doi:10.3969/j.issn.1673-4785.2011.06.002

情感文本分类混合模型及特征扩展策略

夏睿,宗成庆
(中国科学院 自动化研究所,北京 100190)

摘 要:针对篇章级别情感文本分类问题,分析了传统的生成式模型和判别式模型的性能,提出了一种级联式情感文本分类混合模型以及句法结构特征扩展策略.在该模型中,生成式模型(朴素贝叶斯分类器)和判别式模型(支持向量机)以级联的方式进行组合,旨在消除对于分类临界样本,模型判决置信度不足引起的误差.在混合模型的基础上,提出了一种高效扩展依存句法特征的策略.该策略既提高了系统的正确率,又避免了传统特征扩展方法所带来的计算量增加的问题.实验结果表明,混合模型及特征扩展策略与传统方法相比,在算法准确性和效率上,都有显著的提高.

关键词:文本分类;情感分类;混合模型;特征扩展
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1673-4785(2011)06-0483-06

A hybrid approach to sentiment classification and feature expansion strategy

XIA Rui, ZONG Chengqing
(Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In this paper, focusing on sentiment text classification, the performance of generative and discriminative models for sentiment classification was studied, and a hybrid approach to sentiment classification was proposed. The individual generative classifier (naive Bayes, (NB) and the discriminative classifier (support vector machines, SVM) were merged into a hybrid version in a two-stage process in order to overcome individual drawbacks and benefit from the merits of both systems. On the basis of the hybrid classifier, an efficient strategy of incorporating dependency features was also presented. The strategy not only increases the accuracy of the system, but also avoids the defects of increased computing volume brought by the traditional feature expansion method. Experimental results show the apparent advantages of this approach in both classification accuracy and efficiency.

Keywords: text classification; sentiment classification; hybrid model; feature expansion

近10年来文本分类成为自然语言处理和模式识别领域的一个研究热点.传统的文本分类技术关注的是文本的客观内容,如文本主题.基于主题的文本分类技术已有多年的研究基础,发展较为成熟并且得到了广泛应用[1];而情感文本分类所研究的对象是文本的主观内容,如作者的倾向度,近年来逐渐发展成为一种独特的文本分类任务,国内外都有着广泛的研究[2].

情感文本分类的相关研究主要围绕下面2个问题进行:1)设计合适的分类器模型;2)寻找能够有效体现情感信息的特征表示方法.

对于问题1),情感文本分类沿袭了传统的主题文本分类模型,常见的分类器有朴素贝叶斯模型(NB)、支持向量机(SVM)和最大熵模型(MaxEnt).文献[3]对这3种分类器在情感文本分类任务中的性能进行了比较,实验结果显示在电影评论语料(Cornel movie-review dataset)中SVM表现最好,MaxEnt次之,NB最后,不过三者之间的差距并不显著.然而后续研究表明,分类器的性能具有领域依赖性,对不同的领域而言,任何一个分类器性能都无法始终占优[4],例如在多领域情感分类语料(multi-domain sentiment dataset)中,NB性能要优于SVM.因此,对于情感文本分类,生成式模型和判别式模型孰

收稿日期:2011-05-12.
基金项目:国家自然科学基金项目资助项目(60975053);中科院-爱丁堡皇家学会交流项目.
通信作者:夏睿. E-mail:rxia@nlpr.ia.ac.cn.

优孰劣,一直是一个难以回答的问题.

对于问题2),传统的文本分类方法基于词袋模型(bag-of-words, BOW)进行文本表示,以单个词作为特征的基本单元.情感分类有别于主题分类,它需要在特征中体现更多的情感信息,因此,很多研究者立足于挖掘文本中更多能够有效表达情感的信息作为新的特征,如词序及其组合信息[3,5]、词性(part-of-speech, POS)信息[6-8]、高阶 n 元语法(n-gram)[3-4]等,但是这些特征所达到的效果并不明显.也有学者尝试挖掘更深层次的文本信息,比如句法结构信息等[9-11],以期捕捉更加复杂的语法及语义特征(包括否定、转折等),这些方法在一定程度上超过了基于词袋的传统方法,但是系统性能的提高仍然有限.同时,引入句法特征所带来的最大问题就是特征空间的急剧增加,以及分类任务计算量的指数级增加.因此,如何更加有效地利用句法结构特征也是一个亟待解决的难题.

立足于解决上述2个问题,提出了一种基于生成式和判别式模型融合的情感文本分类方法.生成式和判别式分类器以一种级联的方式进行结合,旨在利用判别式模型消除生成式模型对分类临界处样本的判决置信度不高引起的误差.此外,遵循"奥卡姆剃刀"(Occam's razor)原则,在二级判别式分类器上,只对部分临界样本进行特征向量扩展,引入句法结构特征,目的在于向难于分辨的样本中加入更多的情感信息,同时又回避了将所有样本都进行向量扩展所带来的计算量的增加.

1 传统模型

情感文本分类任务的主流分类方法是基于机器学习的统计模型.从建模本质加以区别,可以分为生成式模型(generative model)和判别式模型(discriminative model)2种.

生成式模型对特征和类别的联合概率进行建模,然后利用贝叶斯公式计算后验概率.这一类以贝叶斯决策理论为核心的分类器称作贝叶斯分类器,它们是理论上的最优分类器.其中,朴素贝叶斯分类器(NB)假设了在给定类别的条件下,各个特征项之间相互独立(条件独立性假设),大大简化了类条件概率密度的估计,是一种最简单的生成式模型,在文本分类任务中被广泛应用[12].然而,NB的假设条件过强,在样本的特征相关性较大的情况下,分类性能往往得不到保证.

判别式模型则直接对后验概率进行建模,通常依据一定的准则从样本数据中训练模型参数.支持向量机(SVM)[13]是文本分类任务中常用的判别式模型.SVM的基本思想一是寻找具有最大类间距离的决策面,二是将低维不可分问题转化为高维可分问题,并且通过核函数在低维空间计算并构建分类面.然而,SVM分类器存在容易过学习的缺点,而且,在特征的独立性条件满足较好的情况下,性能不如贝叶斯模型.

2 生成式/判别式混合模型

2.1 NB:足够的置信度?

通过对NB和SVM这2类模型的错误性分析发现,错分样本的分布是交叉的,一部分NB错分的样本SVM可以正确划分,反之亦然.在Kitchen语料(将在4.1节中详细介绍)中抽出100个正例样本和100个负例样本,计算归一化对数联合概率,作分布图,如图1所示,图中加号(+)和点号(·)分别表示2类判决值 $\log(p(\boldsymbol{x}, +))$ 和 $\log(p(\boldsymbol{x}, -))$,虚线左边表示正例样本,虚线右边表示负类样本.

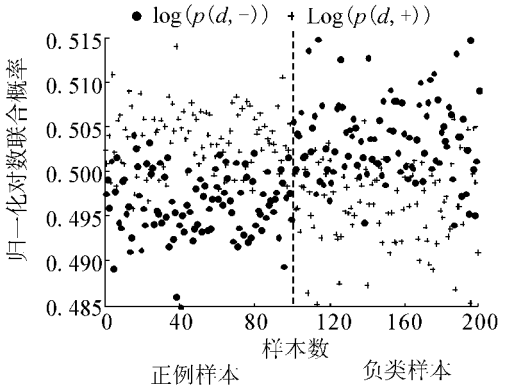


图1 2类归一化对数联合概率分布

Fig.1 Distribution of log-likelihood by NB

由图1可见,纵坐标在0.5附近的水平中轴区域,2类判决值非常接近,正负2类样本中出现错分的比率较大;离水平中轴越远,错分比率越小.给定文档 $\boldsymbol{x}$,为了衡量2类判决的置信度,定义2类归一化对数联合概率距离作为刻画分类判决置信度的一个指标:

$$\mathrm{dist}(\boldsymbol{x}) = \frac{|\log(p(\boldsymbol{x}, +)) - \log(p(\boldsymbol{x}, -))|}{|\log(p(\boldsymbol{z}, +)) + \log(p(\boldsymbol{x}, -))|}.$$

对相同的样本作概率距离分布曲线,如图2所示.与正确划分的样本相比,错分样本的概率距离统计上更加接近0.虽然也有一部分概率距离接近0的样本也被正确划分,认为它们的置信度仍然不高,这样的判决带有很大风险.

用水平线(如图2中纵坐标为0.004的直线)表示置信度阈值,通过设置一个合理的阈值去衡量置信度,如果概率距离高于阈值,表示判决可信,否则认为判决不可信.

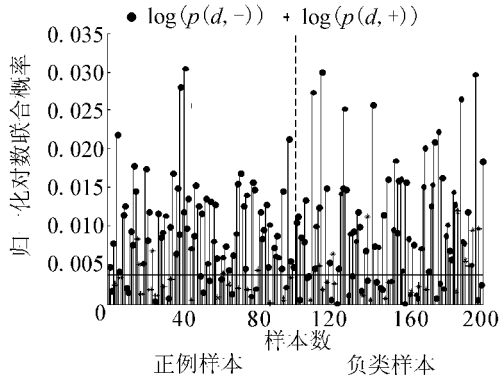


图 2　相同样本归一化对数联合概率距离分布

Fig. 2　Distributions of distance between two-state log-likelihood

2.2　混合模型结构

依据前面的分析,得到这样的结论:当样本处在 2 类空间的临界面附近时,生成式模型 NB 的分类精度不高. 而判别式模型 SVM 基于最大正负样本分类距离准则,相对前者,它对于分类边界处的样本有着较高的判别能力.

基于上述想法,论文提出了一个生成式/判别式混合模型,模型结构如图 3 所示. 其中生成式分类器 NB 作为第 1 级分类器,判别式分类器 SVM 作为 2 级分类器,它们以级联的方式进行组合. 概率距离阈值作为衡量判决置信度的参数,决定 2 个分类器结合的程度. 当 NB 判决的概率距离低于阈值时,转由 SVM 进行二次判别.

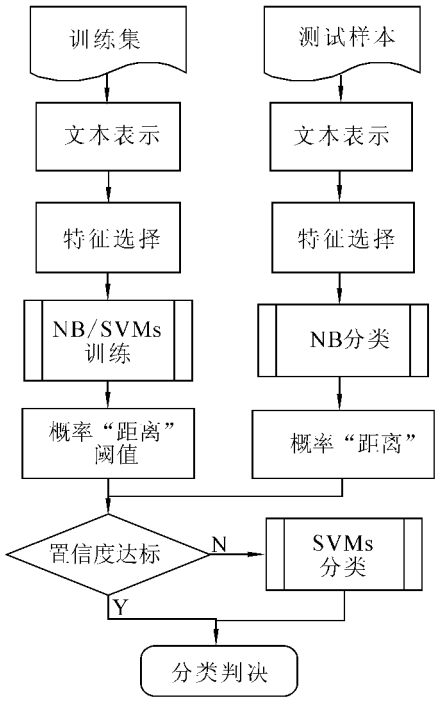


图 3　级联式混合模型结构

Fig. 3　Structure of the hybrid model

3　特征扩展策略

3.1　引入依存句法结构特征

传统的词袋模型(BOW)中,一篇文档被看作一个词袋,完全忽略了词之间的排序信息和句法关系. 虽然高阶 n 元语法,如二元语法(bigrams)和三元语法(trigrams),被用于代替单一的一元语法(unigrams)作为 BOW 的基本特征,然而文献[3]表明在电影评论领域语料中,bigrams 的效果还不如 unigrams,其原因可能是传统的 bigrams 和 trigrams 难于捕捉长距离的依赖关系,对情感分类作用不大.

依存句法信息被认为是情感分类中的有效特征[10,14]. 作为一种句子级粒度的文本结构表示方法,依存句法树利用树中父子节点的关系来表述句子中各词之间的依存关系. 以句子“I definitely recommend this film.”为例,它的依存句法树如图 4 所示.

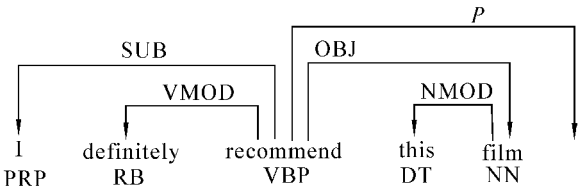


图 4　依存句法结构树示例

Fig. 4　Example of dependency parsing tree

得到依存句法树之后,抽取每个父子节点的词对作为新的特征(如例句中的“definitely recommend”和“film recommend”),这些词对特征能够捕捉句子中词序信息和词之间的长距离依赖关系,经常包含一元语法以外的情感信息. 表 1 中列举了 3 种不同的特征表示方法.

表 1　情感特征表示方法示例

Table 1　Examples of different feature representation

方法	特征表示
Text	I definitely recommend this film.
Unigrams	I, definitely, recommend, this, film
Bigrams	I_definitely, definitely_recommend, recommend_this, this_film
Dependency Pairs	I_recommend, definitely_recommend, this_film, film_recommend

虽然句法结构可以表达更多的文本信息,但是它带来的最大问题就是特征空间变成了原来的平方级,特征空间的急剧增加给后续任务,如特征选择、分类,带来了严重的计算负担.

3.2　依存句法特征扩展策略

为了解决这个问题,在混合模型的基础上,提出了一种高效引入独立依存关系特征的策略:在混合

模型生成式分类器中,概率距离高于阈值的样本有较高的置信度,无需进行特征扩展;而概率距离低于阈值的样本,在第1级分类器中被拒绝判决,在第2级分类器中需要引入句法结构特征以提高其可分性,如图5所示. 该策略不仅能够提高分类精度,而且在效率上也占据优势.

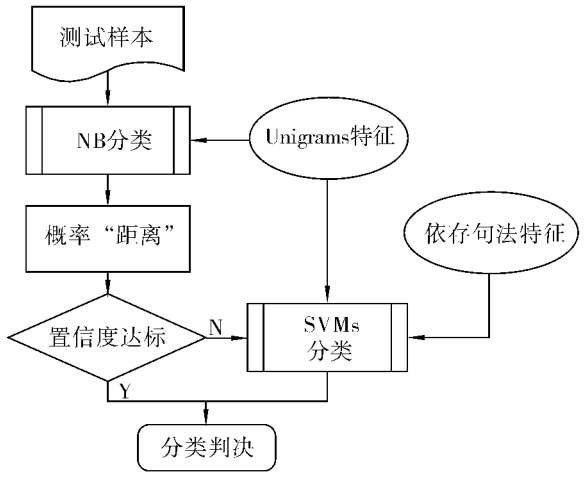


图5 混合模型依存句法特征扩展策略

Fig.5 Structure of the hybrid model with feature expansion strategy

4 实验设计与结果分析

4.1 语料及工具

1)语料:本文选用了多领域情感数据集(multi-domain sentiment dataset)进行实验. 该语料由文献[15]首次引入,之后也得到广泛使用. 该数据集由从Amazon.com抽取的4个领域(Book、DVD、Electronics和Kitchen)的产品评论语料组成,每个领域包含正负例评论文档各1 000篇. 实验采用了全部4个领域的语料.

2)语言分析工具:词性分析是句法分析的预处理步骤之一,选用MXPOST作为词性分析器. 另外,使用MSTParser进行依存句法分析,训练集使用的是宾州书库的WSJ部分.

3)分类器:本文使用开源软件OpenPR-NB[3]和LibSVM[4]作为2种分类器的实现工具. 其中OpenPR-NB的参数设置为多项式模型和拉普拉斯平滑[12],LibSVM采用线性核函数,其他参数均保持默认.

4.2 实验设置

1)交叉验证:每个数据集被平均分成5份,所有的实验结果均经过5倍交叉验证. 交叉验证的每一次循环,4份作为训练集,剩余1份作为测试集.

2)阈值参数训练:阈值是混合模型的一个重要参数,为防止过拟合,参数训练在训练集内使用4倍的交叉验证,最后使用4次循环的均值作为最后的参数,最优参数可以表示为

$$\theta_F^* = \frac{1}{4}\sum_{f\in\bar{F}}\theta_{\bar{F},f}^*,$$

式中:F表示当前测试集,$\bar{F}$表示当前训练集,f表示当前训练集中用于训练参数的开发集,$\theta_{\bar{F},f}^*$在(0, 0.01]范围内以0.005为步长寻找最优值.

4.3 实验1

首先,将NB和SVM作为基线系统,给出Hybrid模型的对比实验结果,如表2所示. 3个系统都以Unigrams作为BOW模型基本特征,分别用U@NB、U@SVM和U@Hybrid表示,特征选择方法使用的是信息增益法(information gain, IG)[16],表2给出了2类实验结果:一类是使用全部特征的分类正确率,表格中用All表示;另一类是经过IG特征选择的最优特征子集的结果,用Best@IG表示.

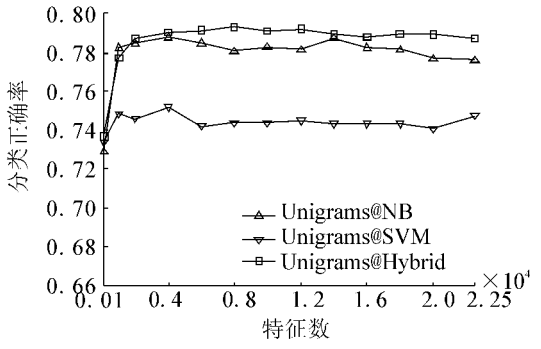
表2 使用Unigrams特征时的系统性能比较

Table 2 The system performance with Unigram features

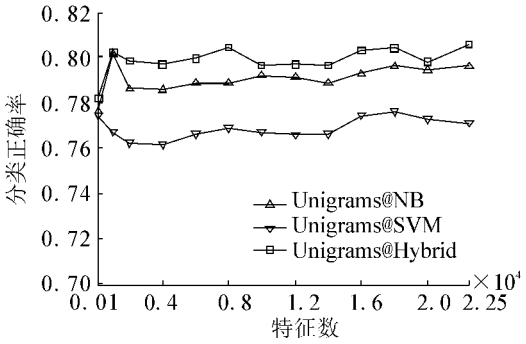
数据集	特征选择	U@NB	U@SVM	U@Hybrid
Book	All	0.776 0	0.747 5	0.787 0
	Best@IG	0.788 0	0.752 0	0.793 0
DVD	All	0.796 0	0.771 0	0.805 5
	Best@IG	0.8015	0.7760	0.8055
Electronics	All	0.817 5	0.804 5	0.820 5
	Best@IG	0.818 5	0.804 5	0.827 0
Kitchen	All	0.828 0	0.828 0	0.853 0
	Best@IG	0.829 5	0.828 0	0.853 0

从表2的结果可以看出,与基线系统NB和SVM相比,Hybrid模型无论是使用全部特征集还是使用最优特征子集,在4个数据集中均表现出了明显的优势.

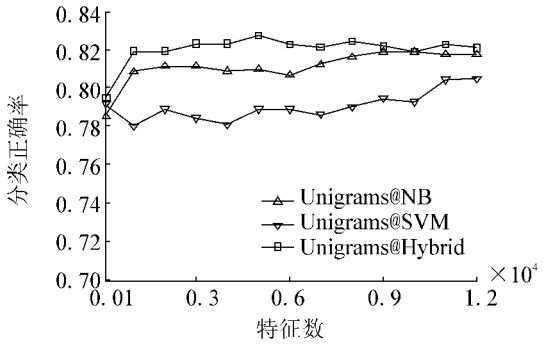
此外,给出了在递增的特征选择子集下,3个模型特征数-分类正确率的曲线,如图6所示.



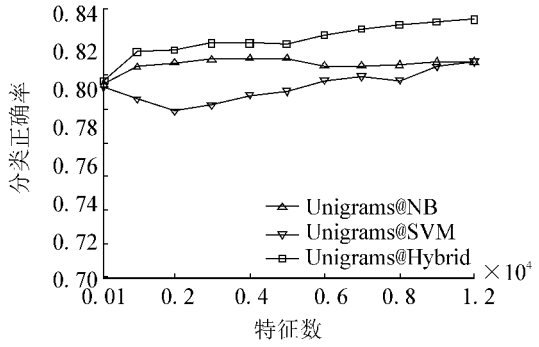
(a)Book领域



(b) DVD 领域



(c) Electronics 领域



(d) Kitchen 领域

图 6　系统在 IG 特征选择下的分类性能

Fig. 6　The accuracy curve under IG feature selection

图 6 中横轴最右边的数值就是使用全部特征的结果,3 条曲线的纵轴最高点就是最优特征子集的结果(参见表 2). 由图 6 可见,无论是在哪个特征子集上,混合模型的曲线均在最上方.

4.4　实验 2

实验 2 中首先在 2 个基线系统上对全部样本都进行依存句法特征扩展(以 unigrams 和依存句法特征的合集作为新的特征集),接着在混合模型中引入第 3 节所述的句法结构特征扩展策略,表 3 给出了综合对比结果.

表 3　句法结构特征扩展后各方法的性能比较

Table. 3　The system performance with feature expansion strategy

数据集	U@NB	U@SVM	U@Hybrid	U+P@NB	U+P@SVM	U@NB & U+P@SVM
Book	0.776 0	0.747 5	0.787 0	0.796 0	0.772 5	0.810 0
DVD	0.796 0	0.771 0	0.805 5	0.812 5	0.780 5	0.823 0
Electronics	0.817 5	0.804 5	0.820 5	0.816 5	0.824 0	0.842 0
Kitchen	0.828 0	0.828 0	0.853 0	0.861 0	0.845 5	0.867 5

其中 U+P@NB 和 U+P@SVM 分别表示 NB 分类器和 SVM 分类器加入依存句法特征的结果. 非常明显地看出,在加入句法结构特征之后,NB 和 SVM 分类器的性能都有了显著提高. 这样的实验结果充分证实了句法结构信息确实是情感文本分类的显著特征.

用 U@NB & U+P@SVM 表示在混合模型上引入句法结构特征的实验结果,表 3 给出了 2 个方向上的结果比较:与混合模型使用原始特征相比,在 2 级分类器上扩展句法特征之后,分类正确率在 5 个领域上均有提高,提高幅值为(1.5 ~ 3.5)%;与 NB 和 SVM 扩展句法结构特征(U@NB、U+P@SVM)两者之中最好的结果相比,各个领域都有(0.5 ~ 2)%的提高. 因此综合来看,混合模型辅以句法结构特征在 2 个方向的比较上都有显著优势.

由于该方法只在 2 级分类器上扩展句法特征的策略,训练语料里绝大部分的样本不需要特征扩展,仅仅需要对分类边界的样本进行扩展,因此该方法大大节省了系统开销.

5　结束语

本文提出了一种情感文本分类混合模型,将生成式、判别式基分类器以一种级联的方式进行组合,旨在消除传统方法对分类边界附近样本由于判决置信度不高而带来的误差. 2 类概率“距离”用于衡量生成式模型判决的置信度,对于置信度不高的样本,生成式模型拒绝判决,交由判别式模型进行分类. 此外,还提出了在 2 级判别式模型中扩展句法结构特

征的策略,通过对难以判决的临界样本增加依存句法信息,以提高其分类的精度,同时只在必要的样本上进行特征扩展.实验表明,与传统方法相比,提出的混合模型及特征扩展策略不仅在分类精度上有显著的、鲁棒的提高,而且在算法效率上,避免了传统特征扩展所带来的高维计算负担,提高了系统效率.

如何更好地将生成式模型和判别式模型融合到一起,以及如何有效地对句法结构特征进行特征选择,是值得进一步研究的问题,这也是下一步即将进行的工作.

参考文献:

[1]宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008: 23-28.

[2]PANG B, LEE L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2: 1-135.

[3]PANG B. Thumbs up? sentiment classification using machine learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, USA, 2002: 79-86.

[4]XIA R. Ensemble of feature sets and classification algorithms for sentiment classification [J]. Information Sciences, 2011, 181: 1138-1152.

[5]RILOFF E, PATWARDHAN S, WIEBE J, et al. Feature subsumption for opinion analysis [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA, 2006: 440-448.

[6]HATZIVASSILOGLOU V, WIEBE J. Effects of adjective orientation and gradability on sentence subjectivity[C]//Proceedings of the International Conference on Computational Linguistics (COLING). Saarbrücken, Germany, 2000: 299-305.

[7]XIA R, ZONG C Q. Exploring the use of word relation features for sentiment classification [C]//Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China, 2010: 1336-1344.

[8]XIA R, ZONG C Q. A POS-based ensemble model for cross-domain sentiment classification [C]//Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). Chiang Mai, Thailand, 2011: 614-622.

[9]GAMON M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis[C]//Proceedings of the International Conference on Computational Linguistics (COLING). Barcelona, Spain, 2004: 841-847.

[10]KENNEDY A, INKPEN D. Sentiment classification of movie reviews using contextual valence shifters[J]. Computational Intelligence, 2006, 22: 110-125.

[11]DAVE K. Mining the peanut gallery: opinion extraction and semantic classification of product reviews[C]//Proceedings of the International World Wide Web Conference (WWW). Budapest, Hungary, 2003: 519-528.

[12]MCCALLUM A, NIGAM K. A comparison of event models for naive Bayes text classification[C]//Proceedings of the AAAI Workshop on Learning for Text Categorization. Madison, USA, 1998: 15-18.

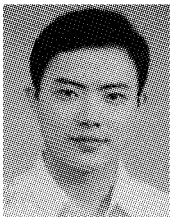
[13]JOACHIMS T. Text categorization with support vector machines: learning with many relevant features[C]//Chemnitz, Germany: Springer, 1998: 237-243.

[14]KUDO T, MATSUMOTO Y. A boosting algorithm for classification of semi-structured text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Barcelona, Spain, 2004:35-41.

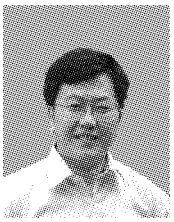
[15]BLITZER J. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification[C]//Proceedings of the Association for Computational Linguistics (ACL). Prague, Czech Republic, 2007: 151-156.

[16]YANG Y, PEDERSEN J. A comparative study on feature selection in text categorization [C]//Proceedings of the Fourteenth International Conference on Machine Learning (ICML). Nashville, USA, 1997: 412-420.

作者简介:



夏睿,男,1981年生,博士,主要研究方向为模式识别、机器学习、自然语言处理和文本挖掘等.



宗成庆,男,1963年生,研究员,博士生导师,中科院自动化所模式识别国家重点实验室副主任.亚洲自然语言处理联合会(AFNLP)执行理事、国际学术期刊 IEEE Intelligent Systems 副主编、ACM Transactions on Asian Language Information Processing 副主编、International Journal of Computer Processing of Languages 副主编、Journal of Computer Science and Technology 编委、《自动化学报》编委、中国中文信息学会常务理事、中国人工智能学会理事,并曾在若干国际学术会议(包括 ACL、COLING 等本领域顶级国际会议)上担任程序委员会及组织委员会主席、Area Chair、委员等职务.主要研究方向为自然语言处理的理论与方法、机器翻译、文本分类等.在大规模口语语料库建设、口语理解与翻译、文本机器翻译和自动分类等方面,提出了一系列新的技术和方法,多次在国际口语翻译权威评测中获得优异成绩.申请国家发明专利10余项。在国内外重要学术刊物和会议上发表学术论文70余篇,出版学术专著1部.