

基于多特征融合的汉语情感分类研究 *

钟 将, 邓时滔

(重庆大学 计算机学院, 重庆 400044)

摘 要: 中文情感分类一般分成基于情感词典和基于特征分类两种方法进行研究,但没有考虑过将两种方法得到的特征进行融合来提高分类效果。基于特征分类的方法忽视了特征词在情感词典的褒贬性以及词倾向性的强弱。用基于特征分类方法得到的文本特征建立朴素贝叶斯模型,根据特征词在情感词典中的褒贬性及其通过对互信息方法得到的词性强弱调整情感词的正负后验概率权重,实现两种特征的融合,提高分类效果并降低了特征维数。

关键词: 文本情感分类; 情感词典; 点对互信息; 特征选择; 朴素贝叶斯

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)01-0098-03

doi:10.3969/j.issn.1001-3695.2012.01.027

Classification approach of Chinese texts sentiment based on integrated features

ZHONG Jiang, DENG Shi-tao

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: Generally the approach of Chinese text sentiment classification was based on the sentiment lexicon or the feature-selection, rather than the integration of the both involved to improve the classification effects. Feature-selection method ignored the emotional tendencies and value of words in the sentiment dictionary. This paper adopted the feature from the method of feature-selection to construct the naive Bayesian model, according to the emotional tendency of the feature in the sentiment dictionary and its value from point mutual information. And adjusted the weights of the positive and negative emotion word posterior probability to achieve the integration, improved the classification results and reduced the feature dimension.

Key words: text sentiment classification; semantic lexicon; point wise mutual information; feature-selection; naive Bayesian

当前互联网上的大量文本中,带有个人观点的文章和言论越来越多,如对产品、人物、新闻等事物的评论,大部分都带有评论者对事物的看法和观点。如何利用计算机自动地对这些评论的看法和观点进行分析和判别,成为当今互联网智能信息处理研究的热点问题。文本中的评论都带有个人情感色彩,如正面和反面、褒义和贬义等,对这些文本进行分析和分类被称为文本情感分类。基于情感的文本分类主要研究如何将文本所表达的情感等主观内容进行分类,判断其是正面的还是负面的。目前的研究工作主要有两种思路:基于情感知识的方法以及基于特征分类的方法^[1]。在基于情感知识的方法中,Hu 等人^[2]分析篇章中评价词语或组合评价单元的极性,然后进行极性加权求和;王素格等人^[3]基于多种词典资源来构建情感词表,使用加权线性组合方法对句子情感进行分类。在基于特征分类的方法中,Pang 等人^[4]首次将机器学习的方法应用于篇章级情感分类任务中;王素格等人^[5]利用多种特征选择和权重计算方法、五种停用词表以及用 SVM 构建分类器对汽车语料的文本情感类别进行研究。

上述方法都是以单独一种方法进行情感分类判断,没有考虑其融合问题。其中基于特征分类的方法忽视了特征词在情感词典的褒贬性以及词性的强弱。本文用基于特征分类方法得到的文本特征构建朴素贝叶斯模型,根据特征词在已有的情感词典中的褒贬性及其通过对互信息(point wise mutual information, PMI)方法得到的词倾向性强弱,调整贝叶斯中情感

词的正负后验概率权重,实现两种方法的融合,提高分类效果并降低了特征维数。

1 情感特征获取与表示

分类器构建需要对进行分词后的文本进行特征选择、计算其特征权值以及表示成文本向量等。使用情感词典可以在中文分词阶段就能完成对文本进行表示,特征分类方法需要单独进行特征选择。

1.1 情感词典及其极性计算

采用情感词典进行特征选择就是只选择分词后的文本中的情感词作为特征,方法简单,在分词处理过程中就可以完成。如果在分词时加入情感词典,就可以很方便地选择出文本中的情感词。同时将情感词典作为特征,特征选择方法简单,可以减少大量的降维计算,节省了时间,提高了分类的速度。用机器学习方法进行文本情感分类时,不用考虑训练语料和测试语料特征的对应问题,使得训练和测试可以分开进行,还可以一次训练、多次分类。

本文以 HowNet^[6]发布的情感词集为基础,通过人工挑选去掉一些不常用的词或者情感倾向不很明显的词,得到 4 435 个褒义词和 4 225 个贬义词作为情感词典,通过语料库的统计特性得到其词倾向性强弱。

对于文本 $d = \{w_1, w_2, \dots, w_n\}$ 的情感倾向属于 $C = \{\text{pos},$

收稿日期: 2011-06-06; 修回日期: 2011-07-15 基金项目: 国家“211 工程”三期建设项目(S-10218)

作者简介: 钟将(1974-),男,重庆人,副教授,博士,主要研究方向为文本分析、数据挖掘、知识管理;邓时滔(1987-),男,四川资阳人,硕士,主要研究方向为文本分类、数据挖掘(dengshitao@cqu.edu.cn).

neg_i},使用 PMI 得到情感词的倾向性,根据词语在语料库中的分类情况判断其联系紧密程度,即将该词与褒义类的互信息量减去该词与贬义类的互信息量后所得的结果。计算方法如式(1)所示。其正负号表示词语的极性,由于情感词典已经将词语的极性确定,主要取绝对值代表了词语的倾向性强度。将 $PV_{PMI}(w)$ 按式(2)进行标准化操作,标准化空间为 $[-1, 1]$ 。

$$PV_{PMI}(w) = PMI(w, pos) - PMI(w, neg) = \log_2 \frac{P(w, pos)/P(pos)}{P(w, neg)/P(neg)} = \log_2 \frac{P(w|pos)}{P(w|neg)} \quad (1)$$

$$pv(w) = \frac{\arctan(PV_{PMI}(w)) \times 2}{\pi} \quad (2)$$

1.2 特征分类

基于特征分类的方法主要是使用机器学习的方法,对文本进行特征选择完成分类任务。特征选择是选择文本中能显著表示文本类别的词汇,去掉无用的词汇。因为文本分词后得到的是一个稀疏矩阵,为了便于计算,一般都要降低矩阵的维度。Yang 等人^[7]提到 chi 统计方法是一种应用比较广泛而且性能不错的特征选择方法,同时使用 chi-max 作为特征选择在文本分类方面表现较好。采用 chi 统计作为特征选择方法,每个特征词 w 对于类别 c_j 的 X^2 值计算方法如下:

$$X^2(w, c_j) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3)$$

其中, A 是特征词 w 和 c_j 类文档共同出现的次数; B 是特征词 w 出现而 c_j 类文档不出现的次数; C 是 c_j 类文档出现但特征词 w 不出现的次数; D 是 c_j 类文档和特征词 w 都不出现的次数; N 是文本总数。采用式(4)作为特征选择的依据,即选用 chi-max 值最大的若干词作为特征词。

$$X^2_{\max} = \max_{c_j \in C} \{X^2(w, c_j)\} \quad (4)$$

1.3 特征权值计算与文本表示

特征权值是特征词在文本中的权重,也可称为词的向量,是分类器分类的重要依据。唐惠丰等人^[8]使用朴素贝叶斯和最大熵方法进行新闻及评论语料的情感分类研究,通过实验认为二值作为特征项权重正确率较好。本文也使用布尔型权值进行情感分类对比,出现为 true,没有出现为 false。一般在分词处理完成后就可以计算特征权值,进行特征选择后输入分类器,用词向量空间表示。

2 朴素贝叶斯模型的构建与特征融合

2.1 朴素贝叶斯模型构建

朴素贝叶斯分类器是解决相应机器学习问题最有实际价值的方法之一。在多数情况下与其他学习算法性能相当,在某些情况下还优于其他算法。特别是应用到文本分类问题,朴素贝叶斯分类是最有效的算法之一。

本文首先采用了特征分类和朴素贝叶斯理论构建情感分类器。对于文本 $d = \{w_1, w_2, \dots, w_n\}$ 的情感倾向属于 $C = \{pos, neg\}$, 在特征相互独立的情况下,其分类式如式(5)。

$$C = \max_{c_j \in C} \{P(c_j) \prod_{i=1}^n P(w_i | c_j)\} \quad (5)$$

其中, $P(c_j)$ 是类别 c_j ($c_j \in C$) 的先验概率, $P(w_i, c_j)$ 是特征词 w_i 在类别 c_j 中的后验概率。对于先验概率 $P(c_j)$, 可以视每类是相同的,也可以预先估计,本文采用数目相同的训练集,得到 $P(c_j) = 1/2$ 。对于后验概率 $P(w_i | c_j)$, 即特征词 w_i 出现在类

别 c_j 中的概率,一般是从训练语料中通过计算进行估计。本文采用词 w_i 在属于类别 c_j 的文本频率除以类别 c_j 的文本中所有词的文本频率之和。为避免 $P(w_i, c_j)$ 等于 0, 可采用 Laplace 转换,由此后验概率计算方法如式(6)所示。

$$P(w_i, c_j) = \frac{df(w_i, c_j) + \delta}{\sum_{i=1}^n df(w_i, c_j) + |V|} \quad (6)$$

其中, $f(w_i, c_j)$ 是词 w_i 在属于类别 c_j 的文本中的文本频率。采用 Laplace 转换,一般情况下常数 V 取所有词的数目, δ 取 1。下面融合特征分类和情感词典的特征。

2.2 特征融合

通过分析发现,在利用基于特征分类和朴素贝叶斯方法进行情感分类时,只利用了每个词在训练语料中的假设概率,却没有考虑这个词在情感词典中的褒贬性及其强弱,即该词对不同类的差异。例如,假设在某一个负面文本中,只有两个通过基于特征分类方法得到的特征词“寂寞”和“太阳”,“寂寞”是情感词典的负面情感词,“太阳”是隐含的情感词。通过训练语料得到,“寂寞”对正面和负面两个类的后验概率分别为 0.3、0.4;“太阳”对正面和负面两个类的后验概率分别为 0.5、0.3。根据朴素贝叶斯公式计算该文本为正面的概率为 0.15,负面的概率为 0.12,将其判断为正面文本。但是文本本身却是负面,因此可以加大情感词典的词“寂寞”的负面后验概率,使其后验概率为 0.6,负面的概率为 0.18,正文被判定为负面。基于以上分析,本文提出将特征分类与情感词典的特征进行特征融合。

对特征分类中出现过的情感词典的情感词 w_i 进行加权;正面的情感词对正面后验概率进行加权;负面的情感词对负面后验概率进行加权。加权公式如式(7)所示。其中 θ 代表了情感词对分类的强弱。情感词对分类的影响大时, θ 偏大;情感词对分类的影响小时, θ 偏小。各种情感评论语料的 θ 不同。 $1 + |pv(w_i)|$ 代表了单个情感词对分类的影响程度, $pv(w_i)$ 可能为正也可能为负,为正值就代表它应该对正面情感产生影响,为负值就代表它应该对负面情感产生影响,其具体阈值代表单个情感词对分类的影响程度。

$$P(w_i, c_j) = p(w_i, c_j) \times \theta \times (1 + |pv(w_i)|) \quad (7)$$

3 实验

3.1 语料与评价指标

实验采用由谭松波提供的关于电脑和酒店的评论情感平衡语料库^[9],正负类各 2 000 篇。本文选择正负各 1 400 篇作为训练集,正负各 600 篇作为测试集。进行情感分类的三种评价指标分别定义如下:

$$precision = \frac{\text{系统判别正确的评论数}}{\text{系统能够判别的评论数}} \times 100\% \quad (8)$$

$$recall = \frac{\text{系统判别正确的评论数}}{\text{总评论数}} \times 100\% \quad (9)$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100\% \quad (10)$$

实验采用了式(10)来衡量实验结果。

3.2 实验结果与分析

实验首先使用情感词典作为情感特征,特征维数为 8 860,分别对电脑和酒店训练语料进行训练,对测试语料进行了分类测试,实验结果如表 1 所示。可以看出,情感词典对分类的效果影响差不多。

然后使用 chi 统计方法选择特征进行贝叶斯分类以及按

式(7)进行特征融合(integration),对同样的语料进行训练和分类测试。在特征维数为2 000时,电脑语料的正面情感词 θ 值为0.5、负面情感词 θ 值为3、酒店语料情感词 θ 值都为12。实验结果如表2所示,融合后的分类效果有所提高。其不同特征维数的实验结果如图1所示。从图中可以看出,关于计算机评论的chi分类效果最高的F1值在89%,融合后的分类效果最高值在91%,融合后的特征还是有所提高;关于酒店评论的chi分类最高的F1值在75%,融合的特征基本在80%以上,融合后的分类效果有明显的提高,说明了本文方法的正确性。

表 1 使用情感词典的分类实验结果(F1/%)

结果	情感词典	
	电脑	酒店
F1 值	77.4	78.8

表 2 特征选择以及特征融合的分类实验结果(F1/%)

方法	电脑	酒店
chi	89.3	71.3
integration	90.3	82.2

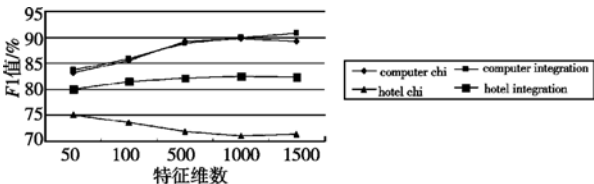


图1 使用特征选择与特征融合在不同语料库上的实验结果

4 结束语

实验结果表明,单纯地使用某一类方法得到的特征进行的分类效果都小于融合特征后的效果。由于缺乏对情感词典词

性倾向性的研究和完备的语料库,如何从已有的语料库中选择更好的特征来构建较好的概率模型以及搜集更多的情感语料,是下一步的工作。

参考文献:

[1] 赵妍妍,泰兵,刘挺. 文本情感分析[J]. 软件学报,2010,21(8): 1834-1848.

[2] HU Ming-qing, LIU Bing. Mining and summarizing customer reviews [C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004:168-177.

[3] 王素格,杨安娜,李德玉. 基于汉语情感词表的句子情感倾向分类研究[J]. 计算机工程与应用, 2009,45(24):153-155.

[4] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2002: 79-86.

[5] 王素格,魏英杰. 停用词表对中文文本情感分类的影响[J]. 情报学报,2008,27(2):175-179.

[6] 知网[EB/OL]. [2011-04-22]. <http://www.keenage.com>.

[7] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publisher,1997: 412-420.

[8] 唐惠丰,谭松波,程学旗. 基于监督学习的中文情感分类技术的比较研究[J]. 中文信息学报, 2007,21(6):88-94.

[9] 谭松波. 中文情感挖掘语料 ChenSentiCorp [EB/OL] (2010-06-29) [2011-04-22]. <http://www.searchforumrg.cn/tan-songbo/corpus-senti.htm>.

(上接第84页) 表 1 控制效果对比

控制律	谐波跟踪			直航镇定		
	T_s/s	$W_e/^{\circ}$	J_E	T_s/s	$W_e/^{\circ}$	J_E
1	14	155	141	15	132	32
2	23	172	157	21	161	49

4 结束语

本文中讨论了存在模型不确定性和未知外界干扰力,并考虑舵机伺服系统因素时,USV 航向非线性系统的跟踪控制问题。该控制方法的优点在于设计中考虑了舵机特性以及未知建模误差和环境干扰力的影响,且具有自适应能力。仿真对比实验验证了本文所提控制律具有良好的自适应能力和跟踪性能。

参考文献:

[1] SKJETNEA R, FOSSEN T I, KOKOTOVIC P V. Adaptive maneuvering, with experiments, for a model ship in a marine control laboratory [J]. Automatica, 2005, 41(2): 289-298.

[2] HU S S, YANG P H, JUANG J Y, et al. Robust nonlinear ship course keeping control by H_{∞} I/O linearization and μ -synthesis[J]. International Journal of Robust and Nonlinear Control, 2002,13(1): 55-70.

[3] 刘雨,郭晨,吕进. 基于 backstepping 的船舶航向滑模控制[J]. 中南大学学报:自然科学版, 2007, 38(Z1): 278-279.

[4] 朱齐丹,周芳,赵国良,等. 基于反步法和滑模观测器的船舶航向控制[J]. 华中科技大学学报:自然科学版, 2009, 37(4): 121-122.

[5] 杜佳璐,郭晨,杨承恩. 船舶航向非线性系统的自适应跟踪控制器设计[J]. 应用科学学报, 2006,24(1): 83-87.

[6] 张显库. 船舶航向保持的非线性逆推鲁棒控制算法[J]. 大连海事大学学报, 2007, 33(2): 80-83.

[7] 罗伟林,邹早建,李铁山. 船舶航向非线性系统鲁棒跟踪控制[J]. 控制理论与应用, 2009, 26(8): 893-894.

[8] 袁雷,吴汉松. 船舶航向非线性系统的多滑模自适应模糊控制[J]. 智能系统学报, 2010, 5(4): 308-312.

[9] YU Zhen-yu, BAO Xin-ping, NONAMI K. Course keeping control of an autonomous boat using low cost sensors[J]. Journal of System Design and Dynamics, 2008,2(1): 389-400.

[10] YU Zhen-yu, BAO Xin-ping, NONAMI K. Mixed H_2/H_{∞} based course following control for a small low cost autonomous boat [J]. ICGST International Journal on Automatic Control and Systems Engineering, 2007,7(1): 51-59.

[11] BAO Xin-ping, NONAMI K, YU Zhen-yu. Combined yaw and roll control of an autonomous boat[C]//Proc of IEEE International Conference on Robotics and Automation. 2009: 188-193.

[12] LIAO Yu-lei, PANG Yong-jie, WAN Lei. Combined speed and yaw control of underactuated unmanned surface vehicles[C]//Proc of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics. 2010: 157-161.

[13] 范尚雍. 船舶操纵性[M]. 北京:国防工业出版社, 2002: 30-39.

[14] SUSSMANN H J. A general theorem on local controllability[J]. SIAM Journal of Control and Optimization, 1987,25(11): 158-194.

[15] 刘金琨. 滑模变结构控制 MATLAB 仿真[M]. 北京:清华大学出版社, 2002: 240-247.