑选一部分最具代表性的数据来训练分类器，取得了较好的分类效果。

李寿山等[8]采用词袋模型对文本进行表示，将主动学习应用到跨领域的文本情感分类中，但是该方法采用的词袋模型往往加入部分与情感分类无关的特征，在一定程度上影响了分类模型的分类效果。在文本情感分类中，具有情感倾向性的词语对于情感分类尤为重要，一个领域的倾向性词语主要包括领域相关情感词和领域无关情感词两类，例如：在“书籍”领域和“笔记本”领域，都出现了包含“高兴”、“难过”等与领域无关的情感特征词，而有些情感词与领域相关，例如：“蓝屏”通常只出现在笔记本领域。因此，本文利用源领域与目标领域无关的情感词以及动态挖掘与目标领域相关的情感词作为特征，以主动学习思想为基础，提出一种基于模型优化的跨领域文本情感分类方法。该方法首先利用互信息绝对差值（Absolute Difference of Mutual Information，ADMI）获取与领域无关的情感词，获得目标领域文本的初始类别标签，在此基础上，再选择目标领域中置信度较高的文本作为主动学习的初始种子样本。为了加快传统主动学习模型在目标领域的优化速度并增强模型的泛化能力，在主动学习每次迭代过程中，不仅将低置信度的文本交给专家标注，还将高置信度的文本一同加入训练集。根据情感词典、评价词搭配抽取规则以及辅助特征词从训练集中动态抽取特征。在中文情感评论语料上进行实验，验证了该方法在减少一定人工标注量的情况下还可以获得较为理想的跨领域文本情感分类结果。

2 相关研究

早期的文本情感分类研究主要集中在单一领域，但是随着数据量的快速增长，人们逐渐意识到跨领域研究的重要性。当前，国内外对跨领域文本情感分类的研究主要从特征迁移和实例分析展开研究。

2.1 特征迁移的跨领域文本情感分类

特征迁移的研究思路是搜索源领域和目标领域的公共特征，用于构建一个统一的跨领域的特征空间。Blitzer 等人[9]提出结构对应学习（Structure Correspondence Learning，SCL）方法，该方法利用源领域和目标领域部分枢纽特征，构建枢纽特征与非枢纽特征的关联模型，在此基础上，构建基于枢纽特征和非枢纽特征的特征空间，用于文本的情感分类。刘康等人[10]首先基于主题的特征翻译模型，构建源领域和目标领域数据的特征空间，然后通过源领域标记数据训练分类器，找出目标领域中富含信息量的样本，最后利用这些有用信息训练适应目标领域的分类模型。张玉红等人[11]提出一种面向跨领域情感分类的特征选择方法，利用对数似然比选取在原始领域富有判别力的特征，并通过对照两个领域的统计信息，选出在目标领域中影响较大的特征，该方法构建的公共特征空间能减少领域间数据分布的差异。魏现辉等人[12]提出基于加权 SimRank 的跨领域文本倾向性分析模型，该方法利用具有领域独立性且带有明显情感极性的枢纽特征作为桥梁，构建枢纽特征与非枢纽特征的二部图，通过加权 SimRank 算法获得潜在的特征空间，用于对文本进行情感倾向性分析。

2.2 实例迁移的跨领域文本情感分类

实例迁移从样本层面解决跨领域文本情感分类问题，其研究思路是以源领域的标记数据为依据，选取目标领域中对分类有价值的实例，用于辅助分类模型的训练。谭松波等人[13]试图利用旧领域标记样本训练分类器，用于目标领域未标记的样例的分类；然后挑选出高可信度的样本，并用这些样本重新训练分类器。Jiang 等人[14]为了减小领域数据分布差异，首先利用目标领域的信息对源领域数据进行评估，剔除源领域中的“噪音数据”。在训练分类器时，赋予目标领域标记数据较高的权重，赋予源领域标记数据较低的权重，用于预测未加标签的数据。Dai 等人[15]推广了传统的 AdaBoost 算法，提出一种具有迁移能力的 Boosting 算法 TrAdaboost，利用 Boosting 的技术过滤掉源数据中那些与目标数据差异最大的数据，其作用是建立一种自动调整权重的机制，相似源数据的权重在训练过程中将会增加，不相似的源数据的权重将会减小。赵传君等人[16]首先标注少量的目标领域文本，采用 SMOTE 和 BootStrapping 方法获得一定数量的目标领域中带标签的数据，然后把源领域带标签数据等量分割，将其中的每一份与目标领域高置信度数据组合，采用 AdaBoost 的方法提升训练分类器，最后将每个模块的分类器进行集成，得到最终的分类器。Liao 等[17]提出 Migratory-Logit 方法，通过引入一个阈值来评估源领域数据样本对目标分类的贡献程度，结合主动学习的方法挑选出最优的一部分源数据，对这部分数据进行建模，得到适用于目标领域数据的分类模型。

3 初始种子选择方法

针对跨领域文本情感分类问题，为了获得目标领域分类模型的初始种子样本，首先利用互信息绝对差值（Absolute Difference of Mutual Information，ADMI）构建领域公共情感词特征集；然后采用委员会投票策略，在源领域包含公共情感词的数据集上训练多个源领域分类模型，得到目标领域数据初始类别标签；最后从中挑选预测结果一致且置信度高的文本作为分类模型的初始种子样本，其具体流程如图 1 所示。

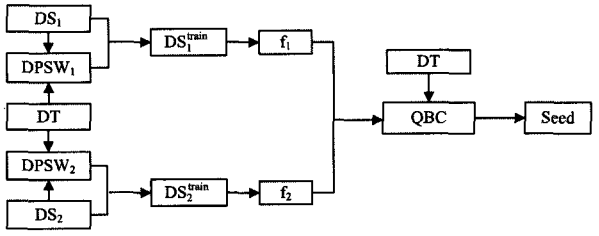

图 1 基于领域公共特征的初始样本选择流程

图 1 中的标记符号说明如下：

DS_i——源领域 i 文本数据集；

DS_i^{train}——源领域包含领域公共情感词的文本数据集；

$DPSW_i$——领域公共情感词（Domain Public Sentiment Words），即在源领域 i 与目标领域都出现且带有明显情感倾向性的词语；

f_i——源领域 i 分类模型；

DT——目标领域无类别标签文本数据集；

$Seed$——主动学习初始种子样本集。

图 1 中的初始样本选择过程由领域公共特征选择过程(Absolute Difference of Mutual Information, ADMI)和委员会投票机制(Query by Committee, QBC)构成。

3.1 领域公共特征选择算法

文本情感分类往往具有领域依赖性,仅通过源领域建立特征空间并不能满足目标领域文本情感分类的要求,因此从源领域挑选出适合目标领域的特征尤为重要。这些特征词不仅需要在源领域和目标领域具有较高的频率,而且需要在源领域和目标领域具有相似的分布。为了获取这些领域公共特征,构建互信息绝对差值,如式(1)所示。

$$ADMI(t)=|MI(t,DS)-MI(t,DT)| \quad (1)$$

其中,$MI(t,DS)$,$MI(t,DT)$分别表示特征 t 对源领域和目标领域的互信息值。$ADMI(t)$越小,该词的领域独立性越高。

根据互信息绝对差值,设计领域公共特征选择算法如下。

输入:源领域数据集 DS={DS_1, DS_2},目标领域数据集 DT,情感词典 SD(Sentiment Dictionary)

输出:DS 相应的领域公共情感词集 DPSW={$DPSW_1$, $DPSW_2$}

1. 初始化 DPSW={$DPSW_1$, $DPSW_2$};//集合 DPSW 中存储领域公共情感词特征集;
2. for t in SD
3. for DS_i in DS
4. if t∈DS_i and t∈DT and numDs(t)+numDt(t)>3//判断情感词 t 属于源领域和目标领域,并且 t 出现的次数大于 3
5. ADMI(t)=|MI(t,DS)−MI(t,DT)|//计算 t 的互信息绝对差值
6. 将 t 加入 $DPSW_i$
7. end if
8. end for
9. Sort($DPSW_i$);//对 $DPSW_i$ 中的情感词按 ADMI(t)值从小到大进行排序
10. 挑选 $DPSW_i$ 中前 200 个领域公共情感词作为 DS_i 最终领域公共情感词集;
11. end for
12. 返回 DPSW

3.2 多领域集成的样本选择算法

不同领域之间的数据分布差异将不同程度地影响跨领域文本情感分类的效率,如图 1 所示,文本通过多领域集成的方法解决领域之间的不适应性问题,利用委员会投票机制 QBC 选择样本,其具体算法如下。

输入:源领域数据集 DS={DS_1, DS_2},相应的领域公共情感词集 DPSW={$DPSW_1$, $DPSW_2$},目标领域数据集 DT

输出:分类模型初始种子样本集 Seed

1. 初始化 Seed=∅;
2. 初始化 DS^{train}={DS_1^{train}, DS_2^{train}};//DS^{train}存储源领域包含领域公共特征的训练数据集;
3. 初始化 f={f_1, f_2};//相应的源领域分类模型
4. for DS_i in DS
5. 从 DS_i 中挑选出包含 $DPSW_i$ 特征项的文本构成 DS_i^{train};
6. 用 DS_i^{train} 训练源领域分类模型 f_i;
7. 用 f_i 预测 DT 的初始类别标签;
8. end for
9. 采用委员会投票策略从 DT 中挑选出各 f_i 预测一致的文本加入 Seed,其中置信度取均值;
10. Sort(Seed);//将 Seed 中的文本按置信度从大到小进行排序;
11. 挑选 Seed 中前 N 个文本作为最终的初始种子样本集并默认其标记正确;
12. 返回 Seed

4 特征选择方法

领域情感词特征对于文本情感分类具有重要的意义,通过情感词典可以获得一定规模的领域无关情感词集,但是与领域相关的情感词集需要构建情感搭配规则并从语料中挖掘。因此,本文将领域相关的情感搭配、领域无关情感词、连词、否定词、程度词融合共同构成分类特征。

情感词典选择大连理工情感词汇本体[18]中的褒义词和贬义词。

连词特征主要由转折连词和承接连词两部分构成,其中表转折的连词有“但是”、“尽管”等,其后面的句子或子句的情感倾向与前面相反;表承接的连词有“并且”、“而且”等,其后面句子或子句的情感倾向与前面一致。

否定词、程度词特征通常和评价词一起使用来表达对其他对象的评价。程度词(比如“非常”、“时常”等)有加强语气的作用,是对其所表达观点的进一步肯定,但是否定词往往能够决定句子的情感倾向程度,比如“酒店价格在当地来讲,不算实惠”这句话中否定副词“不算”使得句子的倾向性与极性词“实惠”的倾向性完全相反,因此加入了此类词汇作为特征,以进一步提高文本情感分类的效果。

情感搭配采用 4 种句法依存关系:主谓关系(SBV)、动宾关系(VOB)、状中关系(ADV)、定中关系(ATT)[19]。本文利用哈工大的句法依存分析工具对语料进行句法依存分析。

5 基于逐步优化的分类算法

在分类模型优化的过程中,本文采用动态变换特征空间的方法。该方法利用情感词典、评价词搭配抽取规则以及辅助特征词对新加入样本的训练集进行动态特征抽取,挖掘出与领域相关的情感词,从而对分类模型不断地优化,使其在分类性能上得到提升。

在选择新样本时,采用双重信息样本抽取策略,即挑选低置信度样本(支持向量)和高置信度文本作为新样本。由于支持向量位于超平面附近,分类模型中缺乏该类文本的特征,加入该类文本对于分类模型的优化尤为重要,但是该类文本相对较少,不能很好地刻画目标领域的数据分布。因此,通过加入高置信度样本、充实训练样本、优化特征空间来提高系统的泛化能力。基本流程如图 2 所示。

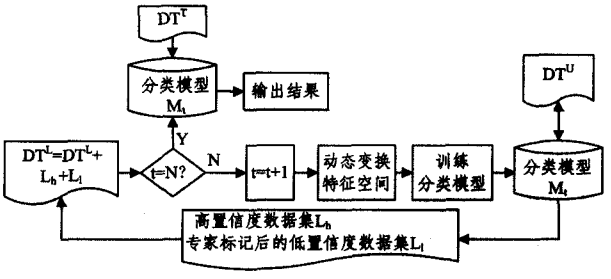

图 2 基于逐步优化的分类算法流程

图 2 中的标记符号说明如下：

DT^L——目标领域已标注数据集；

DT^U——目标领域未标注数据集；

DT^T——目标领域测试数据集；

M_t——目标领域迭代第 t 次的分类模型；

L_h——高置信度数据集；

L_l——专家标注后的低置信度数据集；

t——迭代次数。

根据图 2 的算法流程，基于逐步优化分类模型的算法(Active-Semi-Dynamic)描述如下。

输入：目标领域标记数据集 DT^L，目标领域未标记数据集 DT^U，目标领域测试数据集 DT^T

输出：目标领域测试数据集 DT^T 的分类结果

1. 设定迭代次数为 N；
2. for t=1 to N do
3. 依据情感词典、评价词搭配抽取规则以及辅助特征词对更新后的 DT^L 进行特征选择，动态地变换特征空间；
4. 在 DT^L 上训练分类模型 M_t；
5. 利用分类模型 M_t 对 DT^U 进行预测，按置信度从高到低进行排序；
6. 从排序后的 DT^U 中挑选出 L_h 和 L_l；
7. $DT^L=DT^L+L_h+L_l$；//双重信息样本抽取策略，将 L_h 和 L_l 添加到 DT^L
8. end for
9. 利用分类模型 M_N 对 DT^T 进行预测，获取分类结果。

6 实验数据与预处理

6.1 实验数据

本文实验所使用的语料来自于谭松波发布的中文情感语料库，该语料库包括 3 个领域的中文评论，分别是酒店(Hotel)、书籍(Book)和笔记本电脑(Notebook)，其中每个领域包括正、负评论文本各 2000 篇。实验中，选择 3 个领域的其中 2 个领域作为源领域，另外 1 个作为目标领域。

6.2 预处理及评价指标

针对上述实验数据进行预处理：使用哈尔滨工业大学的 LTP 平台对数据集进行分词以及句法分析，为减小噪音，降低特征维度，本文去除停用词和频率小于 3 的特征。

实验过程中，使用向量空间模型表示每个文本，采用布尔值(Boolean value)作为特征权重，布尔值为 1 时表明文本包含该特征，否则反之。

本文所用到的分类器为台湾大学林智仁等人开发设计的 lib-SVM。在该分类器的参数设置中，除了将核函数设置为线性核函数外，其它参数设置为默认值。

本文以准确率(Accuracy)和人工标注量作为实验结果的最终评价指标。为方便下面的讨论，使用符号“$X\rightarrow Z$”表示由源领域 X 向目标领域 Z 的跨领域情感分析，“$(X,Y)\rightarrow Z$”表示源领域 X 和源领域 Y 联合向目标领域 Z 的跨领域情感分析。

7 实验结果与分析

7.1 传统的 SVM 监督分类方法

传统的 SVM 监督分类方法直接利用训练样本训练 SVM 分类模型，然后对目标领域测试集进行情感分类测试。在该实验部分，源领域和目标领域数据属于同一领域时，将该领域数据分成 5 份，其中 4 份作为训练集，1 份作为测试集；源领域和目标领域数据属于不同领域时，将源领域全部数据作为训练集，目标领域全部数据作为测试集，实验结果如表 1 所列。

表 1 源领域到目标领域的分类准确率(%)

源领域	目标领域		
	Book	Hotel	Notebook
Book	86.875	51.02	54.25
Hotel	60.15	82.15	74.93
Notebook	62.90	74.55	87.625

由表 1 可知：在 Book、Hotel、Notebook 3 个评论数据集上，当源领域和目标领域属于同一领域时，文本情感分类准确率分别是 82.15%，87.625%，86.875%，均达到最高。当源领域和目标领域属于不同领域时，情感分类效果明显降低。这表明传统的监督分类方法适用于单领域的情感分类问题，而不适用于跨领域文本情感分类问题。

7.2 初始种子选择数量对分类模型优化结果的影响

本实验在 50 到 190 范围内改变参数 K，每次增加 20，考察参数初始种子数量对分类模型优化结果的影响。模型优化过程中，当迭代次数超过 20 次时，实验结果基本趋于稳定。因此，选择迭代 25 次时，模型优化结束，结果如图 3 所示。

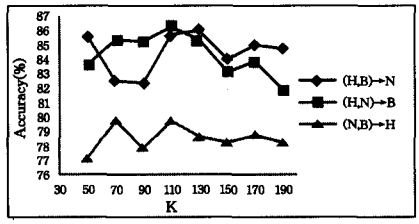

图 3 不同 K 值下分类模型的分类效果

由图 3 可知，不同领域的曲线尽管变化趋势有所差异，但是总体上呈现先波浪式再下降的趋势。当参数 K 小于 90 时，曲线呈波浪形变化，这表明初始种子较少时，这部分样本不能准确表示该领域的数据分布情况，分类模型不稳定；当参数 K 超过 110 时，曲线总呈现下降趋势，这表明初始种子较多时，初始种子类别标签准确率偏低，迭代过程中错误放大，导致分类效果不好；当参数 K 取 110 时，3 个领域都取得较好的分类效果。因此，本文设定初始种子数量为 110。

7.3 多领域集成策略对初始种子准确率的影响

本实验基于领域公共情感词，分别计算$(X,Y)\rightarrow Z$，$X\rightarrow Z$以及 $Y\rightarrow Z$ 初始种子样本的准确率，考察多领域集成对初始种子准确率的影响。由图 3 可知，参数 K 为 110 时，模型优化结果最好，因此，在初始种子数量取 110 时，将单领域和多领域集成得到的初始种子准确率进行对比，实验结果如图 4 所示。

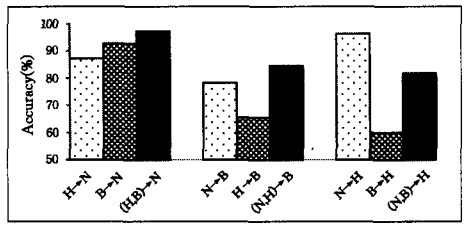

图 4 单领域和多领域集成初始种子准确率

从图 4 可以看出：

1)在基于领域公共情感词的跨领域情感分类中，除 $N\rightarrow H$ 外，两个领域联合得到的种子样本标记准确率明显高于单个领域的，平均准确率提高了 7.88 个百分点。这表明相比单领域，多领域集成的方法减小了领域之间的不适应性，提高了分类模型初始种子的准确率。

2)在 Hotel 中，虽然 $(N,B)\rightarrow H$ 初始种子的准确率低于 $N\rightarrow H$，但是 $(N,B)\rightarrow H$ 初始种子的准确率低于 $B\rightarrow H$，这主要是由于 Book 领域与 Hotel 领域的数据分布差异较大，限制了多领域集成选择初始种子的准确率，表明跨领域研究中，不同领域之间数据分布差异不宜过大，否则影响文本情感分类的效果。

3)3 个领域中，多领域集成选择的初始种子准确率都达到了 81%以上，$(H,B)\rightarrow N$ 甚至达到了 97.27%的准确率，这表明基于领域公共情感词选择初始种子的方法是有效的，保证了下一步分类模型的分类效果。

7.4 3 种分类模型的跨领域文本情感分类结果比较

为了验证本文提出的分类模型(Active-Semi-Dynamic)对跨领域文本情感分类的有效性，设置了以下 3 组对比实验，其中，高置信度阈值 $\zeta=0.995$。

Active-Semi-Dynamic：本文提出的分类模型，即利用第 5 节的动态变换特征空间。每次迭代样本的选取采用第 5 节的方法，即选取专家标注的 10 条低置信度样本和高置信度样本加入训练数据集。

Active-Dynamic：特征选取同 Active-Semi-Dynamic 方法。样本的选取采用传统的主动学习方法，即每次迭代选取 10 条低置信度样本交给专家标注后加入训练数据集。

Active-BOW：特征选取采用词袋模型表示文本，剔除词频小于 5 的词。每次迭代选取样本与 Active-Semi-Dynamic 相同。

上述 3 种方法在 Hotel，Book，Notebook 3 种领域的评论的实验结果如图 5－图 7 所示。

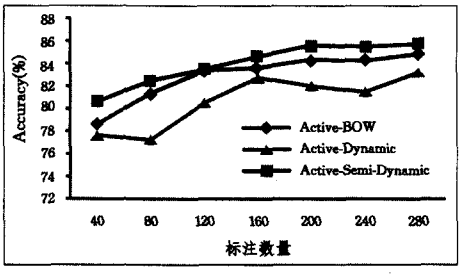

图 5 Notebook 领域分类模型优化效果

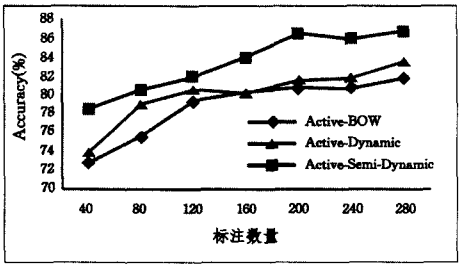

图 6 Book 领域分类模型优化效果

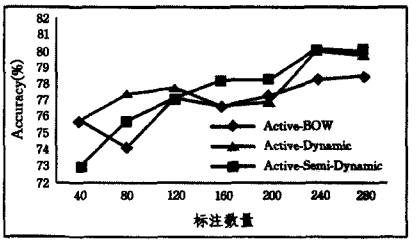

图 7 Hotel 领域分类模型优化效果

从图 5－图 7 可以看出：

1)3 个领域中，Active-Semi-Dynamic 与 Active-Dynamic 相比，第 25 次迭代时，平均精度提高了 2.75 个百分点，表明将高置信度样本加入训练集使得训练样本和特征信息更加丰富，增强了系统的泛化能力，有助于分类模型的训练。

2)3 个领域中，Active-Semi-Dynamic 与 Active-BOW 相比，第 25 次迭代时，平均精度提高了 2.79 个百分点，这表明相比于词袋模型，将情感词典和句法依存分析相结合抽取情感词可以更加准确地刻画文本的情感，融入语义信息来动态变换特征空间能够过滤与文本情感分类无关的词语，改善了跨领域文本情感分类的效果。

结束语 本文基于逐步优化分类模型，提出跨领域文本情感分类方法，其相比传统的主动学习文本情感分类方法，更加充分利用领域间的相关文本信息，减少了很多不必要的冗余标注。该方法首先利用多个领域文本的相关性，分类得到具有公共信息的训练样本，然后通过加入高置信度文本，加入了更多具有领域相关信息的文本，使得特征空间更加全面，从而降低了样本标记的工作量；最后融入语义信息来变换特征空间，过滤掉与情感分类无关的词语，提高了分类效率。实验结果表明本文方法在标注量较少的情况下取得了较好的实验效果，从而说明了该算法的有效性。

分类模型初始种子的选择过程中，本文只用到了两个源领域集成来预测目标领域情感分类，在下一步的研究中，可尝试使用多个领域集成来预测目标领域的情感分类。除此之外，分类模型优化过程中，最优的分类模型仅仅通过迭代次数来决定，下一步需要深入探讨阈值的选择方法，以寻找最优的分类模型。

参 考 文 献

[1] Wang Su-ge, Li De-yu, Wei Ying-jie. 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

[2] Zhao Yan-yan, Qin Bing, Liu Ting. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848(in Chinese)
赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment Classification using Machine Learning Techniques[C]// Proceedings of the Association of Computational Linguistics Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2002: 79-86

[4] Olsson F. A Literature Survey of Active Machine Learning in the Context of Natural Language Processing[R]. Swedish Insti-

tute of Computer Science, 2009

[5] Chen Xiao. Chinese Organization Names Recognition Based on Support Vector Machine[D]. Shanghai: Shanghai Jiao Tong University, 2007(in Chinese)
陈霄. 基于支持向量机的中文组织机构名识别[D]. 上海: 上海交通大学, 2007

[6] Che Wan-xiang, Zhang Mei-shan, Liu Ting. Active Learning for Chinese Dependency Parsing[J]. Journal of Chinese Information Processing, 2012, 26(2): 18-22(in Chinese)
车万翔, 张梅山, 刘挺. 基于主动学习的中文依存句法分析[J]. 中文信息学报, 2012, 26(2): 18-22

[7] Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification[J]. The Journal of Machine Learning Research, 2002, 2(1): 45-66

[8] Li S, Xue Y, Wang Z, et al. Active Learning for Cross-Domain Sentiment Classification[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2013: 2127-2133

[9] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification[J]// ACL, 2012, 31(2): 187-205

[10] Liu K, Zhao J. Cross-Domain Sentiment Classification Using a Two-Stage Method[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1717-1720

[11] Zhang Hong-yu, Zhou Quan, Hu Xue-gang. Feature Selection for Cross-Domain Sentiment Classification[J]. Pattern Recognition and Aitificial Intelligence, 2013, 26(11): 1068-1072(in Chinese)
张玉红, 周全, 胡学钢. 面向跨领域情感分类的特征选择方法[J]. 模式识别与人工智能, 2013, 26(11): 1068-1072

[12] Wei Xian-hui, Zhang Shao-wu, Yang Liang, et al. Cross-Domain Sentiment Analysis Based on Weighted SimRank[J]. Pattern Recognition and Aitificial Intelligence, 2013, 26(11): 1004-1009 (in Chinese)
魏现辉, 张绍武, 杨亮, 等. 基于加权 SimRank 的跨领域文本情感倾向性分析[J]. 模式识别与人工智能, 2013, 26(11): 1004-1009

[13] Tan S, Wu G, Tang H, et al. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis[C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York: ACM, 2007: 979-982

[14] Jiang J, Zhai C X. Instance Weighting for Domain Adaptation in NLP[C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg, PA: ACL, 2007: 264-271

[15] Dai W, Yang Q, Xue G R, et al. Boosting for Transfer Learning [C]// Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon, USA, 2007: 193-200

[16] Zhao Chuan-jun, Wang Su-ge, Li De-yu, et al. Cross-Domain Text Sentiment Classification Based on Grouping-AdaBoost Ensemble[J]. Journal of Computer Research and Development, 2015, 52(3): 629-638(in Chinese)
赵传君, 王素格, 李德玉, 等. 基于分组提升集成的跨领域文本情感分类[J]. 计算机研究与发展, 2015, 52(3): 629-638

[17] Liao X, Xue Y, Carin L. Logistic Regression with an Auxiliary Data Source[C]// Proceedings of the 22nd International Conference on Machine Learning. New York: ACM, 2005: 505-512

[18] Xu Lin-hong, Lin Hong-fei, Pang Yu, et al. Constructing the Affective Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185(in Chinese)
徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185

[19] Chen S, Wang Y. Mining the Emotional Words from Chinese Reviews Based on Part of Speech and Syntax[C]// 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). IEEE, 2012: 1904-1907

(上接第 229 页)

[8] Zhou M, Bao S, Wu X, et al. An unsupervised model for exploring hierarchical semantics from social annotations[C]// Proceedings of the 6th International Semantic Web Conference. 2007: 680-693

[9] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012: 481-492

[10] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th Conference on Computational Linguistics. 1992: 539-545

[11] Ponzetto S P, Strube M. WikiTaxonomy: A Large Scale Knowledge Resource[C]// Proceedings of the 18th European Conference on Artificial Intelligence. 2008, 178: 751-752

[12] Wu F, Weld D S. Automatically refining the wikipedia infobox ontology[C]// Proceedings of the 17th International Conference on World Wide Web. 2008: 635-644

[13] Fellbaum C, et al. WordNet: An electronic lexical database[M]. MIT Press, 1998

[14] Wang H, Wu T, Qi G, et al. On publishing Chinese linked open schema[C]// Proceedings of the 13th International Semantic Web Conference. 2014: 293-308

[15] Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

[16] 百度知道[OL]. [2014-10-11]. http://zhidao.baidu.com

[17] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[R]. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002

[18] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis[C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. 2010: 1606-1611

[19] 网易微博[OL]. [2014-10-11]. http://t.163.com

[20] Zhou Jin, Chen Chao, Yu Neng-hai. Tag Clustering Algorithm Using Object-based Feature Vector[J]. Journal of Chinese Computer Systems, 2012, 33(3): 525-530(in Chinese)
周津, 陈超, 俞能海. 采用对象特征向量表示法的标签聚类算法[J]. 小型微型计算机系统, 2012, 33(3): 525-530

[21] Fernández-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. The Journal of Machine Learning Research, 2014, 15(1): 3133-3181