DOI:10.11830/ISSN.1000-5013.201608010

结合情感词典的主动贝叶斯文本情感分类方法

张敏,陈锻生

(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘要: 提出一种改进的结合情感词典的主动贝叶斯情感分类方法(SLAB).为了证明提出方法的有效性,选用康奈尔影评数据集和互联网电影资料库(IMDB)数据集作为实验数据,并与基于不确定性采样策略的主动学习方法进行比较.结果表明:文中提出的方法在较少的标注训练集下,能够取得更高的分类准确率,一定程度上解决了基于不确定性采样策略的主动学习方法中的误差累积问题.

关键词: 主动学习;文本情感分类;情感词典;朴素贝叶斯;不确定采样策略

中图分类号: TP 391.1 **文献标志码:** A **文章编号:** 1000-5013(2018)04-0623-04

Text Sentiment Classification Based on Semantic Lexicon and Active Bayesian

ZHANG Min, CHEN Duansheng

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: An improved sentiment classification method combined semantic lexicon and active Bayesian (SLAB) is proposed. To demonstrate the effectiveness of our proposed method, the Cornell movie review datasets and Internet movie database (IMDB) datasets are exploited as our experimental data and the active learning method based on the uncertainty of sampling is studied as a comparison. The results show that the proposed method can achieve higher classification accuracy with less labeled training set, which alleviates the influence of error accumulation caused by the active learning method based on the uncertainty of sampling.

Keywords: active learning; text sentiment classification; semantic lexicon; naive Bayesian; uncertainty sampling strategy

为了高效地获得人们对诸如人物、事件、影视剧、产品等有价值的评论信息,文本情感分类应运而生[1].目前,国内外情感分类主流的方法是监督学习方法[2-3],但其不足之处在于需要大量的标注样本对分类器进行迭代训练,否则,根据概率近似正确(PAC)学习理论,算法的泛化能力无法有效提高[4].获取大量的标注样本需要较多的人力、物力,因此,主动学习作为一种能够运用小规模样本数据集获得较理想的分类性能的方法,已成为研究的热点[5-10].主动学习算法的采样策略是算法成功的关键所在.基于不确定性采样策略是目前适用性最广且较为成熟的采样策略[11-12].学习此类训练样本有助于在少量标注训练集条件下快速地提高分类器的分类性能.然而,由于此类训练样本大多不具有类别代表信息,如果不断添加难以判断类别的此类样本进行分类器训练,可能造成分类器分类误差的不断累积,影响分类

收稿日期: 2016-08-08

通信作者: 陈锻生(1959-),男,教授,博士,主要从事机器学习与数据挖掘的研究.E-mail:dschen@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(61370006);福建省科技计划(工业引导性)重点项目(2015H0025);华侨大学研究生科研创新能力培育计划资助项目(1400214023)

器性能的提升.因此,有人提出在选择信息量丰富的样本的同时,选择具有类别代表信息的样本方法,并取得了较好的实验结果[13-15].本文提出一种结合情感词典的主动贝叶斯情感分类方法(SLAB).

1 分类器设计方法

1.1 基于不确定性采样策略的主动学习

对于二分类问题而言,不确定性采样策略认为,未标注样本经过分类器判别为不确定性最大的样本往往处于两类样本的分类界面.因此,反复学习不确定性最大的样本有助于分类器找到样本分界面.同时,根据泛函原理可知,对于线性可分问题,分类间隔中的样本对分类器的影响较大.因此,每次选择不确定性较大的样本更新训练集,候选样本集(未加入分类器学习的所有剩余训练样本)中剩下的样本对分类器的影响逐步减弱,这使利用少量高质量的训练样本达到预期分类准确率成为可能.

所选的分类器模型不同,衡量样本不确定性的度量方式也有所不同.Lewis等[12]提出一种最基本的适用于概率模型的不确定性度量方式.朴素贝叶斯分类器算法由于实现简单、分类效率较高、具有增量学习的特性等优点,大大减少了主动学习中因多次采样和反复训练分类器所需的计算量[16-18],其在文本分类方面也有较好的表现.因此,采用朴素贝叶斯分类器对样本进行分类,而衡量样本不确定性的表达式将使用样本类别的后验概率进行估计,即 $p(c_i|x)\propto p(x|c_i)p(c_i)$,$\mathrm{UNCE}(x)=\min\limits_{i\in\{\mathrm{pos},\mathrm{neg}\}} p(c_i|x)$.其中:$x$ 为预测样本;c_i 为类别;UNCE 为样本的不确定性大小.

1.2 结合情感词典的主动文本情感分类算法

不确定性采样策略是选择当前分类器最不能确定类别的样本,这体现其对含有特殊信息样本的及早重视,但这些特殊信息无法很好地代表一类样本的特性.同时,不确定性采样策略也存在面临选择野点的风险,从而分类器从中学习错误的、导致分类器准确率下降的知识,且随着分类器的反复学习,分类器的误差也将逐渐增大.为了有效地抑制分类器误差的传播,文中提出一种改进的基于不确定性采样策略的主动学习方法 SLAB,如下所示.

算法 1 结合情感词典的主动贝叶斯文本情感分类算法 SLAB

输入:标注样本集 L,未标注样本 U,情感词典 Lexicon

输出:新的标注样本集 L,贝叶斯分类器 F

程序:循环 N 次

1) 从 L 中学习贝叶斯分类器 F;

2) 使用 F 对未标注样本集 U 中每一个样本进行分类,获得 U 中每一个样本类后,验概率;

3) 选择 U 中类后,将验概率最接近 0.5 的 n 个样本,提交给专家进行标注,加入 L' 中;

4) 运用 Lexicon 计算情感分数,选择 L' 中情感分数最大的 $m(m\leqslant n)$ 个样本加入 L,未选中的样本重新放回 U 中.

在不确定性采样策略的基础上,使用情感词典,统计样本中情感特征词个数,文中定义其为样本的情感分数.选择情感分数较大的样本进行人工标注.然后,加入训练集中,重新学习分类器.其中,样本的情感分数越大,表示样本具有更多的情感代表性信息,一定程度上能够代表一类样本的特性.从人工标注样本情感极性的角度也不难发现,样本中带有更多的情感词语更易于判断样本情感极性.因此,采样策略在选择最不确定的样本的同时,也选择情感分数较大的样本,能够弥补不确定性采样策略的不足.

情感词典是对词典中具有感情色彩的词语按照极性、强度、词性等打上不同的标签(积极、消极、中性等),以便在情感分类任务中灵活应用.国内外主要的情感词典有 GI 评价词典、NTU 评价词典、HowNet 评价词典及知网提供的评价词典.文中采用知网提供的英文情感分析用词语集.其中,正面情感、负面情感、正面评价、负面评价、程度级别及主张词共 8 945 个.

2 实验与分析

2.1 数据集及预处理

Pang 等[19]整理的康奈尔影评数据集,以及 Maas 等[20]整理的互联网电影资料库(IMDB)影评数据

集是情感分析中具有一定代表性的平衡语料库. 其中，康奈尔影评数据集中拥有标注的褒贬极性句子各 5 331 句，而 IMDB 则包含训练集及测试集各 25 000 例. 不失一般性，从中随机选择部分语料作为实验数据，实验在 Python 环境下，使用 TfidfVectorizer 提取文档特征，文档的特征向量由词袋模型表示，长度预设为 100. 其中，词项权重使用词频-逆向文档频率(TF-IDF)方法计算. 这些实验数据及算法代码将随同文中发布，以便参考交流.

2.2　评价指标

准确率(η_A)是文本情感分类中的常用评价指标，反映了分类器对整体样本的判定能力. 以 η_A 作为评价指标，$\eta_A = n/N$. 式中：n 为分类正确的测试样本数；N 为总测试样本数.

2.3　实验结果

实验数据选用部分康奈尔影评数据集和部分 IMDB 影评数据集，分类算法采用 Sklearn 计算库里的朴素贝叶斯分类器，主动学习算法采用 pyAL 库实现，实验参数均为默认值，SLAB 算法中每次采样样本数 m 设为 5，其他参数和主动学习算法相同. 康奈尔影评数据集与 IMDB 影评数据集实验比较结果，如表 1 所示. 表 1 中：SLAB 表示结合情感词典的主动贝叶斯方法；UNCE 表示基于不确定的主动贝叶斯方法；RAND 表示基本的贝叶斯分类方法；采样数 m 表示经采样策略或随机选择样本更新训练集的标注样本个数；INCR 表示 SLAB 方法相对于 UNCE 方法分类器准确率提高的程度.

表 1　实验结果比较

Tab. 1　Comparison of experimental results

m	康奈尔影评数据集				IMDB 影评数据集			
	RAND	UNCE	SLAB	INCR	RAND	UNCE	SLAB	INCR
10	52.50	52.50	61.57	9.07	56.54	56.54	51.43	−5.11
50	56.98	59.43	60.35	0.92	60.88	59.74	62.84	3.10
100	64.42	62.79	67.18	4.38	60.26	63.14	64.83	1.69
150	66.87	65.65	70.44	4.79	59.74	64.31	65.81	1.50
200	67.18	69.11	72.99	3.87	59.48	64.82	66.67	1.85
250	69.72	71.46	74.82	3.36	60.07	65.76	67.13	1.37
300	71.66	74.21	75.64	1.43	60.81	66.20	67.03	0.82
350	73.19	74.62	75.64	1.02	62.32	66.49	66.99	0.50
400	74.01	75.64	77.68	2.04	62.20	66.88	67.56	0.68
450	74.41	76.25	78.90	2.65	62.33	67.33	67.90	0.57
500	75.74	76.66	78.80	2.14	62.40	67.73	68.17	0.43

2.4　实验结果分析

由实验结果可知：分类器的准确率在应用主动学习贝叶斯方法方法(SLAB，UNCE)的情况下比传统的朴素贝叶斯方法(RAND)有明显地提高，同时，SLAB 又一定程度上优于 UNCE. SLAB 表现最好的原因在于，它主动选择了最有利于分类器提高性能的样本作为训练样本，多次循环训练分类器使其分类准确率不断提高，进而在较小的标注代价下获得分类器的强泛化能力. SLAB 相对于 UNCE 的准确率提高程度不是很明显的一个可能原因是情感词典的构成.

目前，可用的情感词典中收录的多为常规情感词，且并不针对某一特定领域. 众所周知，不同领域会含有特殊的情感表达词. 因此，影评数据集如果能够应用本领域的情感词典应该更能体现 SLAB 方法的优势.

分类器的准确率并不是随着训练样本的数目增加而单调递增，而可能呈现局部最优. 这是由于当采样策略选择“不好”的样本加入训练集时，分类器可能学习到不可靠的知识，不可靠的分类器反过来也将影响下一步的采样策略选择样本，如此循环最终导致样本分类表现不好. 因此，作为分类器的初始化训练样本，初始种子样本的采样策略是否需要特殊处理也是一个值得研究的问题.

值得注意的是，文中提出的方法适用于平衡语料，而当语料的正负样本数不相同时，则会影响机器学习中分类模型的建立，进而影响主动学习结果. 为了减小非平衡语料对主动学习结果的影响，有学者认为此时标注少类样本更有意义，并提出以样本的确定性作为选择少类样本的衡量指标.

3 结束语

提出一种结合情感词典的主动贝叶斯文本情感分类方法.实验结果表明:在较少的等量标注训练集下,提出的方法相较于基于不确定采样策略的主动学习方法可取得更高的分类准确率.不断地选择"好的"样本有助于提高主动学习的分类器性能;反之,当分类器选择"坏的"样本则可能导致分类器误差的不断累积,采样策略中初始种子样本的选择影响后续样本的选择,也最终会影响分类器的性能.因此,今后的研究重点在基于主动学习的文本情感分类中如何选择"好的"初始种子样本.

参考文献:

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

[2] WANG Sida,MANNING C D. Baselines and bigrams: Simple, good sentiment and topic classification[C]// Meeting of the Association for Computational Linguistics. Stroudsburg:ACL,2012:90-94.

[3] WU Fangzhao,SONG Yangqiu,HUANG Yongfeng. Microblog sentiment classification with contextual knowledge regularization[C]// Conference on Artificial Intelligence. Halifax:AAAI Press,2015:2332-2338.

[4] 吴伟宁,刘扬,郭茂祖,等.基于采样策略的主动学习算法研究进展[J].计算机研究与发展,2012,49(6):1162-1173.

[5] ZHU Weizhong,ALLEN R B. Active learning for text classification: Using the LSI subspace signature model[C]// International Conference on Data Science and Advanced Analytics. New Jersey:IEEE Press,2014:149-155. DOI:10.1109/DSAA.2014.7058066.

[6] CETIN M,AMASYALI M F. Active learning for Turkish sentiment analysis[C]// Innovations in Intelligent Systems and Applications. New Jersey:IEEE Press,2013:1-4.

[7] KUMAR A,KANSAL C,EKBAL A. Investigating active learning techniques for document level sentiment classification of tweets[C]// 7th International Conference on Communication Systems and Networks. Bangalore: IEEE Press,2015:1-6. DOI:10.1109/COMSNETS.2015.7098727.

[8] 赵建华,刘宁.结合主动学习策略的半监督分类算法[J].计算机应用研究,2015,32(8):2295-2298.

[9] ANGLUIN D. Queries and concept learning[J]. Machine Learning,1988,2(4):319-342. DOI:10.1007/BF00116828.

[10] SETTLES B. Active learning literature survey[J]. University of Wisconsinmadison,2010,39(2):127-131.

[11] SUN Lili,WANG Xizhao. A survey on active learning strategy[C]// International Conference on Machine Learning and Cybernetics. New Jersey:IEEE Press,2010:161-166. DOI:10.1109/ICMLC.2010.5581075.

[12] LEWIS D D,GALE W A. A sequential algorithm for training text classifiers[J]. Sigir,1994,29(2):3-12.

[13] HUANG Shenjun,JIN Rong,ZHOU Zhihua. Active learning by querying informative and representative examples [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(10):1936-1949. DOI:10.1109/TPAMI.2014.2307881.

[14] DONMEZ P,CARBONELL J G,BENNETT P N. Dual strategy active learning[C]// European Conference on Machine Learning. Berlin:Springer,2007:116-127.

[15] ZHAO Xu,YU Kai,TRESP V,*et al*. Representative sampling for text classification using support vector machines [C]// European Conference on Information Retrieval. Berlin:Springer, 2003:393-407. DOI:10.1007/3-540-36618-0_28.

[16] 杨鼎,阳爱民.一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J].计算机应用研究,2010,27(10):3737-3739.

[17] 宫秀军,孙建平,史忠植.主动贝叶斯网络分类器[J].计算机研究与发展,2002,39(5):574-579.

[18] HOULSBY N,HUSZÁR F,GHAHRAMANI Z,*et al*. Bayesian active learning for classification and preference learning[J]. Computer Science,2011:10-13.

[19] PANG B,LEE L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proceedings of the 43rd Annual Meeting of the ACL. Morristown:ACL,2005:115-124.

[20] MAAS A L,DALY R E,PHAM P T,*et al*. Learning word vectors for sentiment analysis[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Morristown:ACL,2011:142-150.

(**责任编辑**:钱筠 **英文审校**:吴逢铁)