情报学报 ISSN 1000-0135
第35卷第4期 349-357, 2016年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 4, 349-357 April 2016

doi:10.3772/j.issn.1000-0135.2016.004.002

基于修正 G^2 特征筛选的中文微博情感组合分类

杜亚楠 刘业政

(合肥工业大学管理学院,合肥 230009)

摘要 新词的涌现、热词的漂移、海量碎片化及中文常用词特性带来的高维稀疏性成为中文微博情感分类的主要困难。本文提出了一种新颖的方法用以解决上述问题:构造表情符号词典用来自动获取微博的情感标签,解决海量微博数据标注的问题;引入修正的 G^2 检验进行特征筛选,进行降维,控制稀疏性;采用多阶段判断的抽样策略保证基分类器的多样性,最后采用加权多数投票的方式对基分类器结果进行融合,解决特征和情感漂移及碎片化问题。实验表明本文方法可以快速有效的获取训练标签,保留下强区分能力的特征,并实现较高的精度,在中文微博情感分类上是一个有竞争力的方法。

关键词 表情符号词典 修正 G^2 检验 多阶段判断抽样 加权多数投票 组合分类器

Ensemble Emotion Classifier for Chinese Tweets Based on Modified G^2 Feature Selection

Du Yanan and Liu Yezheng

(*School of Management, Hefei University of Technology, Hefei* 230009)

Abstract New emerging words, hot words shifting, fragmentation and large dimension and sparse features generated from large amounts of most commonly used Chinese words are the main obstacles to emotional analysis of Chinese tweets. In this paper, we propose a novel approach for emotion classification of Chinese tweets to address above issues. Using emoticons construct a dictionary, then auto get Tweets emotional label. Modified G^2 test was introduced for feature selection. The multistage judgment sampling strategy was employed to ensure the diversity of base classifiers. The weighted majority voting was applied to combine the ensemble classifier. Our approach is a swarm intelligence approach. The experiments demonstrate that the method can quickly and efficiently obtain emotional labels, retain strong distinguish features, more important, achieve higher accuracy. It is a competitive method for emotion classification of Chinese tweets.

Keywords dictionary of emoticons, modified G^2 test, multistage judgment sampling, weighted majority voting, ensemble classifier

1 引 言

在线社交网络如 Twitter、Face Book 和微博等日益流行,海量用户通过这些社会化媒体平台针对广泛的主题分享信息、交换意见,这为企业、政府及社会学者提供了丰富的数据资源。从海量碎片化的信息中提取和分析用户的观点、态度和情感对商务智能应用[1,2]及组织决策[3]都有重要的意义和作用。例如,消费者在做出购买决定前可以通过情感分析来帮助其决策;市场营销人员可以通过情感分析来研究用户对其产品和服务的态度,进而为客户关系管理及市场营销提供重要的数据支持;组织部门可以利用情感分析获取公众对事件的关键反馈并做出

收稿日期:2015年10月12日

作者简介:杜亚楠,女,1985年生,博士研究生,主要研究方向:电子商务、数据挖掘等,E-mail:duyanan2046@163.com;刘业政,男,1965年生,博士,教授,博士生导师,主要研究方向:电子商务、数据挖掘、智能决策理论方法及社会性网络等。

合理的决策。因此,一种适合海量碎片化中文微博数据的情感分析技术显得迫切而重要。

情感分析一直是近年来研究的热点,但是相比社交网络而言大部分的成果是关于产品和电影评论[4~6]。这是因为在社交网络上,用户的表达更加口语化,且充满了短句、讽刺、缩写等,从而导致精度不高。相较于英文语言的Twitter情感分析[6~9],中文微博的情感分析工作更不成熟[10,11]。这是因为中文微博有其自身独有的处理难点,如:①英文是空格自动分割,词边界准确。中文是连续书写,由于微博用词特性的随意,分词结果的误差造成语义更加难以理解。②Twitter允许用户发布信息不超过140个字符,而中文微博要求用户不超过140个中文字符。然而,中文常用词有56 008个,包括3181个单音节词,40 351个双音节词,6459个三音节词,5855个四音节词以及162个5音节词,词特征选项庞大,而每则微博里用词有限,这将导致数据特征超高维且稀疏。③缺乏统一规范的研究语料库,使得对比实验很难规范展开。④再加上微博情感分析固有的困难:从其他语言及动漫中泊来的新词、热词等不断涌现,伴随新词、热词而来的是特征及情感的同时漂移,从流数据中获取最有效的特征变成一个具有挑战性的工作。因此,中文微博的情感处理更加复杂。

中文情感分类的主要任务集中在特征筛选,标签获取和分类器构造上。特征选择是情感分析里的一项重要任务,它可以显著的影响算法的精度。以本文中语料为例,语料库共368 614条记录,一条微博是一条记录。经过分词和去停用词后,得到了171 018个特征。但是并不是所有的特征对于情感分类而言都是有用且重要的,有时候某些特征反而会成为干扰。我们希望保留下来具有强分类能力的那些特征,现有的针对在线评论的特征选择方法并不适合大规模的微博流数据。选取有效的特征筛选方法,不会导致过过滤和欠过滤显得尤为重要。传统的情感分类方法大致可以分为有监督和无监督两类。基于字典的方法是典型的无监督情感分类[3,12],其原理是使用情感词典和帖子中的词进行匹配进而确定帖子的情感倾向,情感词典的完备性直接影响情感分类的精度。有监督的学习最关键的是从带标签的数据集中训练出鲁棒的情感分类器。然而,在实际中人工获取标签既耗时且成本又高。实践发现,情感符号如表情符号、产品评级等都和情感具有强相关性[10,13]。在社交网络中,很容易获得大量含有情感符号的数据,但是获取高质量的情感标签却很难。在获取标签时寻求一种质量和效率的平衡显得至关重要。自然语言理解是一项很复杂的工作,尽管有很多方法可以处理这一任务,但是选取适合自身任务的方法也是一种挑战。由于应用背景的不同,很难有统一的标准来衡量孰优孰劣。但是大家普遍认可的是群智能要比个体更智慧,这为中文微博分类器的设计提供了一个很好的灵感。

我们把情感定义为“喜”、“怒”、“哀”、“惊”、“惧”、“恶”这六大类。在本文中,我们关注中文微博的情感分类。本文的主要贡献如下:①构建表情符号词典自动获取微博的情感标签,从而节省训练标签获取的人力和财力成本,且具有较高的客观性。②引入修正的G^2检验联合情感词词典进行特征选择。该方法可以保留强分类能力的特征而不至于过过滤,并尽可能消除无效特征的干扰。③使用多阶段判断式抽样策略产生训练集,保证基分类器的多样性。本文方法很容易扩展到流数据并实现并行化。

本文的组织结构如下:在第2节里讨论相关研究工作;在第3节里给出中文微博情感分类器构造的详细步骤;第4节进行实验和讨论;第5节结论和工作展望。

2 相关工作

情感分析分为三个不同的层次:文档级、句子级及短语级。微博情感分析类似于短语级和句子级。近年来,在微博情感分析上出现了很多新技术。在本节中,我们将在标签标注、特征选择和算法设计三方面对相关工作进行讨论和分析。

在传统的情感分析中,标签通常是人工获取的。后来,为了获取高质量的基准评分,亚马逊的土耳其机器人[Amazon's Mechanical Turk (AMT①)]投入使用[14]。人口统计学特征表明,相比较于传统抽样和网络抽样,AMT的参与者更加多样化且具有代表性[15]。随之AMT被广泛应用于多个领域[16,17]。AMT是“人工的人工智能”,由于Twitter上的话题数量庞大,使用这种方法获得足够多的数据来训练分类器显得力不从心。此时,一种新颖的方法即使用情感符号来获取带噪声的标签出现了[13],这最早被用在Twitter的情感分类上,是一种模糊监督学习。类似的工作如[7,18],使用微博结尾的正向情感

① http://aws.amazon.com/mturk/

符号如“:)”、“:-)”作为正向标签,负向符号如“:(”、“:-(”作为负向标签,然后训练出不同的机器学习模型如朴素贝叶斯、最大熵、支持向量机等来处理微博语料。比起人工获取标签,使用情感符号虽然充满噪声但是显著提高了效率。

特征项是文本表达的基本单位,特征筛选独立于分类器之外。特征筛选应遵循以下四条基本原则:①文本可以用选定的特征进行识别;②选取的特征不宜过大;③筛选出的特征应该具备强分类能力;④特征容易获取。常用的特征筛选方法有 TF * IDF、信息增益、互信息、期望交叉熵、卡方检验等。这些方法都是通过特征评估函数对特征进行打分,然后通过阈值的设定,选取一定数目的特征。特征评估函数的选取对于特征选择非常重要,主流的特征评估函数有距离、信息熵、相关性、一致性等[19~21]。

作为最流行的无监督情感分析方法[22,23],基于字典的方法通过文本中的情感词来确定文本的情感或者极性。文献[22]引入词语的语义指向及上下文变化来进行深入分析。文献[24]采用基于词典的方法进行初始的情感分类,但是为了改进低召回率的缺点,人为地对结果进行修正,最后使用修正后的结果训练分类器对新实例进行学习。文献[25]从 WordNet 图中提取出节点权重向量,然后把这些权重引入到 SentiWordNet,从而最终确定 Twitter posts 的极性。文献[23~25]都是改进了的基于词典的学习方法。

无监督的学习方法有很多种。贝叶斯分类器是一种简单的概率分类器[11,26,27],通过从语料中计算先验概率来获取规则。贝叶斯分类器有两种概率模型,一种是基于二项式分布,一种是基于多项式分布。前者是使用词语出现或者不出现作为特征,后者是用词频或者词的 TF * IDF 作为特征。文献[28]表明当词语规模较大时,多项式分布模型表现更优。相比二项式分布它可以平均降低 27% 的误差率。SVM[29]是另外一种流行的情感分类器,其原理是使结构风险最小化。它具有处理高维特征数据的潜力,却无法识别出哪些词对于分类是更重要的。最大熵方法是一种基于分布估计的技术。这种方法假设:当我们对数据知之甚少时,它服从均匀分布,换句话说就是具有最大熵。这种方法不需要假设特征是独立的,但是很容易造成过过滤。文献[30]分析了 PANAS-t、Emoticons、SentiStrength、SentiWordNet、SentiNet、SASA 和 Happiness Index 这些方法的精度和召回,并对其赋以不同的权重,使用多种方法的组合来获取更好的结果。在中文微博的分析上,文献[10,11]给出了典型的无监督学习和有监督学习方法。文献[31]使用带标签的英文数据和分类器做协同训练,从而进行中文微博的分析。

通过对国内外研究现状的分析我们发现,基于词典的无监督学习不能很好处理语境相关的词语,且面对不断涌现的新词热词,情感词典的完备性很难保证。例如,这两句话中“车的空间很大”和“车的油耗很大”,“大”对应的情感极性是完全相反的。再者,如“稀饭”和“喜欢”,“稀饭”是“喜欢”在社交媒体中的新生形式,在中文中二者是近音词且表达的是同样的意思。但是情感词典很难做到完备去覆盖不断涌现的新生情感词。对于无监督的学习方法,获取一组高质量的情感标签耗时耗力,代价昂贵。且目前的情感分类更多的关注于正负两极分类,情绪分类研究较少。无监督学习和有监督学习还面临共同的障碍,即在海量碎片化的微博流数据中筛选出高质量的特征。故而,设计适合中文微博情感分类的方法显得迫切重要。

3 基于修正 G^2 特征筛选的中文微博情感组合分类器构造

本文设计的中文微博情感分类器工作流程如图 1 所示,在后续的章节我们对具体细节进行阐述。

3.1 预处理

因为微博的自然语言特性,我们需要对其进行预处理使其标准化。预处理过程必须且必要,它可以提取出微博的相关内容,抛弃无关的内容。预处理的程度对分类器的精度会产生很大影响,本节对关键预处理步骤进行描述。

(1)去除 URLs,指向符和标签

为了在有限的篇幅内分享更多的信息,很多微博中都会包含 URLs。但是对于情感分析而言,URLs 可能仅仅是陈述事实,对情感分类的贡献非常有限,而且获取 URLs 的内容代价也很大。因此,在预处理阶段,我们将去除所有的 URLs。此外,在微博中,用户会使用“@”在用户名前,用来指向指定用户获取关注。但是这种指向对情感分析是没有意义的,也需要移除。在微博中用“#”作为开始和结束,来表明该条微博属于某个话题。微博的话题归属对情感分析也是没有贡献的,故而我们在该步移

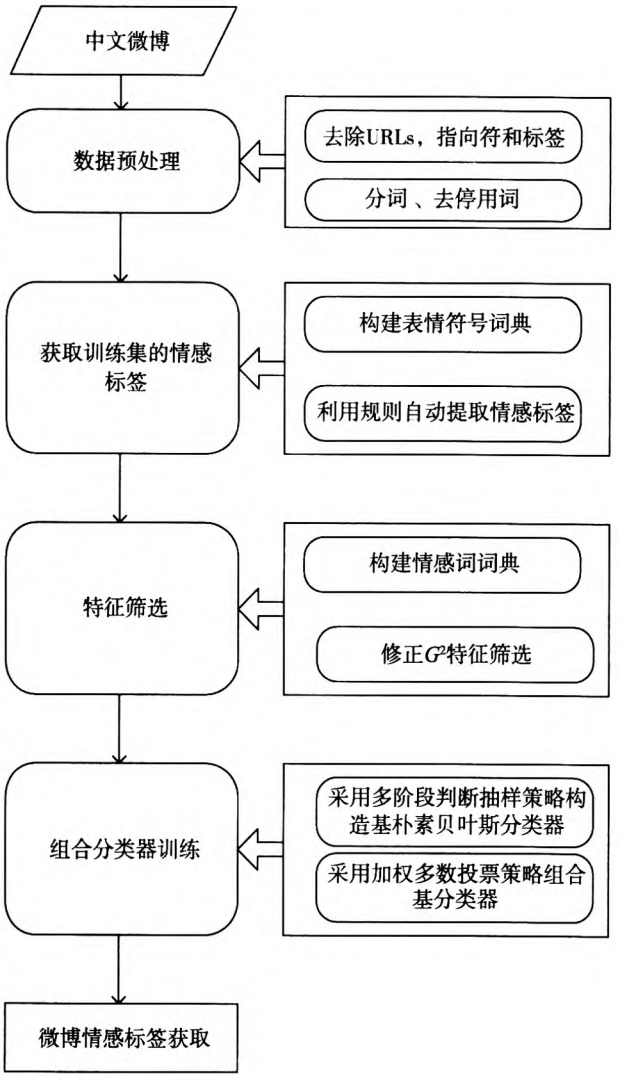

图1 基于修正 G^2 特征筛选的中文微博情感组合分类器工作流程图

除这些标签。

(2)分词、去停用词

词是进行文本处理的最小单位。因为中文是连续书写的,为了进行后续分析必须对文档进行分析获取准确的词边界。在本文中,我们使用 NLPIR (2014 version)系统进行分词和词性标注。

去停用词是文本处理的常规步骤。使用停用词词表去除“的”、“地”、“得”、“是”这些助词,“个”、“条”、“块”这样的量词等。

3.2 训练集情感标签获取

(1)表情符号词典构建

表情符号经常在微博中被用以表达感情,对情感分析极为有用。本文使用新浪微博最常用的表情符号构建表情符号词典。该词典共收录1018个表情符号。类似 AMT,我们设计了一个标注网页,募集志愿者对这些表情符号进行标注。每个 IP 用户只能标注1次,每个表情符号被至少标注100次。获得标注结果后,使用卡方检验来验证标注的有效性,舍弃无效的标注结果。通过这种形式,表情符号被分为6种不同的类别(喜、怒、哀、惊、惧、恶)。表1给出了本文构建的表情符号词典的特征描述。

表1 情感符号词典特征描述

情感	收录表情符号数量	典型表情符号示例
喜	619	
怒	52	
哀	166	
惧	71	
恶	43	
惊	67	

虽然表情符号会随着时间的发展增加或者改变,但是这个增量都是有限的。为了获取更加完备的表情符号词典,可以每间隔一段时间利用同样的方法对词典进行补充和完善。

(2)自动提取微博的情感标签

我们使用表情符号词典来自动提取微博的情感标签。若微博中出现的表情符号属于某一类,该微博就属于某一类。例如:“支付宝收益还不错,心情美美哒”,该条微博包含的表情符号属于“喜”这一类,则这条微博被赋以标签“喜”。在通常情况下,一条微博可能包含不止一个表情符号,此时我们使用如下规则(1)来自动提取微博的情感标签。

$$label = \arg\max_{c}(N_{EC}(c)) \quad (1)$$

这里,$c \in C = \{$喜,怒,哀,惊,惧,恶$\}$,$N_{EC}(c)$是该条微博中属于类别 c 的表情符号的个数。

当微博中属于不同类别的表情符号的个数相等时,使用在最后出现的表情符号所属的类别作为整条微博的类标签。例如:“怎么只有3.8了,”,选取所属的类别作为微博的类标签,其类标签是“哀”。

3.3 基于修正 G^2 检验的特征筛选

(1)情感词词典构建

为了防止在特征筛选中过过滤掉情感词,我们试图建立一个较完备的情感词词典。情感词是进行情感分类的基础,我们选取大连理工大学信息检索

实验室①在数据堂②上发布的情感本体作为起点。将该情感本体中“乐”和“好”合并成“喜”这一类，其余的保持不变，从而得到初步的情感词词典。接下来，我们使用 Hownet③ 和 HIT－CIR④ 的同义词林对该词典进行扩充。最后加入人工干预对该情感词词典进行修正。至此，情感词词典构建完成。表 2 给出了本文构建的情感词词典的特征描述。

表 2　情感词词典特征描述

情感	收录情感词数量	典型情感词示例
喜	10 382	高兴、优秀、表扬、祝福、精彩
怒	1 092	生气、发疯、烦闷、大怒、气愤
哀	3 349	悲伤、痛苦、遗憾、后悔、失望
惧	230	恐慌、恐惧、害怕、吓唬、担心
恶	1 182	憎恶、痛恨、讨厌、反感、厌恶
惊	10 141	奇怪、震惊、奇妙、奇迹、惊讶

(2)修正 G^2 检验特征筛选模型

微博文本经过预处理后，获取了数目庞大的特征。本文中使用 multinomial 模型构建特征值空间。

传统的卡方检验被经常用于特征约减，这种特征约减的方法适用于特征值是 Bernoulli 模型的情况：

$$\chi^2(f,c)=\frac{N(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \tag{2}$$

得分较低意味着特征 f 和类别 c 是独立的。当特征空间很大时，Bernoulli 模型的分类效果逊色于 multinomial 模型。但是传统的卡方检验不适用于 multinomial 模型。本文中我们使用一种新颖的检验方式 G^2 检验[32](pp.611－615) 适用于 multinomial 模型进行特征筛选。

G^2 检验是卡方检验的另一种形式，我们通过一个小例子来简单解释下 G^2 检验。假设样本包含 M 条数据以及两个随机变量 X_1 和 X_2。令 s_i^a 表示样本中满足 $X_i=a$ 的实例数，s_{ij}^{ab} 表示样本中同时满足 $X_i=a$ 和 $X_j=b$ 的实例数。G^2 统计量定义如下：

$$G^2=2\sum_{a,b}s_{ij}^{ab}\ln\left(\frac{s_{ij}^{ab}M}{s_i^a s_j^b}\right) \tag{3}$$

该检验中自由度 df 计算如下：

$$df=(r_i-1)(r_j-1) \tag{4}$$

式中，r_i 是变量 X_i 的取值空间大小。p 是 G^2 检验返回的 p 值，α 是显著性水平。当 $p<\alpha$ 时我们拒绝原假设“X_1 和 X_2 是独立的”。换句话说，如果 $p<\alpha$，则两个随机变量相关，否则，他们独立。

为了适用于文本计算的，我们对 G^2 统计量做出小小的调整。如下：

$$G^2=2\sum_{F,C}s_{F,C}^{f,c}\ln\left(\frac{(s_{F,C}^{f,c}+1)N}{s_F^f(s_C^c+1)}\right) \tag{5}$$

式中，f 是特征集 F 中的元素，c 是类集 C 里的元素，$s_{F,C}^{f,c}$ 是属于类别 c 的语料中出现特征 f 的实例数目，s_F^f 是整个语料中出现特征 f 的实例数目，s_C^c 是语料中属于类 c 的实例数目，N 是整个语料库中实例的数目。引入 Laplace 平滑避免计数为 0 时出现奇异值。我们使用上述方法对特征空间进行约减。为了避免将超低频的有效情感词过滤掉，我们强制保留在情感词词典中出现的特征。整个上述特征筛选的过程我们称之为修正的 G^2 检验。

3.4　组合分类器的构造

在线社交网络为各种实际应用提供了丰富的资源，但是这种数据动态、海量、碎片化。这些特性使得训练单个稳健的强分类器显得力不从心，故而本文关注于训练组合分类器。在本文中，我们使用 AdaBoost 的思想，利用多阶段调整抽样的策略训练不同的基分类器，然后使用加权投票的方法把基分类器得到的结果进行融合。考虑到速度和易操作性，每个基分类器我们使用朴素贝叶斯模型来学习。组合分类器的构造框架如图 2 所示。

符号说明：$D=\{X_i\mid i=1,2,\cdots,N\}$ 是预处理后的语料集。在语料集 D 中有 N 个实例，记为 $|D|=N$。X_i 是由词 $\{w_j\}$ 组成的序列，w_j 是一个词。C 是情感类别集，$c_k\in C, k=1,2,\cdots,6$。

(1) 基朴素贝叶斯分类器的构造

由语料集 D 随机生成训练集 $D_T=\{X_i\mid i=1,2,\cdots,|D_T|\}$。因为在概率上每种情感发生的可能性都是相等的，故 c_k 的先验概率相等，即 $P(c_k)=1/6$，$k=1,\cdots,6$。对于不同的词 w_j，通过它在每类情感 c_k 中出现的次数计算其先验概率，见公式(6)：

$$P(w_j\parallel c_k)=\frac{n^{c_k}(w_j)+1}{\sum_q(n^{c_k}(w_q)+1)} \tag{6}$$

其中，$n^{c_k}(w_j)$ 是包含词语 w_j 的微博在训练集中属于

① http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx

② http://www.datatang.com/data/45805

③ http://www.keenage.com/html/c_index.html

④ http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

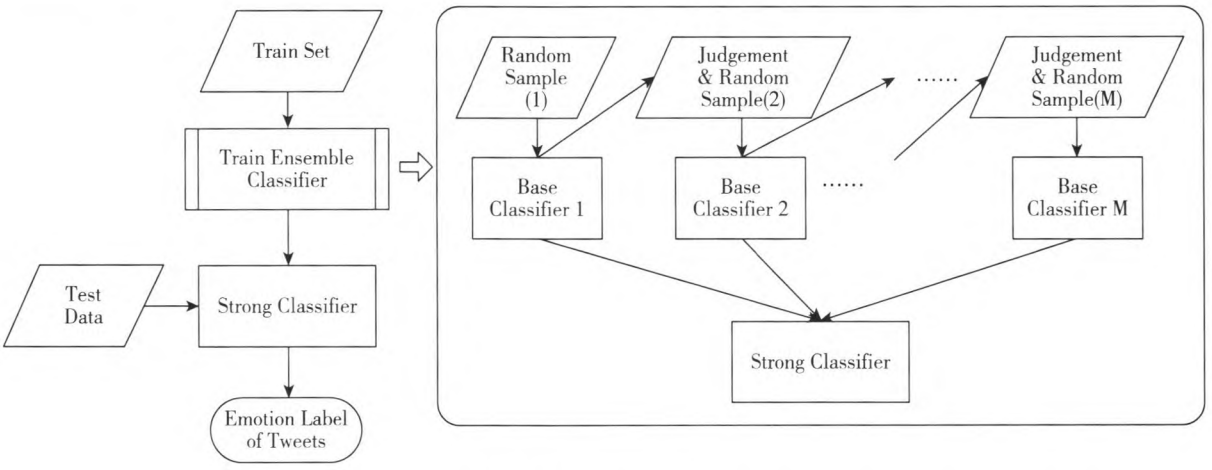

图2 组合分类器构造框架图

类别 c_k 的数目。使用 Laplace 平滑避免出现零值。对于新微博 $X=\{w_j\}$，公式(7)给出了情感标签的学习结果。

$$c(X) = \arg\max_k P(c_k)\prod_j P(w_j \parallel c_k) \quad (7)$$

该步骤训练出了单个基朴素贝叶斯分类器。

为了验证基分类器的性能，随机产生验证集 $D_{test} \subset D - D_T$，$D - D_T$ 表示原语料集去掉训练集 D_T 后的结果。对于验证集中的每条微博 $X_i \in D_{test}$，我们有其原始的类别标签 y_i 以及通过第一个基分类器学习得到的标签 $c(X_i)$。下面给出了度量基分类器性能的测量指标。

定义1. 敏感性：$s_k = \dfrac{n(y_i = c_k \parallel c(X_i) = y_i)}{n^{c_k}}$，(8)

其中，$n(y_i = c_k \parallel c(X_i) = y_i)$ 是在测试集 D_{test} 中同时满足 $c(X_i = y_i)$ 和 $y_i = c_k$ 的实例数目。

定义2. 召回率：$r = \dfrac{1}{6}\sum_{k=1}^{6} s_k$ (9)

定义3. 精度 $p = \dfrac{n(c(X_i) = y_i)}{|D_{test}|}$， (10)

其中，$n(c(X_i) = y_i)$ 在测试集 D_{test} 中满足 $c(X_i = y_i)$ 的实例数目。

定义4. F-measure：$f = 2pr/p + r$ (11)

(2)基分类器组合策略

影响组合分类器性能的因素主要有两个，一个是基分类器的性能，一个是基分类器的多样性。为了保证每个基分类器的性能，引入阈值控制函数 γ。我们希望基分类器的效果至少要比随机猜想好，故只保留结果优于随机猜想的基分类器，即当基分类器的召回 $r > \gamma \geq \dfrac{1}{6}$ 时，保留。为了最大化单个基分类器之间的多样性，我们采取多阶段判断抽样的方式获取训练集 $D_1, D_2, \cdots, D_l$，然后在这些数据集上训练基分类器。通过以上两个措施，得到了一组弱分类器。最后，我们采用加权多数投票的方式对这些基分类器的结果进行融合。即对于新微博实例 $X=\{w_j\}$，其情感类标签

$$c(X) = \arg\max_k\left(\sum_{i=1}^{l} s_k^{(D_i)} P^{(D_i)}(c_k) P^{(D_i)}(w_j \parallel c_k)\right) \quad (12)$$

上标 (D_i) 表示基分类器是在训练集 D_i 上获得的。构建基分类器的过程类似于看医生，患者询问多个大夫获取信息，专家的意见比普通医师更具权威。最后的诊断结果由每个医生的意见加权投票获得。整个组合分类器构建的伪代码如图3所示。

4 实 验

为了对比本文提出的方法和当前主流算法的结果，我们在两个数据集上进行测试。数据集1来源于文献[10, 11]，数据集2是我们编写爬虫，从新浪微博收集而来。在表3中，我们对两个数据集的特性进行了描述。

4.1 G^2 特征筛选对结果的影响

为了对比本文提出的修正 G^2 特征筛选对结果的影响，我们在两个数据集上使用基分类器进行对比实验。我们选择了最常用的两种特征筛选方法，频数和互信息。在频数方法中，我们对候选特征进行频数计数，去掉排名最高的10%和排名最低的20%，只留下中间的70%作为特征集。在互信息方法中，我们抛弃得分最低的30%。在 G^2 检验中，我们设置显著性水平为 $\alpha = 0.05$。

The ensemble classifier algorithm

1. Initialization

   Training set is denote as D_{train}, Test set is denote as D_{test}. We train ensemble classifier on D_{train}.

   Base classifier number $l=1$

   Randomly sample n instances from D_{train}, $n<\frac{|D_{train}|}{2}$,denote as D_l

   Randomly sample $n/4$ instances from $D_{train}-D_l$,denote as T_l

   Let $Temp=\{\}$; The set of falsely predicted instances set $F=\{\}$

2. Train base Naïve Bayes classifier BC_l on D_l

3. Validate the performance of base classifer BC_l on T_l

   Put the falsely predicted instances into F

4. Randomly select $n-|F|$ instances into $Temp$ from $D_{train}-\bigcup_{i=1}^{l}D_i\cup\bigcup_{i=1}^{l}T_i$

5. Update $l=l+1$, $D_l=F\cup Temp$, $Temp=\{\}$, $F=\{\}$

6. If $\left|D_{train}-\bigcup_{i=1}^{l}D_i\cup\bigcup_{i=1}^{l-1}T_i\right|>n$, update T_l, randomly select n instances from $D_{train}-\bigcup_{i=1}^{l}D_i\cup\bigcup_{i=1}^{l-1}T_i$

7. Repeat step 2-6, until $\left|D_{train}-\bigcup_{i=1}^{l}D_i\cup\bigcup_{i=1}^{l-1}T_i\right|\le n$,then update $T_l=D_{train}-\bigcup_{i=1}^{l}D_i\cup\bigcup_{i=1}^{l-1}T_i$

   Execute Step 2 and 3, then break out.

8. Combine all the base classifiers by weighted majority voting.

图 3　组合分类器构造伪代码

表 3　数据集描述

Data	Totle number	Number (joyful)	Number (sad)	Number (anger)	Number (disgusting)	Number (fear)	Number (surprise)	Number (different words)	Number (features)
Dataset 1	368 614	221 877	60 771	56 922	29 044	—	—	171 018	55 110
Dataset 2	220 880	90 771	33 230	52 029	13 570	11 600	19 680	33 508	18 716

为了验证结果的有效性,我们进行了 10 次交叉验证,表 4 和表 5 给出了平均结果。

表 4　不同特征筛选方法对结果的影响(数据集 1)

	不进行特征筛选	频数	互信息	G^2 检验
特征数量	119 712	89 712	76 349	55 110
精度	0.519	0.528	0.659	0.678
召回	0.396	0.301	0.562	0.594
F 值	0.449	0.383	0.607	0.633

表 5　不同特征筛选方法对结果的影响(数据集 2)

	不进行特征筛选	频数	互信息	G^2 检验
特征数量	33 508	203 455	18 236	18 716
精度	0.638	0.645	0.757	0.789
召回	0.589	0.602	0.702	0.72
F 值	0.613	0.623	0.728	0.753

从结果很容易看到,无论哪种方法都提高了结果的精度,且互信息的表现优于频数。G^2 检验的方法明显优于前两者,且适合流特征的筛选。

4.2　分类器性能评估

在本节中,我们对比了有监督学习和无监督学习的两种改进算法,并进行了 10 次交叉验证。①我们使用的是本文 3.1 节中构建的情感词词典,进行无监督的基于词典的学习。②MoodLens:半监督中文微博情感计算方法。在表 6 和表 7 中我们对两个数据集上的结果进行了展示。

很容易看出本文方法的表现整体优于 Lexicon-based 和 MoodLens,后两者都是单个强分类器。结果展示了群智能优于个体智能,组合学习显著改善了情感分类的结果。这个结果很容量理解,在算法的设计阶段,组合分类收集每个基分类器的结果,然后以加权投票的方式进行融合。每个基分类器都是

表6 在数据集1上不同分类器的性能

Method	Precision	Recall 1	Recall 2	Recall 3	Recall 4	Average Recall	*F*
Lexicon-based	0.522	0.576	0.446	0.458	0.392	0.468	0.494
MoodLens	0.643	0.741	0.582	0.497	0.313	0.533	0.583
Our approach	0.678	0.758	0.599	0.563	0.456	0.594	0.633

表7 在数据集2上不同分类器的性能

Method	Precision	Recall 1	Recall 2	Recall 3	Recall 4	Recall 5	Recall 6	Average Recall	*F*
Lexicon-based	0.72	0.766	0.692	0.713	0.644	0.653	0.663	0.688	0.704
MoodLens	0.759	0.891	0.687	0.729	0.594	0.483	0.626	0.668	0.711
Our approach	0.789	0.862	0.729	0.817	0.654	0.55	0.708	0.72	0.753

一个专家,可能某个基分类器给出的结果是错误的,但是组合分类器只有在大多数权威都是错误的时候才返回错误的结果。因为多阶段调整抽样策略的引入基分类器呈现了显著的多样性。

5 结论和工作展望

本文提出了一种新颖的情感分类的策略,该方法适合中文微博流数据。在本文中,构建了情感符号词典用以自动提取微博的情感标签,使用修正的 G^2 检验获得了强区分能力的特征,并巧妙的引入多阶段判断抽样的策略产生多样化的训练集,以保证基分类器的多样性,最后使用加权多数投票的融合方式获取最终的学习结果。理论和实验都表明,本文方法在中文微博情感分类上是一种有竞争力的方法,尽管学习是建立在有噪声的标签之上,该方法在精度和易操作性上获得成功,对商业智能和市场营销具有积极的意义。

本文工作尚有不足之处,其一,在情感标签的标注质量上。获取高质量,噪声小的情感标签是大数据环境下微博情感分类的关键。在未来可以从心理学角度,对表情符号和不同类别的情感之间的相关性上展开更为深入的研究,从理论和实践的双重角度提出更有效的情感标签获取方法。其二,在基分类器的融合策略上。组合分类器提升性能的关键在于基分类器的多样性及不同基分类器的融合策略上。未来可以尝试采取更多样的基分类器构造方式和融合策略,进行知识发现。

参 考 文 献

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C]// Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[2] Oghina A, Breuss M, Tsagkias M, et al. Predicting imdb movie ratings using social media [M]//Advances in information retrieval. Springer Berlin Heidelberg, 2012: 503-507.

[3] O'Connor B, Balasubramanyan R, Routledge B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series[J]. ICWSM, 2010, 11(122-129): 1.2.

[4] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[5] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5 (1): 1-167.

[6] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining [C]//LREc. 2010, 10: 1320-1326.

[7] Bollen J, Mao H, Pepe A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena[J]. ICWSM, 2011, 11: 450-453.

[8] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment[J]. ICWSM, 2010, 10: 178-185.

[9] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data [C]//Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011: 30-38.

[10] Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals[C]//Proceedings of the 22nd international conference on World Wide Web.

International World Wide Web Conferences Steering Committee, 2013: 607-618.

[11] Zhao J, Dong L, Wu J, et al. Moodlens: an emoticon-based sentiment analysis system for chinese tweets [C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1528-1531.

[12] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis [C]// Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005: 347-354.

[13] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009, 1: 12.

[14] De Choudhury M, Counts S, Gamon M. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media[C]//ICWSM. 2012.

[15] Buhrmester M, Kwang T, Gosling S D. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? [J]. Perspectives on Psychological Science, 2011, 6(1): 3-5.

[16] Aldao A, Nolen-Hoeksema S. The influence of context on the implementation of adaptive emotion regulation strategies[J]. Behaviour Research and Therapy, 2012, 50(7): 493-501.

[17] Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk [J]. Behavior Research Methods, 2012, 44(1): 1-23.

[18] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage? [C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1833-1836.

[19] Dash M, Liu H. Consistency-based search in feature selection[J]. Artificial Intelligence, 2003, 151(1): 155-176.

[20] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. The Journal of Machine Learning Research, 2003, 3: 1289-1305.

[21] Wu X, Yu K, Ding W, et al. Online feature selection with streaming features [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35 (5): 1178-1192.

[22] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2): 267-307.

[23] Ding X, Liu B, Yu P S. A holistic lexicon-based approach to opinion mining [C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 231-240.

[24] Zhang L, Ghosh R, Dekhil M, et al. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis[J]. Hp Laboratories Technical Report. 2011 June 21: 1-8.

[25] Montejo-Raez A, Martínez-Cúmara E, Martin-Valdivia M T, et al. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter [C]// Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics, 2012: 3-10.

[26] Prasad S. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods [R]. Technical Report, 2010.

[27] Gokulakrishnan B, Priyanthan P, Ragavan T, et al. Opinion mining and sentiment analysis on a twitter data stream [C]//Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on. IEEE, 2012: 182-188.

[28] Bhuta S, Doshi A, Doshi U, et al. A review of techniques for sentiment analysis Of Twitter data[C]//Issues and Challenges in Intelligent Computing Techniques(ICICT), 2014 International Conference on. IEEE, 2014: 583-591.

[29] Sharma A, Dey S. A comparative study of feature selection and machine learning techniques for sentiment analysis [C]//Proceedings of the 2012 ACM Research in Applied Computation Symposium. ACM, 2012: 1-7.

[30] Gonçalves P, Araújo M, Benevenuto F, et al. Comparing and combining sentiment analysis methods [C]// Proceedings of the first ACM conference on Online social networks. ACM, 2013: 27-38.

[31] Wan X. Co-training for cross-lingual sentiment classification [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 235-243. .

[32] Neapolitan R E. Learning bayesian networks[M]. New York: Pearson Prentice Hall Upper Saddle River, 2004: 611-615.

(责任编辑 贾 佳)