基于监督学习的微博情感分类方法

朱晓光 聂培尧 林培光
(山东财经大学计算机科学与技术学院 山东 济南 250014)

摘 要 随着在线社交网络的快速发展,微博平台上聚集了大量的包含情感的主观句。微博情感可影响受众的观点形成,作用于商务智能、政策制定,甚至是股票市场。微博情感分类是指如何从微博中自动抽取出情感极性和不同的情感分类,如喜爱、愤怒、惊奇等。结合情感词汇本体和同义词词林,从微博中抽取不同类别的特征,运用监督学习方法进行情感分类,在学习过程中优化不同的模型,并分别进行误差和拟合分析,比较不同模型的性能。分类算法在 NLP&CC 2013 的评测任务中取得了具有竞争性的结果。

关键词 微博 情感分类 监督学习 情感词汇本体 同义词词林

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.08.057

SUPERVISED LEARNING-BASED MICROBLOGGING SENTIMENT CLASSIFICATION METHOD

Zhu Xiaoguang Nie Peiyao Lin Peiguang
(*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, Shandong, China*)

Abstract With the rapid development of online social networks, microblogging platforms gather a lot of subjective sentences which contains the sentiment. The sentiment in a microblog can affect the formation of audiences' opinion, and can act on business intelligence, policy development, and even the stock market. Microblogging sentiment classification refers to how to automatically extract the emotion polarity and different emotional categories, such as love, anger, surprise, etc, from microblogs. Combined with emotional lexicon ontology and synonymy thesaurus, we extracted the characteristics with different categories from microblogging text, and used the supervised learning method to classify the microblogging sentiment. Different models are optimised in learning process. The error analysis and fitting analysis are processed respectively as well, and the performances of different models are compared. This classification algorithm achieved competitive result in NLP & CC 2013 evaluation task.

Keywords Microblogging Sentiment classification Supervised learning Emotional lexicon ontology Synonymy thesaurus

0 引 言

互联网的快速发展使用户有更多的途径在互联网上表达自己的观点,观点也成为 Web 文本中的一种重要信息。尤其是在线网络的发展过程中,用户不仅表达自己的观点与情感,同时受 Web 中已存在观点的影响。在过去的十年里,非结构化数据,尤其是 Web 文本的情感分析得到了越来越多的关注[1]。

本文主要针对微博中的文本进行情感分类,使用的语言资源有情感词汇本体和同义词词林。对于情感的划分,本文采用了情感词汇本体中的 7 类情感,分别是:愤怒、厌恶、恐惧、高兴、喜好、悲伤、惊讶,这是在 Ekman 的 6 大类情感的基础上划分的[2]。《同义词词林》是梅家驹[3]等人于 1983 年编纂而成,这本词典中不仅包括了词语的同义词,也包含了一定数量的同类词,即广义的相关词。

微博情感分析的相关研究中,Tsytsarau[1]等人比较了过去十年的情感分析算法并比较了算法性能,将其分为词典、统计、语义和机器学习四类。其中,词典和机器学习占较大比重,同时近年来大量的研究转向了微博平台 Twitter。事实上,在使用机器学习进行多情感分类时,需要使用基于词典或规则的方法来获得情绪值。Pang[4]等提出并评估了三种监督分类方法:朴素贝叶斯、最大熵和支持向量机(SVM),其中 SVM 取得了最好的性能。

除了监督学习方法,Zhou 等[5]使用一种新的半监督学习方法 ADN 来解决标注数据不足的问题,在半监督学习过程中,应用 Active Learning 来选择待标记的数据,文中还提出了一种 Information ADN 方法,根据原始数据中的信息密度来选择需要标注的数据。除此之外,表情符号还可以用来获得情感标注数据,Pak 等[6]使用两类表情符号:快乐和悲伤,来收集微博情感预料,训练分类器并用以识别正性与负性情感。孙艳等[7]提出了一种无监督的主题情感混合模型(UTSU 模型),对每个句子采样情感标签,对每个词采样主题标签,无需对样本进行标注,就可以得到各主题的情感词,从而对文档进行情感分类。

收稿日期:2013 - 10 - 21。教育部人文社科一般项目(10YJC880076);教育部人文社会科学专项任务项目(12JD710120,13JDSZ2084);山东省自然科学基金项目(ZR2010FL008);济南市科技局科技明星计划项目(2013010)。**朱晓光**,硕士生,主研领域:智能信息处理。**聂培尧**,教授。**林培光**,副教授。

情感分析的另一个主要问题是文本特征抽取,除了传统的 N-gram 模型,在文本情感分析的研究中,句法、文体特征和情感词典等特征都已有广泛的应用。Zhai 等[8]研究了中文情感分类中的特征选择,提出了使用子字符串特征来进行情感分析,并通过字符串聚合来降低特征维度。此外,还有表情符号和主题等特征,在 Fu[9]的 MSA-COSRs 模型中,使用 LDA 模型获取文档的全局主题,然后抽取出与情感有关的局部主题来进行情感分析,以获得评论的情感极性。谢丽星等[10]在微博情感分析中也引入了主题相关特征,并识别了微博的主题发散性特征,如微博中不同句子属于不同主题、微博与其评论的主题不相同等。在微博的情感特征抽取中,还包含与其平台特性相关联的特征,如缩略词、表情符号和用户特征等。

在线网络的情感分析中,大部分方法用于分析文本的情感极性及其强度,准确度较高。主要方法包括 SVM、多项式朴素贝叶斯(MNB)、PMI。除了情感极性程度的分析,情感分析还包括多类别情绪分析,与情感极性分析方法类似,其方法包括情绪规则库、情绪细化和机器学习。张晶等[11]以情绪因子中的常用情绪词和情绪短语为基础构建情绪词典,并针对特殊的情绪表达形式,结合标点符号和表情符号在情绪分析中的功能,建立情绪规则库。欧阳纯萍等[12]提出一种基于多策略融合的细粒度情绪分析方法以细化情感,如高兴、失望等,并结合机器学习进行情感分析。其中,微博的短文本特性给情感分类带来的巨大的挑战,使得文本特征无法充分表达微博的情感[13]。针对微博的特征,本文使用同义词词林对微博的词项进行转化,并使用情感词汇本体提取词汇的情感值。在 2013 中文信息处理的会议中,CCIR-COAE、NLP&CC 都有情感分析的评测任务,NLP&CC 的情感分类评测任务中,将微博文本标注为了 7 类情感来进行情感分类。

1　监督学习模型

本文使用了前馈神经网络和 SVM 作为监督学习模型。其中,单隐层神经网络具有很强的学习能力,能够逼近复杂非线性函数,同时能够解决传统参数方法无法解决的问题。相对于神经网络,SVM 不需进行复杂的模型调整,但求解多类别分类任务需要训练多个不同的模型。

1.1　前馈神经网络

本文使用的前馈神经网络的基本模型[14]如图 1 所示。

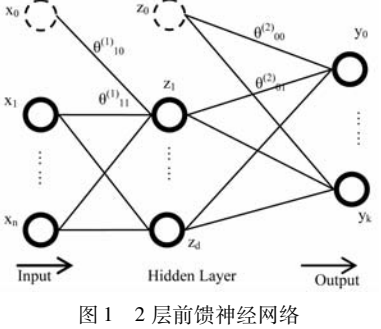


图 1　2 层前馈神经网络

其中,箭头表示信息传递的方向,误差则按其反向传递;x 表示神经网路的输入,包含 2118 个特征和 1 个误差调整结点 x_0,即:$x \in \mathbb{R}^{2119\times 1}$;$z$ 表示隐藏结点,其个数为 $d+1$,通过交叉验证进行调整;$y \in \mathbb{R}^{8\times 1}$,包含 7 类情感和无情感 None;$\theta$ 表示需要被优化的参数,其中 $\theta^{(1)} \in \mathbb{R}^{d\times n+1}$,$\theta^{(2)} \in \mathbb{R}^{k\times d+1}$,$\theta_{ij}^{(c)}$ 表示第 $c+1$ 层的第 i 个结点到第 c 层的第 j 个结点的连接。

根据图 1 的学习模型,第 k 个输出结点 y_k 的输出如下:

$$y_k(x,\theta) = \sigma\left(\sum_{j=1}^{d}\theta_{kj}^{(2)}h\left(\sum_{i=1}^{n}\theta_{ji}^{(1)}x_i + \theta_{j0}^{(1)}\right) + \theta_{k0}^{(2)}\right) \quad (1)$$

其中,$\sigma(a) = 1/(1+e^{-a})$,即 logistic 函数,$h(\cdot)$可以选择 logistic 函数或 tanh 函数,本文选择了前者作为隐层结点的转换函数。关于前馈神经网络的成本函数和误差传递,可以参照文献[14]。

1.2　支持向量机

SVM 模型是最优边界分类器的一种,与前馈神经网络的形式不同,其目标函数的形式如下:

$$w,b = \arg\max_{w,b}\left\{\frac{C}{\|w\|}\min_n[t_n(w^{\mathrm{T}}\phi(x_n)+b)] + \frac{1}{2}w^{\mathrm{T}}w\right\} \quad (2)$$

其中,w、b 为 SVM 模型的参数,C 为正则化参数,与神经网络中的 λ 的倒数 $1/\lambda$ 功能相同,$t_n \in \{-1,1\}$,即第 n 条训练数据的标注值,对于新数据 x 的预测值形式为参数 w 的仿射函数。在实际应用中,可以采用已有框架进行 SVM 模型训练,本文采用 Libsvm[15]框架,并结合 one-vs-the-rest 方法进行多类别情感分类,对情感类逐一建立模型,应用其概率输出计算最终结果。对于新数据 x,根据模型计算 $y_k(x)$,其值相当于神经网络输出层的输入值,x 的最终预测值为:

$$k = \arg\max_k(y_k(x)) \quad (3)$$

2　监督学习过程

2.1　微博文本预处理

微博平台提供一种即时消息发布服务,其文本信息一般不超过 140 字。例如:“人生路上,我们会无数次被自己的决定或碰到的逆境击倒”。本文首先对其微博文本进行分词和词性标注,分别使用斯坦福大学的分词和词性标注工具。中文分词基于条件随机场序列模型,分词准确率约为 95%[16]。词性标注使用了条件对数模型,同时结合多种词汇特征和领域词汇的词性[17]。词性标注工具可以直接标注中文句子,但不如结合中文分词使用时的准确度高。

本文使用 s 表示训练数据集中的一条微博,$w = [w_1, w_2, \cdots, w_d]$表示对 s 分词后的词项序列,$t = [t_1, t_2, \cdots, t_d]$表示 s 的词性序列,d 表示 s 分词后的词项个数,包括标点和特殊符号。一条微博的分词结果示例如下:

人生 路/NN 上/LC ,/PU 我们/PN 会/VV 无数/CD 次/M 被/SB 自己/PN 的/DEG 决定/NN 或/CC 碰到/VV 的/DEC 逆境/NN 击倒/VV

本文使用的词性为 34 种,比最初的词组结构标注所使用的标注集多出一个 URL 标签。词性标注中,更多的标签可以提供更有效的信息,但会增加标注的错误率。

2.2　特征抽取

微博特征的选择结合了同义词词林[3]和情感词汇本体[2],前者对微博词项进行概念化,用于抑制学习算法的过拟合,后者用于词汇情感值的获取。同义词词林对词汇进行三层编码,示例如下:

Aa01 = 高个子　矮子　巨人……;Aa03 = 美女　丽人……

其中第 1 层用大写英文字母表示;第 2 层用小写英文字母表示;第 3 层用二位十进制整数表示。此外,哈工大对同义词词林进

行了扩展,增加了词群和原子词群两层编码:第 4 层用大写英文字母表示;第 5 层用二位十进制整数表示[18]。例如:

Aa01A05 = 匹夫 个人;Aa01C01 = 众人 人人 人们

本文采用三层编码的同义词词林,使用 $Syno(w)$ 表示词项 w 的同义词编码,例如:

$$Syno(\text{巨人}) = Ac01$$

情感词汇本体将词汇分为 21 个情感小类,如快乐(*PA*)、尊敬(*PD*)、愤怒(*NA*)。除了情感类,情感词汇本体还包括情感极性和强度两个维度,此外,有的情感词汇还存在辅助情感分类。本文不区别对待词汇的主从情绪,将其进行合并,示例如表 1 所示。

表 1 情感词汇本体示例及表示

编号	词汇	情感分类	强度	极性
i	sw	Emotion(sw)	Strength(sw)	Polarity(sw)
1	脏乱	NN	7	2
7	战祸	ND	5	2
7	战祸	NC	5	2
11	清莹	PH	5	1
214	祭拜	PD	5	0

其中,*sw* 表示一个情感词汇,$Emotion(sw)$ 表示词项 *sw* 的情感类,$Emotion \in \{PA, PE, \cdots, NL, PC\}$;$Strength \in [1,9]$,表示情感强度,其中 1 表示情感最小,9 表示强度最大;$Polarity \in \{0,1,2,3\}$,分别表示中性、褒义、贬义和兼有褒贬两义。在情感词汇本体中,词汇数量最多的是褒义词,最少的是双性词,只有 78 个。

本文抽取了 5 组特征,特征矩阵 X 的示例如表 2 所示。

表 2 特征矩阵 *X* 的形式

	F1		F2		F3		F4	F5
w	$F1_{PA}$	$F1_{PD}$	$F2_{RARE}$	$F2_{Ac02}$	$F3_{RARE}$	$F3_{Ac02}$	$F4_{NN}$	$F5_1$
$W^{(1)}$	1	0	0	2	1	0	2	2
$W^{(2)}$	0	1	1	0	0	0	4	5
$W^{(3)}$	3	1	2	0	1	0	5	3

1) 特征组 *F*1 表示微博 w 中不同情感小类的词汇数量,其运算公式如下:

$$F1_{em}^{(w)} = \sum_{i=1}^{d} I(Emotion(w_i) = em) \quad em \in Emotion \tag{4}$$

其中 d 表示微博 w 的词项数,$I(dis)$ 表示指示函数,当判别式 *dis* 为真时 $I(dis)$ 等于 1,否则为 0。特征组 *F*1 有 21 个特征,对应情感词汇本体中的 21 个小类。

2) 特征组 *F*2 表示微博中不同词林的词汇数量:

$$F2_{syn}^{(w)} = \sum_{i=1}^{n} I(Syno(w_i) = syn) \quad syn \in Syno \tag{5}$$

其中 $Syno(w)$ 表示词项 w 的同义词词林编码。*Syno* 表示 *F*2 所使用的同义词词林编码集合,是满足以下条件的编码 *syn* 的集合:

$$\sum_{w \in trainset} \sum_{i=1}^{d} (I(w_i) = syn) \geqslant \alpha \tag{6}$$

其中 α 表示特征选择所使用的阈值,在训练集中出现频率低于 α 的同义词词林编码将被 *RARE* 替代,如表 2 中特征组 *F*2 和 *F*3 的第一列。α 可以有效地抑制训练过程中的过拟合,同时提高算法效率,其值的设置可以通过交叉验证或人工设置,本文手动将 α 设置为 5。

3) 特征组 *F*3 表示微博中不同词林的情感词汇数量:

$$F3_{syn}^{(sw)} = \sum_{i=1}^{d} I(Syno(sw_i) = syn) \&\& I(Emotion(sw_i) \in Emotion) \tag{7}$$

其中,*sw* 表示情感词汇,即情感词汇本体中的词汇。*F*3 也可以设置特征选择的阈值,与 *F*2 的方法相同,使用 *RARE* 代替出现频率较的情感词的词林编码。本文手动将 *F*3 的阈值 α 设置为 2。

4) 特征组 *F*4 表示微博中不同词性的词项数量,使用词性标注结果的词性集合,例如 $F4_{NN}^{(w)} = 2$ 表示微博 w 中有两个普通名词,$F4_{IJ}^{(w)} = 1$ 表示 w 中有 1 个感叹词。为了使程序的健壮性更好,我们在 *F*4 中也加入了 *RARE* 特征,防止由于预处理中的异常而出现的其他词性。

5) 特征组 *F*5 表示微博中不同极性的词项的数量,例如:$F_1^{(w)} = 6$,表示 w 中有 6 个中性词。

特征矩阵 X 的 5 组特征分别包含 21、1348、710、35、4 个特征,总特征数为 2118。

2.3 模型优化

使用监督学习进行情感分类的第一步是对标注集的数据进行预处理,得到微博的词项序列 W 和词性序列 T。然后使用上文的特征提取方法从 W 和 T 中提取出不同的特征,合并成特征向量 X。对 X 进行归一化,本文以特征组为粒度,对不同特征组的特征进行了单独归一化处理;然后对 X 进行分割,形成训练集 X_{train} 和验证集 X_{val}。将标注值转换为向量形式,如 $t = 2$ 转换为 $t = [0,1,0,0,0,0,0]$。X 和 t 的形式如下:

$$\begin{aligned} X &= [X^{(1)}, X^{(2)}, \cdots, X^{(n)}]^{\mathrm{T}} \quad X \in \mathbb{R}^{n \times m} \\ t &= [t^{(1)}, t^{(2)}, \cdots, t^{(n)}]^{\mathrm{T}} \quad t \in \mathbb{R}^{n \times K} \end{aligned} \tag{8}$$

其中 n 是标注集中微博的数量,m 是特征维度,K 是类别数。划分出训练集后,选择学习模型进行优化。对于前馈神经网络,其过程如下:

1) 选择参数 d 和 *iter* 来训练神经网络,根据预测结果设置模型参数的优化区间,选择最优的参数。除代码中给出的参数,还包括成本函数中的正则化参数 λ,用于决定误差和参数值在成本中的权重。本文未通过交叉验证来获得参数 λ,而是手动将 λ 设置为 1。此外本文选择了较小的参数 d 和 *iter*,因为特征矩阵 X 是稀疏矩阵,同时较少隐层结点能在很大程度上提高学习效率。此外,标注集的规模有限,较多的隐层结点容易造成训练过程中出现过拟合。

2) 对每一组参数,运行一轮完整的训练和预测,得到这一组参数对应的交叉验证误差。训练神经网路时,首先随机初始化参数 $\theta \in [-0.1, 0.1]$。初始化 θ 后,选择隐层节点数 d 和迭代次数 *iter* 来训练模型。训练结束后,根据式(1)得到神经网络的输出 $y_{out} \in \mathbb{R}^{(0.2 \times m) \times 8}$,然后计算:

$$y_{d,iter} = \arg\max_{i \in [1,8]} (y_{out}^{(i)}) \tag{9}$$

在交叉验证阶段,预测的准确度可以取正确的预测结果所占的比例,也可以按情感分类取 F_β 值。

3) 选定参数 d 和 *iter* 之后,使用标注集 X 进行神经网络的训练,得到优化的参数 $\theta^{(1)}$ 和 $\theta^{(2)}$。对测试数据进行相同的数据预处理和特征选择,使用神经网络进行分类,输出模型参数于预测结果 y,算法结束。

根据上述描述,前馈神经网络的模型优化的算法如下:

Inputs: Training set S,t
Initial: Randomly initial θ
Define: $[X_{train}, X_{val}]$ = feature(S)
Algorithm: for d in {25, 50, 80, …}, iter in {50, 100, 150,…}:
For i in [1,iter], x in X_{train}
θ = adjust(x,θ,t,d)
$y_{d,iter}$ = predict(X_{val},θ)
end for
end for
[d, iter] = $\underset{d,iter}{arg\ min}$error($y_{d,iter}$,$y_{gold}$)
θ = train(X,θ, t, d, iter)
y = predict(X_{test},$\theta^{(1)}$,$\theta^{(2)}$)
Output: θ,d,y

其中,S 表示标注的微博语料,X_{train},X_{val}分别表示训练集、交叉验证集;$t \in [1,8]$ 表示人工标注的微博情感,其中 1 ~ 7 表示 7 类情感,8 表示无情感 None;d 表示隐层结点数,*iter* 表示迭代次数;*adjust* 表示参数 θ 的调整过程,即一次误差反向传递过程;*train* 表示一次训练过程。

对于 SVM 模型,优化过程比较直观。首先对处理训练数据的目标值,在训练类别 k 的模型时,将目标值 $t=k$ 的数据的目标值设置为 $t=1$,其余的设置为 $t=-1/(k-1)$;然后设置核函数和概率输出,对每个类别进行模型优化;在新数据的预测过程中,将新数据输入每个类别的模型中,得到 $y_k(x)$,然后使用式(3)计算情感类别。

2.4 算法复杂度

在情感分析过程中,中文分词具有较高的空间复杂度,主要的空间消耗在于中文分词模型和词典的读取。对于分词后的微博文本,其词性标注复杂度较低,使用动态规划可以将在线性时间内找到最优词性标注序列[17]。

神经网络具有很强的学习能力,但结点的增加会使算法的复杂度快速增加,尤其是隐层结点个数。当神经网络的结点较少时,会有较快的收敛速度。但是当结点增大到一定规模时,收敛速度会急剧下降,从而导致一些神经网络算法不可用。如模型优化的算法所示,迭代次数 *iter* 和训练集规模 m 会使算法的复杂度线性增长,隐层结点的增加会使连接数线性增长,并使得算法的复杂度非线性增长。

SVM 算法的时间复杂度一般在 $O(n^2)$ 和 $O(n^3)$ 之间,有些算法的复杂度为 $O(n^2v)$,其中 n 为训练集的规模,v 为支持向量个数。另一方面,SVM 的复杂度很大程度上依赖核函数的复杂度。在多类别分类任务中,SVM 的时间复杂度与类别数目成正比,如 one-vs-the-rest 方法。

图 2 对比了不同分析过程的时间消耗,其输入为相同的特征矩阵,其中,$F[60]$ 表示隐层节点数为 60 的 FNN 模型,$S[Radial]$表示核函数为径向基函数的 SVM 模型。由图 2 可以看出:文本预处理具有较高的时间消耗;FNN 的优化时间随隐层节点数的增加而非线性增长;不同核函数的 SVM 模型优化时间不同。图中 FNN 迭代了相同的次数,并未收敛到一定的误差。

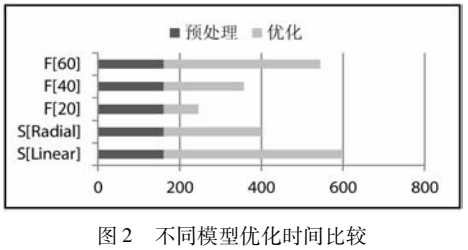


图 2　不同模型优化时间比较

3 实　验

本实验所使用的数据来自 NLP&CC 2013 评测任务:中文微博情绪识别,是对一条微博的整体情感进行识别。标注集有4000条微博,共 371 697 字。提取出的特征矩阵 $X \in \mathbb{R}^{4\,000\times 2\,118}$,标注结果矩阵 $t \in \mathbb{R}^{4\,000\times 1}$,选择参数 d 和 *iter* 的不同组合 $d \in \{25,50,80\}$,$iter \in \{10,20,\cdots,150\}$。

本文中交叉验证和测试数据的预测准确度采用宏平均和微平均 F_β 值,其中 β 表示重要性因子,用于调节准确率和召回率的权重,宏平均给予每个情感分类相同的权重,而微平均给予每条微博相同的权重[19]。本文中宏平均的计算公式如下:

$$P = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{n}\frac{I(t^{(j)})\&\&I(y^{(j)}=j)}{I(y^{(j)}=j)+c} \tag{10}$$

$$R = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{n}\frac{I(y^{(j)})\&\&I(t^{(j)}=j)}{I(t^{(j)}=j)+c} \tag{11}$$

$$F_\beta = \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R} \tag{12}$$

其中,t 是标注值,y 是预测值;n 是预测数据的大小,K 是类别数,此处未包含类别 None;c 是平滑因子,取较小的正值,用于频率较小的情感类出现除零异常,如恐惧和惊奇。使用宏平均 F_1 值运行交叉验证得到图 3 所示的学习曲线。

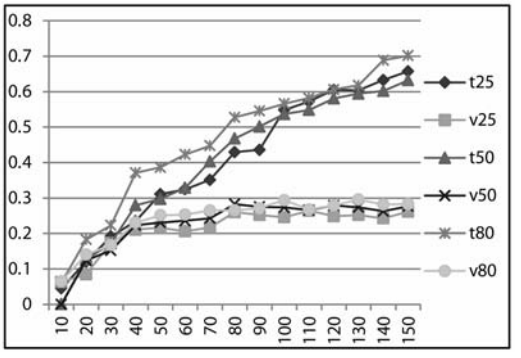
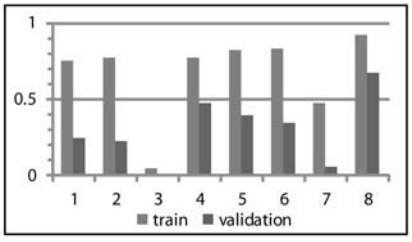


图 3　神经网络模型的学习曲线

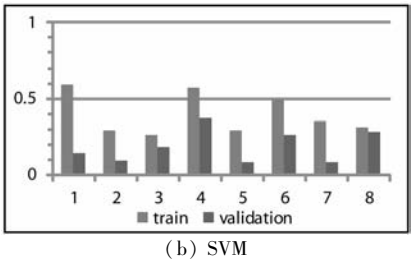
其中横轴表示迭代次数,纵轴表示宏平均值,*t*25、*v*25 分别表示隐层结数 $d=25$ 时训练集、交叉验证集的平均值。从图 3 可以看出,不同隐层结点的神经网络均出现了不同程度的过拟合,主要原因在于:1) 微博的特征数接近标注集的规模,导致神经网络快速的拟合训练集。2) 本文选取的情感词和词性等特征不能充分的表现微博的情感,尤其是出现频率较小的情感类。3) 成本函数中的正则化参数 λ 偏小,同时,在数据和特征不充分的情况下,仅调整 λ 不足以提高交叉验证的 F_1 值。

在交叉验证中,除了宏平均之外,图 4 给出了各情感类的 F_1 值,其中横轴表示情感类,纵轴表示 F_1 值。从图中可以看出训练集的 F_1 值明显高于交叉验证集,并且交叉验证的 F_1 值波动较大,这是由于部分情感出现频率较低影响了预测的召回率,从而降低了 F_1 值。

此外,本文使用的前馈神经网络在微博情感分类的性能上要优于 SVM。对比不同类别的准确度可以发现,SVM 在低频率情感类的预测中性能较好,拟合程度较之神经网络要低。完成交叉验证之后,使用全部标注集训练神经网络,测试集的预测结果如表 3 所示。



(a) 前馈神经网络



(b) SVM

图4 各情感类的 F_1 值

表3 测试集预测结果与对比

评测对象	正确率	召回率	F_1 值
主观句判断	0.7479	0.6335	0.6866
宏平均	0.2653	0.2201	0.2406
微平均	0.3655	0.309	0.3349
宏平均 max	0.2704	0.3064	0.2873
宏平均 avg	0.2145	0.1933	0.2033
微平均 max	0.3133	0.3746	0.3412
微平均 avg	0.2481	0.2684	0.2579

其中,前三行为本文的预测结果,后面为 NLP&CC 评测的最优结果和平均结果。从表3可以看出,FNN 预测的结果中,正确率明显高于召回率;情绪句判断具有一定的误差,这也直接影响了情绪分类的召回率,因为情绪句已被识别为无情感 None;情绪分类的微平均大于宏平均 F_1 值,这是因为宏平均的计算中赋予每类情感相同的权值,因此各情感类的 F_1 值的波动会降低宏平均 F_1 值,如图4(a)中第3类情感的交叉验证 $F_1=0$;另一方面,F_1 值较小的情感类出现频率较低,因此微平均权重较低,使得情绪分类的微平均 F_1 增加。

本文在情感分析的过程中,结合同义词词林进行特征抽取,将词项映射到同义词词林编码,使用词林编码作为特征空间的维度,提高了优化效率并抑制了拟合现象。同时使用情感词汇本体提取词汇的情感类和强度,组合成不同特征并分组进行特征矩阵的归一化处理。在模型优化阶段,借助不同的监督学习模型进行分类并优化模型参数。对比表3的数据可得,本文的微平均值取得了较好的预测结果。

4 结 语

本文应用监督学习算法进行微博的情感分类,取得了一定的效果。但相对于语音,微博文本没有语调和语速等表达情感的重要特征,同时文本中大量的修辞也给情感分析带来了巨大的挑战。因此,应用监督学习进行情感分类,还需要大量的标注集和更深层的情感特征。对于微博的情感分析,其文本的内在特征也非常重要,此外还有微博用户的特征、微博的主题,都在不同程度上影响微博文本的情感分类。

从实验结果来看,本文的学习模型存在一定的过拟合现象,并且微博情感分类的准确度较低。因此,在下一步的研究中除了情感的语言特征的发现和情感标注,还需要利用半监督学习方法提取微博文本的情感特征,如微博的主题和语境,以提高情感分类的准确度。

参 考 文 献

[1] Tsytsarau M, Palpanas T. Survey on mining subjective data on the web [J]. Data Mining and Knowledge Discovery, 2012, 24(3): 478-514.

[2] 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.

[3] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林[M]. 上海辞书出版社, 1983.

[4] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Found. Trends Inf. Retr., 2008, 2(1-2): 1-135.

[5] Zhou S, Chen Q, Wang X. Active deep learning method for semi-supervised sentiment classification[J]. Neurocomputing, 2013, 120: 536-546.

[6] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]. LREC, 2010, 17-23, May 2010, Valletta, Malta.

[7] 孙艳, 周学广, 付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报: 自然科学版, 2013, 49(1): 102-108.

[8] Zhai Z W, Xu H, Kang B D, et al. Exploiting effective features for chinese sentiment classification[J]. Expert Systems With Applications, 2011, 38(8): 9139-9146.

[9] Xianghua F, Guo L, Yanyan G, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37: 186-195.

[10] 谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报, 2012, 26(1): 73-83.

[11] 张晶, 朱波, 梁琳琳, 等. 基于情绪因子的中文微博情绪识别与分类[J]. 北京大学学报: 自然科学版, 2014(1): 79-84.

[12] 欧阳纯萍, 阳小华, 雷龙艳, 等. 多策略中文微博细粒度情绪分析研究[J]. 北京大学学报: 自然科学版, 2014(1): 67-72.

[13] Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network[J]. Expert Systems with Applications, 2013, 40(16): 6266-6282.

[14] Bishop C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. Springer-Verlag New York, Inc., 2006.

[15] Chang C, Lin C. LIBSVM: A Library for Support Vector Machines [J]. ACM Trans. Intell. Syst. Technol., 2011, 2(3): 21-27.

[16] Tseng H. A conditional random field word segmenter for sighan back off 2005[C]//Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 2005, 171.

[17] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 173-180.

[18] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602-608.

[19] Aixin S, Ee-Peng L. Hierarchical text classification and evaluation [C]//Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001: 521-528.