

基于 OCC 模型和贝叶斯网络的情绪句分类方法



徐源音¹ 柴玉梅¹ 王黎明¹ 刘 箴²

1 郑州大学信息工程学院 郑州 450001

2 宁波大学信息科学与工程学院 浙江 宁波 315211

(lucyquanquanxu@163.com)

摘 要 情绪句分类是情绪分析研究领域的核心问题之一,旨在解决情绪句类别的自动判断问题。传统基于情绪认知模型(OCC 模型)的情绪句分类方法大多依赖词典和规则,在文本信息缺失的情况下分类精度不高。文中提出基于 OCC 模型和贝叶斯网络的情绪句分类方法,通过分析 OCC 模型的情绪生成规则,提取情绪评估变量并结合情绪句中含有的表情符号特征构建情绪分类贝叶斯网络;通过概率推理,可以实现句子级文本的情绪分类,并减小句中信息缺失所带来的影响。与 NLPCC2014 中文微博情绪分析评测的子任务情绪句分类评测结果的对比表明,所提方法具有有效性。

关键词: 情绪分析;OCC 模型;贝叶斯网络;情绪句分类

中图法分类号 TP391

Emotional Sentence Classification Method Based on OCC Model and Bayesian Network

XU Yuan-yin¹, CHAI Yu-mei¹, WANG Li-ming¹ and LIU Zhen²

1 School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

2 School of Information Science and Technology, Ningbo University, Ningbo, Zhejiang 315211, China

Abstract Emotional sentence classification is one of the core problems in the field of emotional analysis. It aims to solve the problem of automatic judgment of emotional sentence categories. Traditional emotional sentence classification methods based on OCC sentiment recognition models mostly rely on dictionaries and rules. In the absence of textual information, the classification accuracy is relatively lower. This paper proposed an emotional sentence classification method based on OCC model and Bayesian network. By analyzing the emotion generation rules of OCC model, it extracts emotional assessment variables and combines the emotion features contained in the emotion sentence to construct a Bayesian network of emotion classification. Through probabilistic reasoning, it is possible to identify a variety of emotion categories that the text may want to express and reduce the impact of missing text information. Compared with the NLPCC2014 Chinese Weibo emotion analysis evaluation sub-task emotional sentence classification evaluation results, the results show that the proposed method is effective.

Keywords Emotional analysis, OCC model, Bayesian network, Emotion sentence classification

1 引言

文本情绪分析^[1]是自然语言处理领域的一个热门研究方向,其主要任务是自动识别用户在文本中所表达的情绪类别。情绪^[2]是多种感觉、思想和行为综合产生的心理和生理状态,泛指喜欢、厌恶、愤怒、惊讶、伤心等。如今,互联网已成为人们获取信息、反馈信息的平台,针对热点话题,人们可发表自己的观点,随意评论商品的好坏。情绪分类是情绪分析领域一直被关注的热点问题,具有很大的研究价值和应用价值,从这些海量数据中捕获有效信息、了解用户情感,将有助于政府部门的舆情监控以及企业和个体的管理与决策。快节奏的生活培养了大众碎片化阅读的习惯,且发表的网络言论通常较短。例如,微博的长度大多不超过 140 字;其评论则更为精

简,一般表现为简单句、语气词、网络用语等。现有的情绪分类方法按文本规模可分为篇章级、句子级、词语级。从简短的句子、随意组合的网络用语中正确识别文本所表达的情绪,实现句子级文本的情绪分类,是情绪分析的一项基础任务。

传统的情感分析^[3]主要是指识别用户的情感倾向,即正向、负向或者中性;情绪分析则是细粒度的情感分析任务,可以识别多种情感。情绪分析的研究方法有基于词典和规则的方法、基于机器学习的方法、基于情绪认知 OCC 模型^[4]或者情感和适应性模型 EAM(Emotion and Adaption Model)^[5]的方法等。Wen 等^[6]提出了一种基于类序列规则的情绪分类方法,该方法将文本看作一个文本序列,通过挖掘文本的类序列规则以及规则特性实现文本情绪分类。ABDUL-MAG-EED 等^[7]构建了大规模的细粒度英文情绪数据集,并且使用

到稿日期:2019-02-20 返修日期:2019-06-18 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(U1636111)

This work was supported by the National Natural Science Foundation of China(U1636111).

通信作者:柴玉梅(ieymchai@zzu.edu.cn)

深度学习模型识别 24 种细粒度的情绪类型。对于类似“哈哈,小明今天受到惩罚了。”这种因主体不同而有情绪转折的文本,现有基于规则和机器学习的情绪分析方法通常无法正确将其归类。对“小明”而言,表达的情绪应为伤心、难过;很容易因“惩罚”一词将此句情绪归为难过;但文本的主体应为作者本人,其带有强烈的开心、兴奋的情绪。基于认知模型的文本情绪分析方法可识别文本中描述的事件与行为的相关主体,通过认知模型情绪生成规则中的一系列诱发条件确定最终的情绪类别。基于 OCC 模型进行情绪分析的已有方法通常使用词典、知识库来识别情绪诱发条件,通过定义相应的情绪分类规则识别情绪,但由于网络言论具有随意性,常常存在句子信息缺失或句中关系复杂的情况,这样会导致因情绪诱发条件缺失或无法有效、合理地识别而无法准确辨别文本情绪的问题。

为了解决上述问题,本文基于 OCC 模型提出了一种新的句子级文本情绪分类方法。该方法选取一组情绪评估变量作为节点,并结合句中所含有的表情特征,使用 MCMC(Markov Chain Monte Carlo)^[8] 算法、期望最大化(Expectation Maximization, EM)^[9] 算法分别进行结构学习、参数学习,构建情绪分类贝叶斯网络。对提取的情绪评估变量提出了一种基于语义关系的赋值方法,通过对文本进行语义分析,梳理句中包含的事件关系来确定事件优先级,并依此对各部分的情绪表达赋予不同的权重。通过贝叶斯网络进行概率推理,可以解决在句子级文本情绪分类中情绪评估变量不完整的问题,同时得到每种情绪发生的概率,不同于一般基于认知模型的情绪分类方法。传统的情绪分析方法一般只能识别一种情绪,但人们通过文本表述的情绪有时不是单一的。贝叶斯网络基于概率推理的这种特性正适应于文本情绪分析,依据所得概率可识别人们想要表达的多种情绪类别。

2 相关工作

文本情绪分析是自然语言处理领域在近年的一个热门研究方向,句子级文本的情绪分类任务被众多研究者所关注^[10-12]。Zhao^[13]通过构建大规模的情绪词典,制定情感短语模板及其计算规则,以识别句子的细粒度情感类别。Wang 等^[14]提出了一种基于标签排序的情绪分类方法,其通过计算句子各情感标签的排序值来实现分类,在 NLPCC 2104 微博文本情绪分析评测中取得了最优成绩。Santos 等^[15]提出了一种新的卷积神经网络情绪分类方法,该方法使用字符、词、句子级的特征对短文本进行情绪分析,在两个不同领域的语料集中均取得了不错的效果。Wang 等^[16]使用 skip-gram 语言模型来学习分布式词表达作为卷积神经网络的输入特征,在没有任何人工标注特征的环境中解决多标签情绪分类问题。不同于上述情绪分类方法,本文选用基于情绪认知模型和贝叶斯网络相结合的方法来实现句子级文本的情绪分类。

认知评估理论是认知心理学研究中应用广泛且较为成熟的情绪分析理论^[17]。1988 年,Ortony 等^[4]提出了一种情绪认知结构模型(OCC 模型)。Steunebrink 等^[18]针对 OCC 模型中一些冗余、有歧义的地方进行改进,建立了新的模型图和规则。Udochukwu 等^[19]基于 OCC 情感认知模型,去除冗余的规则,识别与事件和智能体有关的情绪,制定了 16 种情绪推理规则,最终合并为 5 类情绪,实现了英文文本的情绪分

析。Pang^[20]分析了传统情绪分析方法存在的不足,构建了一种基于 OCC 情绪认知模型的文本情绪分析方法,同时考虑了性格对情绪的影响。Huangfu 等^[21]采用统计方法自动构建情感维度词典,结合 OCC 模型中的情绪类型与情绪维度值的对应关系来实现情绪分类。上述方法大都是对 OCC 模型中的情绪识别规则做简单改进,并将其直接应用于情绪分析任务。本文则从 OCC 模型的情绪产生规则中提取相应的情绪评估变量用于构建情绪分类贝叶斯网络,实现句子级文本的情绪分类任务。Gao 等^[22]在 OCC 模型的基础上,通过构建 22 种情绪类型的词典以及制定细粒度的情感规则,来实现基于贝叶斯概率的情绪因果构成比计算。Werhli^[23]直接以 OCC 模型图来构建贝叶斯网络的拓扑结构,以基于 BDI 体系结构的认知环境模型为基础,选择标准数据库中多代理模型的示例,对每个代理创建网络,给定先验概率后进行模拟实验。该方法测试了情绪在 3 个假设情况下的引入,并比较了这些情绪影响代理行为的方式。本文使用贝叶斯网络,通过结构学习、参数学习刻画情绪评估变量、文本表情特征、情绪类别之间的依赖关系,通过概率推理实现文本情绪类别的分类。

3 基础理论

3.1 情绪认知评估理论

OCC 模型将情绪看作对事件的结果、智能体的行为和对象属性的认知评价。该模型通过一系列情绪诱发条件定义了 22 种情绪产生规则,如文献^[18]中的表 1 所列。从表中的规则可以看出,OCC 模型将 22 种情绪分为 4 类,分别为基于事件结果的情绪、基于智能体行为的情绪、基于事物属性的情绪以及 4 种事件与智能体的混合情绪。如图 1 所示,“不被公司需要会被解雇”是一件令人不愉快的事情,“担心”说明事情如果发生会不开心,“将”“被解雇”说明事情结果与自身相关。依照 OCC 模型中情绪 Fear 的生成规则(规则 1),将识别出这句话带有害怕、担忧的情感。

规则 1 IF (displeased about) the prospect of an undesirable event THEN Fear.



图 1 情绪生成规则示例图

Fig. 1 Generation of emotion types example

3.2 相关概念

定义 1(文本结构) 文本结构定义为 $S = \{w_1, w_2, \dots, w_N\}$, w_i 表示一条文本中的第 i 个词语对象。词语对象 $w_i = \langle cont, pos, ne, arg, rel, par, semRel, semPar \rangle$, 其中, $cont$ 表示词语内容, pos 代表词性, ne 为命名实体信息, $arg = \langle type, beg, end \rangle$ 表示序号从 beg 至 end 构成的词组的语义角色类别为 $type$, rel 表示依存句法关系, par 为关系指向的父节点词序号, $semRel$ 表示语义依存关系, $semPar$ 为关系指向的父节点词序号。

定义 2(事件关系集) 根据哈尔滨工业大学语言云平台中的语义依存关系,将事件关系集 $eveRel$ 定义为如表 1 所列的 19 种事件关系的集合,即 $eveRel = \{eCoo, eSelt, eEqu,$

ePrec, eRect, eSucc, eProg, eAdvt, eResu, eSum, ePref, eInf, eConc, eSupp, eCond, eCau, eMetd, ePurp, eAban}。

表 1 LTP 语义依存关系的事件关系标注集

Table 1 Tag set of event relations		
类型	标签	例子
并列关系	eCoo	和
选择关系	eSelt	是...还是...
等同关系	eEqu	他们→三个人
先行关系	ePrec	首先、先
顺承关系	eSucc	随后、然后
递进关系	eProg	况且、并且
转折关系	eAdvt	却、然而
原因关系	eCau	因为、既然
结果关系	eResu	因此、以致
推论关系	eInf	才、则
条件关系	eCond	只要、除非
假设关系	eSupp	如果、要是
让步关系	eConc	纵使、哪怕
手段关系	eMetd	
目的关系	ePurp	为了、以便
割舍关系	eAban	与其、也不
选取关系	ePref	不如、宁愿
总括关系	eSum	总而言之
分叙关系	eRect	例如、比方说

定义 3(谓词集) 给予一条文本 d , 其结构为 S 。若 $w_i.rel = HED$, 则核心谓词为 $w_i.cont$, 并将其存入谓词集 V 。对于所有词语对象 W , 如果 $w.semRel \in eveRel$ 且 $w.semRel \in$

V , 则 $V \leftarrow w.cont$ 。最终得到的 V 即为谓词集。

定义 4(事件优先级) 给定谓词集 V , 得到三元组 $\langle w_i, cont, w_i.semRel, w_{w_i.semPar}.cont \rangle$, 其表示 $w_i.cont$ 与 $w_{w_i.semPar}.cont$ 之间的事件关系为 $w_i.semRel$ 。根据事件关系的不同倾向性, 计算得到谓词对应的权值 $\mu_{w_i.cont}, \mu_{w_{w_i.semPar}.cont}$ 越大则事件的优先级越高。

定义 5(情绪评估变量) 从情绪认知评估 OCC 模型的情绪产生规则中提取一组用以描述事件状态及主体对不同情境的的反应的评价标准, 将其定义为情绪评估变量。

定义 6(情绪分类贝叶斯网) 情绪分类贝叶斯网 $ECBN = \langle G, \Theta \rangle$, 其中 $G = \langle X, R \rangle$ 表示一个有向无环图, X 为情绪评估变量 EV 、特征节点 E 和情绪类别 ET 的集合, R 表示结点之间的依赖关系; Θ 表示所有条件概率分布的参数集合。 $p(x_i = k_r | \pi(x_i) = j)$ 表示结点父集 $\pi(x_i)$ 属性值组合为 j 时, 结点 x_i 取第 r 个属性值的条件概率, 联合条件概率为 $p(x_1, x_2, \dots, x_N) = \prod_i^N p(x_i | \pi(x_i))$ 。

4 基于 OCC 模型的情绪评估变量提取与赋值

4.1 情绪评估变量提取

图 2 为 OCC 模型的情绪结构图^[18], 它将 22 种情绪分为 4 类, 并描述了每种情绪的诱发条件。

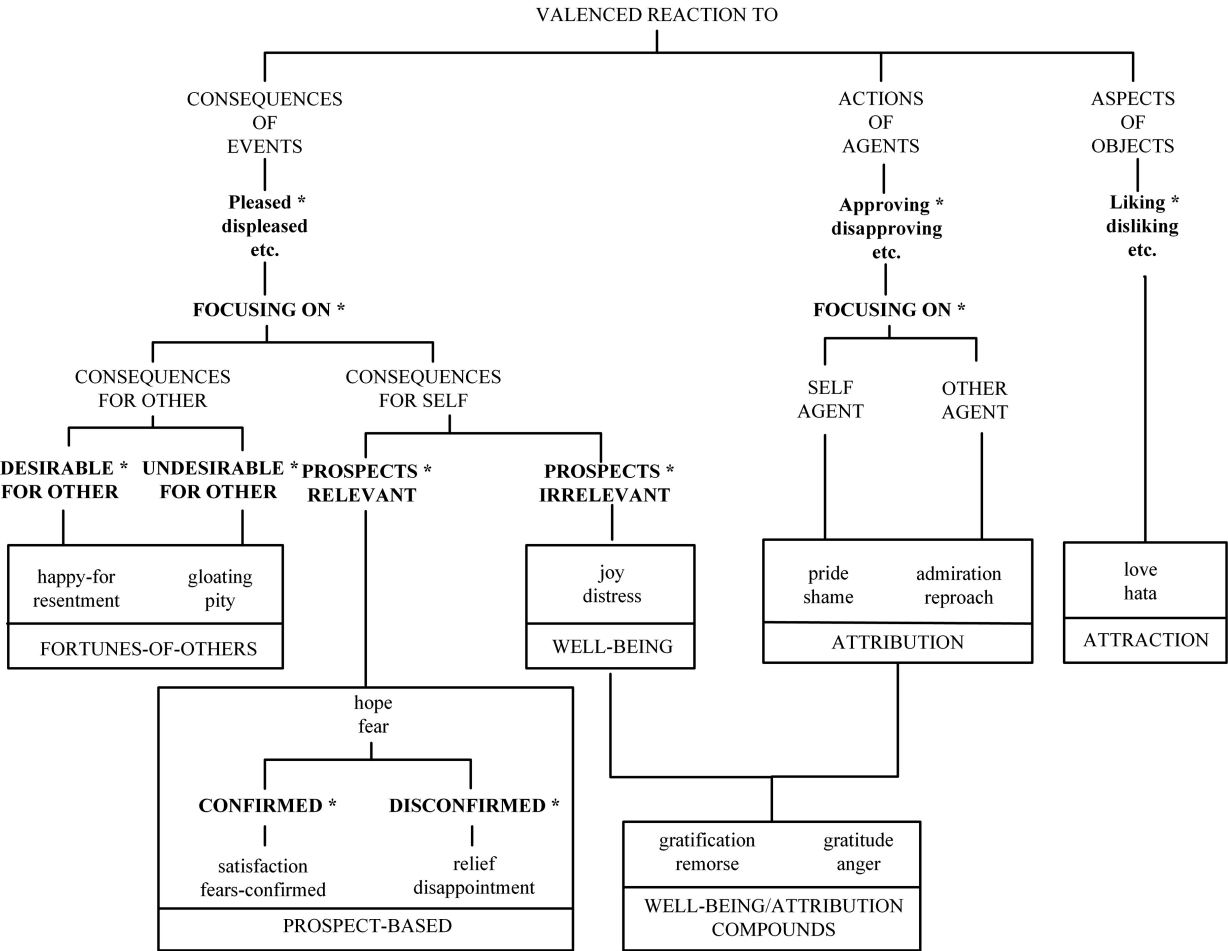


图 2 OCC 模型结构图
Fig. 2 Emotional rules of OCC model

本文方法为句子级情绪分类方法, 而现有网络文本具有随性、简洁的特点, 往往无法精确地将此类情绪句划分为 22

种情绪中的一种。因此, 在情绪句情绪类别的选择上, 本文去除 OCC 模型的 4 种混合情绪, 将余下的 18 种情绪归为 6 种

基础情绪,分别为 like, happiness, fear, anger, disgust, sadness,情绪对应如表 2 所列,并在此基础上添加常见情绪 surprise,共计 7 类。

表 2 情绪类别对应表

情绪类别	对应情绪
like	love, admiration, pride
happiness	joy, happy-for, hope, satisfaction
fear	fear, fear-confirmation
disgust	hate, gloating
sadness	pity, disappointment, distress, shame
anger	resentment, reproach

OCC 模型从心理学角度出发定义了一系列的情绪诱发条件,如图 2 中加 * 处所示。liking 和 disliking 表示作者对对象属性的态度;pleased 和 displeased、approving 和 disapproving 分别表示作者对事件结果、智能体行为的态度;focusing on 表示事件或行为的相关主体是自己还是其他人;desirable, undesirable 描述事件与他人相关时事件的属性;prospects relevant 指事件未来发展的影响性,表示时间状态为将来;confirmed, disconfirmed 指事件是否确定。不同的情绪诱发条件组合对应不同的情绪类别,如“明天要下雨,可能不能去春游了。”表示作者不满于一件令自己期待的事情可能不会发生而产生情绪 disappointment。

基于上述分析,提取构建 ECBN 所用的情绪评估变量。1)作者对事物属性的态度(*au_obj*):like, dislike;2)作者对事件的态度(*au_eve*):pleasure, displeasure;3)事件的相关主体(*rel*):self, other;4)事件的状态(*time*):confirmed, unconfirmed;5)事件属性(*rel_eve*):desirable, undesirable。为使贝叶斯网络结构能够精简且有效地表达情绪评估变量与情绪类别之间的依赖关系,确定网络节点即情绪评估变量时应尽量减少冗余。由于本文选择了 7 种情绪对文本进行分类,简化了情绪类别,因此在分析情绪的诱发条件时不再细分,将事件和智能体的行为统称为事件。

网络文本的随意性较强,含有较多“一流的!”“哈哈…”这种由词或词组构成的独立句子,简称其为独词句。对于独词句而言,通常直接表达情感,不包含对对象、事件的描述,因此添加一个变量 *words* 用于区分是否为独词句,以更好地提取句子的情绪评估信息。对于添加的情绪类别 surprise, OCC 模型中没有明确对应的情绪诱发条件,对此本文添加情绪评估变量 *sur_mood* 用以表示文本中关于情绪 surprise 的描述信息。

至此,本文共提取了 7 个情绪评估变量,如表 3 所列。

表 3 情绪评估变量表

情绪评估变量	属性值
<i>words</i>	是否为独词句
<i>rel</i>	self(1), other(2)
<i>au_obj</i>	like(1), dislike(2), none(3)
<i>au_eve</i>	pleasure(1), displeasure(2)
<i>sta_eve</i>	desirable(1), undesirable(2)
<i>time</i>	confirmed(1), unconfirmed(2)
<i>sur_mood</i>	low(1), high(2), none(3)

根据情绪评估变量与情绪类别间的依赖关系,可构建贝叶斯网络以实现情绪分类。用一个简单的实例来说明,上述例句“明天要下雨,可能不能去春游了。”对应图 3 中的分析,“春游”对应图中变量 *sta_eve* 的值为 desirable,而“不能去春游”对应变量 *au_eve* 的值为 displeasure,事件的相关主语为 self,事件的时态为将来,此时句子表达伤心的情绪,即上述变量与 sadness 之间具有一定的依赖关系,可通过已知信息 *T* 计算情绪 sadness 的条件概率 $P(\text{sadness}|T)$ 来判别句子是否表达该情绪。



图 3 贝叶斯网络结构示例图

Fig. 3 Bayesian network structure example diagram

4.2 情绪评估变量赋值

4.2.1 选用情感词典与构建常识知识库

目前常用的中文情感词典有 4 种,分别为 HowNet 情感极性词典、大连理工大学情感词汇本体库、清华大学极性词典、台湾大学情感词典。归纳整理得到基础情感极性词典 pos 和 neg。从大连理工大学情感词汇本体库中抽取类别为“惊”的词语构建情绪词典 surprise,用以辅助识别情绪 surprise。

HowNet 知识库^[24]包含 Entity, Attribute, Event 等 11 种概念类,本文提取 Entity 类作为基础 object 词典。参考文献[25]中常识知识库的构建方法,通过分析 HowNet 知识库的概念结构,完善对应常识词典,使用哈工大构建的同义词词林^[26]构建其他常识词典。词典描述如表 4 所列。

表 4 常识词典的具体描述

Table 4 Description of common sense dictionaries

类型	描述	常例
object	客观事物	肖邦、木块、月季等
other	其他人称	你、她、公众等
ptime	已确定	早已、曾经、昔日等
ftime	未发生	今后、来年、可能等

4.2.2 面向情绪评估变量的句法语义分析

为了挖掘文本中的事件关系、语义信息等内容,本文首先使用哈尔滨工业大学的语言云平台(LTP)对文本进行预处理,再从情绪评估变量的角度分析文本的语义关系。

(1)事件的相关主体识别

通过语义角色标注,可识别句中谓词的论元,如施事、受事、时间等。A0 通常表示动作的施事;A1 通常表示动作的影响;A2—A5 依照谓词分别对应不同的含义,此处不做过多分析。若谓词为形容词,A0 通常为形容词描述的对象,如“你今天真漂亮”,“漂亮”为对象“你”的评价词;若谓词为动词,则 A0 为动作的主体,A1 为动作影响的对象或结果,如“我被小明教会了游泳”,发出动作“教”的对象为“小明”,动作影响的对象为“我”。由此,总结得出情绪评估变量 *rel* 的赋值规则。

规则 2 If $match(w, other) \parallel (w.pos == 'nh')$ Then $rel=2$ Else $rel=1$

如果句中谓语含有语义角色 A0,则 *w* 表示 A0 指代的

词,否则 w 为该句主语。

(2) 计算事件优先级

对于简单句中情绪评估变量的赋值,可通过依存句法分析句子的结构,理清对象、主体以及事件的关系。然而,对于句子结构复杂的文本,仅依靠依存句法分析的标注关系无法识别子句中的事件逻辑关系。本文借助语义依存关系中的递进、转折、假设等 19 类事件关系(见表 1),分析事件间的情绪诱发关系,确定事件的优先级。

根据子句间的关系,复句^[27]可分为联合复句和偏正复句两类。从语义角度分析,联合复句通常表示子句间的并列、选择、递进等关系,各子句间不分主次;偏正复句一般通过关联词表达子句间的因果、假设、转折等关系,子句间有主从之分,通常正句是全句的语义中心。按照上述分析,将表 1 中所列的事件关系分为 3 类:即事件关系不分主次,事件关系指向节点为中心事件,事件关系的父节点为中心事件。

1) 子句间的事件关系不分主次

当事件关系属于 $equSem = \{eCoo, eSelt, eEqu, ePrec, eRect, eSucc, eProg\}$ 时,关系所对应的两个事件是平等关系。如 $eCoo$ 表示并列,常见形式为“事件 A 和事件 B”,如“吃饭和跳舞”中“吃饭”和“跳舞”没有主次之分;也可标注对象之间的关系,如“张森和我一起去吃饭”中“张森”和“我”被标注为 $eCoo$ 。为了分析复句中的中心事件,本文在这里主要考虑事件之间的关系,忽略包含的对象关系,后文不再重复声明。

2) 事件关系指向节点为中心事件

当事件关系属于 $childSem = \{eAdvt, eResu, eSum, ePref, eInf\}$ 时,关系指向的事件为文本表达的主要意思。例如, $eResu$ 表示结果关系,常见形式为“因为(由于)…,所以(于是)…”,如“因为今天预报有雨,所以妈妈才不让我出去打篮球。”标注的事件关系为“有 $\rightarrow$ 让($eResu$)”,则这句话表达的中心是妈妈不让我出去打篮球,“预报有雨”是在描述结果发生的原因。

3) 事件关系的父节点为中心事件

当事件关系属于 $parSem = \{eConc, eSupp, eCond, eCau, eMetd, ePurp\}$ 时,关系的父节点所描述的事件为文本表达的主要意思。例如, $eSupp$ 表示假设关系,常见形式为“如果(要是、假如)…就…”,如“如果你感觉好点了,那就不用输液了。”LTP 标注结果为“感觉 $\leftarrow$ 输液($eSupp$)”。前者是一种假设条件,正句则表示在这种假设成立的条件下产生的结果为“不用输液”。选择事件关系的父节点“输液”所在的子句作为中心。

根据上述分析,现使用式(1)计算谓词权重,以此表示事件优先级 $\langle V, \mu_V \rangle$ 。

$$\mu_{v_i} = num(v_i, semRel \in \{equSem, parSem\}) + \alpha[num(par(v_i), semRel \in parSem) + num(v_i, semRel \in childSem)] + num(par(v_i), semRel \in \{childSem, equSem\}) \quad (1)$$

其中, α 表示对中心事件的加权,本文设为 2; $num(T)$ 指满足条件 T 的数目; $par(v_i)$ 表示 v_i 的父节点。

4.2.3 情绪评估变量赋值算法

分析句子级文本并对情绪评估变量赋值时,通常要考虑文本是否为独词句。对于没有完整主谓宾关系的独词句,默

认选择相关主体为自己,对象属性的描述为空,而对其他情绪评估变量则通过逐词匹配相应词典进行赋值;对于简单句和复句,则根据事件优先级对每个子事件的分析结果进行加权处理,从而得到最终的变量属性值。情绪评估变量赋值的具体过程如算法 1 所示。

算法 1 情绪评估变量赋值算法

输入:情感词典、常识词典、结构为 S 的文本 D

输出:情绪评估变量

```

1. if D 为独词句
2.   words = rel = 1
3.   au_obj = 3
4.   for  $w_i$  in S
5.     通过匹配词典 pos, neg, ptime, ftime, surprise 为变量 au_eve, sta_eve, time, sur_mood 赋值
6.   end
7. else
8.   words = 2
9.   按规则 2 对事件相关主体 rel 赋值
10.  通过匹配词典 surprise 为变量 sur_mood 赋值
11.  对满足条件  $w_i, semRel \in \{mTime, dTime, Time, mMod\}$  的词  $w_i$ , 匹配词典 ptime 和 ftime, 并计数为 P 和 F
12.    time = T(F - P)
13.    根据式(1)计算得到  $\langle V, \mu_V \rangle$ 
14.    if 主语  $\in$  object
15.      for  $v_i$  in V:
16.        计算情感值 obpos $_{v_i}$ , obneg $_{v_i}$ 
17.      end
18.      au_obj = T( $\sum_V \mu_{v_i} (obpos_{v_i} - obneg_{v_i})$ )
19.    end
20.    for  $v_i$  in V:
21.      按规则 2 对子事件  $v_i$  的相关主体 rel $_{v_i}$  赋值
22.      if rel $_{v_i} = 2$ 
23.        计算情感值 othpos $_{v_i}$ , othneg $_{v_i}$ 
24.      else
25.        计算情感值 selpos $_{v_i}$ , selneg $_{v_i}$ 
26.      end
27.    end
28.    sta_eve = T( $\sum_V \mu_{v_i} (othpos_{v_i} - othneg_{v_i})$ )
29.    au_eve = T( $\sum_V \mu_{v_i} (selpos_{v_i} - selneg_{v_i})$ )
30. end
31. 输出情绪评估变量 EV

```

对于函数 $T(value)$, 若 $value$ 为正则输出为 1, 若 $value$ 为负则输出为 2, 若 $value$ 为 0 则输出为 3 (若 sta_eve 和 au_eve 为 0 则输出为 null)。

5 基于 OCC 模型和贝叶斯网络的情绪句分类方法

上节已从 OCC 模型中提取了 7 个情绪评估变量作为 ECBN 的情绪评估节点。由于网络文本的句子结构较为随意,表情符号较多,本节将提取基于表情符号的文本特征来完善网络节点,并基于 MCMC 算法逐级学习节点间的依赖关系,依照所提融合策略构建完整的网络拓扑结构。为减小文本信息缺失所带来的影响,本文使用 EM 算法对 ECBN 进行

参数学习,得到完整的 ECBN 实现句子级文本的情绪分类。

5.1 文本特征变量的提取与赋值

为了提高文本情绪分类的精度,依据网络文本多包含表情且随意性较强的特点,本文从文本中提取表情特征 E 作为 ECBN 的特征节点。文本特征变量赋值算法的具体实现如算法 2 所示。

算法 2 文本特征变量赋值算法

输入:文本 D ,权重 φ

输出: $\exp_var, E_j(j=1,2,\dots,7)$

```

1. ejc=True
2. if 文本  $D$  不包含表情符号
3.    $\exp\_var=E_j=1(j=1,2,\dots,7)$ 
4. else
5.    $\exp\_var=1$ 
6.   for 每个表情  $emo_k$  in  $D$ :
7.     if  $\sum_{j=1}^7 Nemo(n_{kj} \neq 0) \leq 2$ 
8.       ejc=False
9.        $E_j(n_{kj} \neq 0)=2, E_j(n_{kj}=0)=1$ 
10.    end
11.  end
12.  if ejc
13.    对  $\sum_i Nemo(n_{ki} \neq 0) + n_{ki} \times \varphi$  按倒序排序,取前两位对应情绪类别的表情特征  $E$  并将其赋值为 2,其他为 1
14.  end
15. end
16. 输出  $\exp\_var, E_j(j=1,2,\dots,7)$ 

```

其中, n_{kj} 表示表情 emo_k 所在的文本属于情感 j 的数目,权值 φ 设为 $1/\max(n_{kj})$, $Nemo(n_{kj} \neq 0)$ 表示符合条件 $n_{kj} \neq 0$ 的次数。算法的时间复杂度取决于句子中表情符号的个数,表示为 $O(k)$ 。

5.2 构建 ECBN

情绪分类贝叶斯网络 ECBN 由文本情绪类别与选取的情绪评估变量和特征节点组成。以情绪 surprise 为例, $\exp_var$ 表示文本是否包含表情符号,与其他表情特征节点间具有依赖关系; E_7 表示文本所含表情符号是否对应情绪 surprise,如果是则可能诱发相应情绪;当 $words$ 表示独词句时,节点 rel 和 au_obj 的属性值取默认值, sur_mood 表示文本惊讶情绪的强弱与其他基于 OCC 模型提取的情绪评估变量一并影响情绪的产生。在上述有依赖关系的节点间建立有向弧,得到情绪 surprise 的贝叶斯网络结构图。逐个分析 7 类情绪,构建完整的情绪分类贝叶斯网络结构示意图,如图 5 所示。

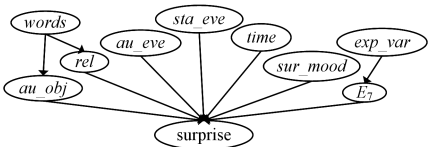


图 4 单个情绪的贝叶斯网络结构图

Fig. 4 Bayesian network structure for single emotion

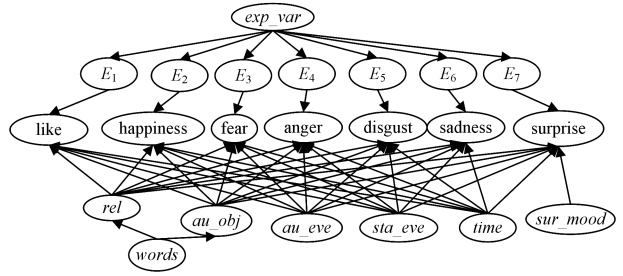


图 5 ECBN 拓扑结构图

Fig. 5 ECBN network structure

依照上述分析,学习节点间的依赖关系是合理构建 ECBN 拓扑结构的关键步骤。本文基于 MCMC 算法学习网络节点之间的依赖关系,构建 ECBN 的网络结构。Metropolis-Hastings(MH)算法^[8]是 MCMC 算法的一种形式,Giudici 等^[28]将其应用于贝叶斯网络结构的学习。为了避免训练中情绪类别之间出现过强的依赖关系,本文选择逐类学习每种情绪的贝叶斯网络结构。用 G_{emo} 表示单类情绪的网络结构(emo 表示情绪类别), Y 为完整的样本集, G^{i+1} 由 G^i 在不产生有向环的条件下进行一次边操作(增加、删除、改变方向)所得。使用 MH 算法构造一个马尔可夫链,其平稳分布为 $P(G|Y)$,则从 G^i 到 G^{i+1} 的接受概率为:

$$A(G^i, G^{i+1}) = \min\{1, \frac{P(G^{i+1}|Y)}{P(G^i|Y)}\} \quad (2)$$

其中:

$$\frac{P(G^{i+1}|Y)}{P(G^i|Y)} = \frac{P(G^{i+1})P(Y|G^{i+1})}{P(G^i)P(Y|G^i)} = \frac{P(Y|G^{i+1})}{P(Y|G^i)} \quad (3)$$

迭代数次后,根据所得的一组贝叶斯网络结构,按式(4)取 G 为最终的网络结构。

$$G = \arg \max_{G^i} \sum(G^i), i=1,2,3,\dots,n \quad (4)$$

为增强情绪评估节点与情绪类别之间的依赖关系,并削弱评估节点之间的关联,本文针对上述学习过程得到的 7 种情绪的网络结构 G_{emo} 提出了一种融合策略,即选择与情绪类别节点直接相连的有向边, $distance(*, ET)$ 为 1;统计不含情绪类别节点间的依赖强度(有向边频数),若其小于均值则删除;最终根据所得有向边构建完整的情绪分类贝叶斯网络结构 ECBN。基于 MH 算法和融合策略构建 ECBN 网络结构的具体描述过程如算法 3 所示。

算法 3 ECBN 结构学习算法

输入:依情绪类别分割后的变量集 X ,完整样本集 Y ,初始值 m 和 n

输出:情绪贝叶斯网络结构 ECBN

```

1. for each emotion
2.   初始化网络结构  $G^0$ 
3.   for  $i=0$  to  $n$ 
4.     在不产生有向环的条件下,对  $G^i$  进行一次边操作,得到候选结构  $G'$ 
5.     在  $[0,1]$  区间中依照均匀分布采样得到  $u \sim U[0,1]$ 
6.      $A(G^i, G') = \frac{L(G', a)L(G', a)}{L(G^i, b)L(G^i, b)}$ 
7.     if  $u \leq A(G^i, G')$ 
8.        $G^{i+1} = G'$ 
9.     else
10.       $G^{i+1} = G^i$ 

```

```
11.      end
12. end
13. Gemo = arg maxGi sum(Gi), i = 1, 2, ..., n - m
14. end
15. select dir-edge(distance( * , ET) = 1)
16. if sum(dir-edge((E, EV), (E, EV))) ≥ 4
17.     select dir-edge
18. end
19. ECBN = (E, EV, ET, dir-edge)
20. return ECBN
```

算法 3 中, $L(G^i, a) = \prod_b \frac{\Gamma(N'_{ab})}{\Gamma(N_{ab} + N'_{ab})} \frac{\prod_{c=1}^{R_a} \Gamma(N_{abc} + N'_{abc})}{\Gamma(N'_{abc})}$ 表示 G^i 关于节点 x_a 的边际似然, 用 $\Pi_a = b$ 表示在数据集中 x_a 的父节点的第 b 个实例, c 表示节点 x_a 的属性值, N_{abc} 表示满足 $x_a = c$ 且 $\Pi_a = b$ 的数目, $N_{ab} = \sum_c N_{abc}$, N'_{abc} 表示 Dirichlet 分布的超参数(设为 1)。算法 3 中, 对图进行一次边操作的时间复杂度为 $O(N^2)$, 则算法总的时间复杂度为 $O(nN^2)$ 。

本文将 EM 算法用于 ECBN 的参数学习, 将训练样本集 H 分割为完整样本集 Y 和含有变量缺失值的样本集 D 。每个节点的条件概率初始值由完整样本集 Y 计算得出, 使用 EM 算法对含有变量缺失值的样本集 D 进行参数学习的具体过程如算法 4 所示。

算法 4 ECBN 参数学习算法

输入: ECBN 网络结构, 完整样本 Y , 有未知变量的样本集 D , 收敛阈值 δ

输出: 条件概率表 CPT

```
1. for 每个变量  $x_i$  in  $V$ :
2.   for 每个父节点  $\pi(x_i)$ :
3.     for  $j$  in  $\pi(x_i)$  属性值:
4.       for  $k_r$  in  $k$ :
5.         初始化  $t = 0$ 
6.         对于  $H$  中完整的数据集  $Y$ , 计算  $\theta_{ijk_r}^t$ , 并将其作为初始值
7.         对于缺失数据集  $D$ , 计算:
8.            $m_{ijk_r}^t = \sum_{g=1}^N P(x_i = k_r, \pi(x_i) = j | D_g, \theta_{ijk_r}^t)$ 
9.            $\theta_{ijk_r}^{t+1} = \frac{m_{ijk_r}^t}{\sum_{i=1}^{M_i} m_{ijk_r}^t}$ 
10.        while(  $|\theta_{ijk_r}^{t+1} - \theta_{ijk_r}^t| > \delta$  ) {
11.           $t = t + 1$ 
12.           $\theta_{ijk_r}^{t+1} = \arg \max_{\theta} Q(\theta_{ijk_r}, \theta_{ijk_r}^t) = \frac{m_{ijk_r}^t}{\sum_{i=1}^{M_i} m_{ijk_r}^t}$ 
13.        }
14.      将参数  $\theta_{ijk_r}^{t+1}$  放入条件概率表 CPT
15. 输出条件概率表 CPT
```

算法 4 中, t 为迭代次数; θ 为 ECBN 中节点的条件概率, 表示为 $\theta_{ijk_r} = P(x_i = k_r, \pi(x_i) = j); D_g$ 表示第 g 个样本实例, 共有 N 个; V 表示所有的变量(节点) $\{x_1, x_2, \dots, x_{x_2}\}$, k_r 表示 x_i 的第 r 个取值, 共有 M_i 个; $\pi(x_i)$ 表示 x_i 的父集。

5.3 情绪分类

5.2 节已完成 ECBN 的网络结构构建和参数学习工作, 得到了完整的情绪分类贝叶斯网络。使用 ECBN 进行分类的核心就是在给定 T , 即已知情绪评估变量和特征变量对应属性值的情况下, 计算情绪类别 emo_k 的条件概率 $P(emo_k | T)$ 。联合树算法^[29]是一种被广泛应用的贝叶斯网络精确推理算法。应用联合树算法计算每类情绪的条件概率, 根据所得概率和规则 3 对文本表达的情绪进行归类。

规则 3 If $P(emo_k | T) > \lambda$ Then $T_{emo} = emo_k$ 。其中, $k = 1, 2, \dots, 7$ 分别表示 7 种情绪类别; λ 代表情绪的产生阈值。

6 实验与分析

6.1 数据集与评价指标

为了保证实验结果的可靠性, 本文选用 NLPCC 2014 中文微博情绪分析评测数据集作为实验数据, 该数据集包含 14 000 条训练数据和 40 000 条测试数据。除去个别无法使用 LTP 在线语义分析的情绪句, 训练数据含有 15 690 条情绪句, 测试数据含有 3 822 条情绪句, 每句话表达了 1~2 两类情绪, 每类情绪的分布情况如表 5 所列。

Table 5 Experimental data sentiment distribution							
	like	happiness	anger	sadness	disgust	surprise	fear
训练数据	4 632	3 063	2 154	2 628	3 627	906	336
测试数据	1 705	707	289	326	735	282	71

本文选用与评测任务相同的评价标准¹⁾, 以多标签分类的平均精度为评价指标:

$$Average_{Precision} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y_i|} \sum_{k=1}^{|Y_i|} \frac{|(emo_k \in Y_i | rank(x_i, emo_k) \leq rank(x_i, y))|}{rank(x_i, emo_k)}$$

(5)

其中, $|Y_i|$ 表示情绪句 x_i 的情绪个数, $rank(x_i, y)$ 表示情绪句 x_i 中情感 y 的位置。评测结果报告中根据计分规则的不同, 又将平均精度分为严格标准和宽松标准²⁾。

6.2 结构学习

使用 LTP 对评测训练数据集中包含的情绪句进行句法语义分析, 再通过算法 1 和算法 2 得到变量赋值后的结果, 即为 ECBN 的训练数据。15 690 条情绪句共有 13 401 条完整数据, 将其按情绪类别分割后输入算法 3 逐次进行单类情绪的贝叶斯网络结构学习, 然后经步骤 15、步骤 16 得到完整的 ECBN, 如图 6 所示, 其中参数 m 和 n 分别设为 110 和 2 200。选取实验中间结果即个别情绪的网络结构简要阐述算法中的边操作环节。图 7、图 8 分别为情绪 happiness 和 disgust 网络结构的实验结果图(图中只绘制了与情绪相关的节点)。按算法 3 中的步骤 15 选择 $distance(*, ET) = 1$ 的有向边, 按步骤 16 对图中的其他有向边进行删减, 其中 $dir-edge\langle au_eve, sta_eve \rangle$ 在 7 类情绪网络结构图中出现的频次小于均值, 因此

1) <http://tcci.ccf.org.cn/conference/2014/dldoc/evatask1.pdf>
2) <http://tcci.ccf.org.cn/conference/2014/index.html>

删除该边而保留其他满足条件的有向边。图 6 中基于 OCC 模型提取的情绪评估变量与情绪类别之间的依赖关系,与上文分析得到的网络结构图 5 大致相同;不同之处在于,图 5 中的文本表情特征只与对应的情绪类别相关,而实验表明表情特征 $sadness(E_6)$ 与情绪 happiness 之间、表情特征 $happiness(E_2)$ 与情绪 disgust 之间有较强依赖关系,即文本表情特征不只与对应的情绪类别相关,与其对应的相反情绪也有较强的相关性。

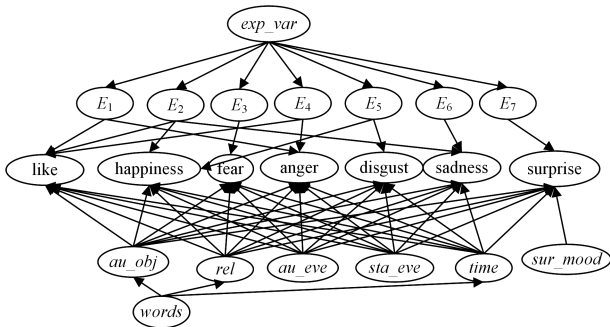


图 6 结构学习实验结果图(ECBN)

Fig. 6 Structure learning experimental results figure (ECBN)

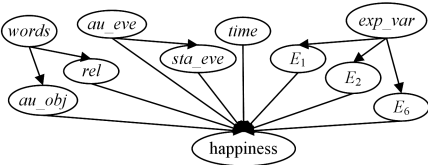


图 7 情绪 happiness 的贝叶斯网络结构图

Fig. 7 Bayesian network structure of emotional happiness

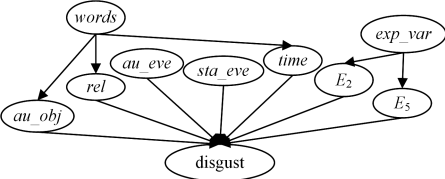


图 8 情绪 disgust 的贝叶斯网络结构图

Fig. 8 Bayesian network structure of emotional disgust

为了验证所得网络结构的合理性,本文选用图 5 所示的结构、图 6 所示的结构以及图 6 中去除有向边 $E_2 \rightarrow like$ 和 $E_4 \rightarrow like$ 的网络结构,分别构建情绪分类贝叶斯网络 ECBN5,ECBN6 以及 ECBN6',参数学习算法均选用算法 4,使用联合树算法进行概率推理,首要情绪取概率最大的情绪类别,次要情绪的判别阈值初步设为 0.6,实验结果如图 9 所示。

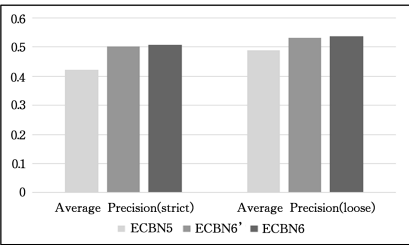


图 9 不同网络结构的分类结果对比

Fig. 9 Comparison of classification results of different network structures

由图 9 可知,实验调整后的网络结构 ECBN6 较文中分析所得结构 ECBN5,在严格与宽松标准下的平均精度均有大幅提升。另外,本文使用网络结构 ECBN6'取得的实验结果,在严格与宽松标准下的平均精度均高于 ECBN5,略低于 ECBN6,说明情绪类别与其他情绪的表情特征之间的依赖关系有助于提升分类精度,即实验所得网络结构具有合理性。

6.3 参数学习

根据算法 4,用含有缺失值的训练数据对贝叶斯网络进行参数学习,阈值 δ 设为 0.001,最大迭代次数为 10。将缺失值用固定值填充,使用贝叶斯估计算法对贝叶斯网络进行参数学习。使用测试数据对两者得到的贝叶斯网络模型分别进行实验,主要情绪取概率最大的情绪类别,次要情绪判别阈值设为 0.6,结果如图 10 所示。EV_(EM) 较 EV,以及 ECBN_(EM) 较 ECBN,在宽松标准和严格标准的平均精度上均有所提高(EV 为去除文本表情特征节点的网络结构)。提升幅度较小是由于训练集中含有缺失值的数据仅有两千多条,所占比例相对较低。

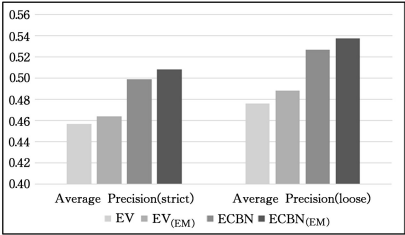


图 10 不同参数学习方法的分类结果对比图

Fig. 10 Classification results of different parameter learning methods

6.4 情绪判别阈值的选择

实验中,通过设置情绪判别阈值来得到文本的情感类别,对最终计算得到的每个情绪类别的概率进行排序,第一位为首要情绪,第二位若高于阈值则记为次要情绪,否则为无。平均精度随阈值变化的情况如图 11 所示。

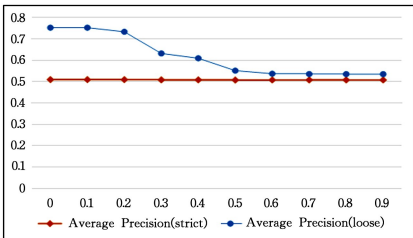


图 11 平均精度随情绪判别阈值的变化趋势

Fig. 11 Trend of average accuracy with emotional judgment threshold

从图 11 中可以看出,严格标准的平均精度基本稳定,不随阈值的变化而变化。宽松标准下的平均精度随阈值变化的浮动较大,在阈值为 0,即选取所有第二位的情绪时,平均精度最大,为 0.7527。此时,虽能匹配较多的正确情绪,但也会带来一些损失,将错误的情绪归给文本。随着阈值逐渐增大,宽松标准的平均精度逐渐变小并趋于稳定。通过分析实验结果可知,平均精度在 0.2~0.4 之间波动较大的原因为不能较好地识别表意相近的情绪,且主次区分不明显。例如,较多句子的情绪为“happiness,none”或“like,none”,而本文得到的结

果为“like, happiness”或“happiness, like”。通过上述分析,本文取宽松标准平均精度的平均值所对应的阈值 0.4 作为情绪判别阈值。

6.5 对比实验

本文共设计了两个对比实验来验证情绪分类贝叶斯网络的有效性。实验 1:情绪评估变量与文本表情特征相结合对实验结果的影响。实验 2:与 NLPCC 2014 微博情绪分析评测子任务 2 的开放测试做对比,验证本文方法的有效性。实验结果如图 12 所示。

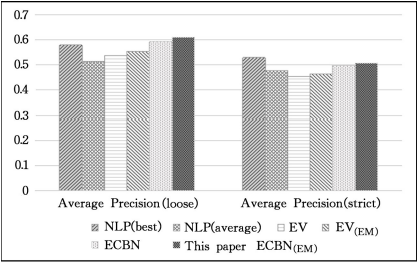


图 12 实验对比结果

Fig. 12 Experimental comparison results

EV 为仅使用从 OCC 模型中提取的情绪评估变量构建的情绪分类贝叶斯网络,其在宽松标准下的平均精度为 0.5377,高于 NLP 中位成绩,接近 NLP 最好成绩;而在严格标准下,其效果低于 NLP 中位成绩。为此,本文结合文本表情特征构建 ECBN 分类模型。ECBN 相比 EV,在宽松标准下的平均精度提高了 5.6%,在严格标准下的平均精度提高了 4.6%;ECBN_(EM) 相比 EV_(EM),在宽松标准下的平均精度提高了 5.5%,在严格标准下的平均精度提高了 4.4%。实验结果表明,准确地把握网络文本的特点,提取有效的文本特征,能提高情绪句分类的精度。

本文使用与评测封闭测试要求相同的情感词典;数据预处理的工具为哈尔滨工业大学的云平台,与评测封闭测试要求使用的数据预处理工具 Stanford parse 不同,因此只能与评测开放测试结果做对比。图 12 中显示的是 NLP 评测开放测试的最好成绩^[14]和中位成绩。对比实验结果,对宽松标准的平均精度而言,本文方法所得结果为 0.6096,高于 NLP 最好成绩,说明 ECBN 所识别的情绪能尽可能覆盖句子表达的多种情绪,具有优越性。对于严格标准,本文模型所得平均精度为 0.5085,高于 NLP 中位成绩,低于 NLP 最好成绩,这是由于 ECBN 对相近情绪的区分度较低,这与上文情绪判别阈值选择的分析一致。综上分析,本文方法在情绪句的情绪类别分类中具有一定的优越性,但对句子情绪的主次之分以及相近情绪的区分还有待提高。

结束语 本文分析了文本情绪分类的研究现状以及 OCC 模型在情绪分析中的应用,提出了一种基于 OCC 模型和贝叶斯网络的句子级文本情绪分类方法。从 OCC 情绪认知模型中提取合理的情绪评估变量,结合文本表情特征使用结构学习、参数学习算法学习情绪分类贝叶斯网络 ECBN,实现文本情绪分类。实验结果表明该方法具有有效性。

OCC 模型在文本情绪分析上的研究还存在很多问题需要进一步探讨。从文中的实验结果可以看出,本文在主要情

绪和次要情绪的分辨中尚有不足。后续将进一步改进情绪评估变量的赋值算法,改善 ECBN 的网络结构,提高情绪分类结果的精度,并考虑在段落级文本的情绪识别与分类中开展进一步的工作。

参 考 文 献

[1] LI R, LIN ZH, LIN H L, et al. Text Emotion Analysis: A Survey. Journal of Computer Research and Development[J], Journal of Computer Research and Development, 2018, 55(1): 30-52.

[2] SHI L. A Summary of Some Issues in Emotion Research[J]. Advances in Psychological Science, 2000, 8(1): 63-68.

[3] ZHAO Y Y, QIN B, LIU T. Text sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.

[4] ORTONY A, CLORE G L, COLLINS A. The Cognitive Structure of Emotions: Factors Affecting the Intensity of Emotions [J]. Contemporary Sociology, 1988, 18(6): 2147- 2153.

[5] LAZARUS R S. Emotion and adaptation [J]. Journal of Nervous & Mental Disease, 1991, 181(3): 17-23.

[6] WEN S, WAN X. Emotion classification in microblog texts using class sequential rules[C]// Proc of the 28th AAAI Conf. on Artificial Intelligence. Quebec: AAAI Press, 2014: 187- 193.

[7] ABDUL-MAGEED M, UNGAR L. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks[C] // Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL Press, 2017: 718-728.

[8] HASTINGS W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970, 57(1): 97-109.

[9] LAURITZEN S L. The EM algorithm for graphical association models with missing data[J]. Computational Statistic & Data Analysis, 1995, 19(2): 191-201.

[10] DAS D, BANDYOPADHYAY S. Sentence-Level Emotion and Valence Tagging[J]. Cognitive Computation, 2012, 4(4): 420-435.

[11] WANG L, REN F, MIAO D. Multi-label emotion recognition of weblog sentence based on Bayesian networks[J]. IEEJ Transactions on Electrical & Electronic Engineering, 2016, 11(2): 178-184.

(更多参考文献见文章首页二维码)



XU Yuan-yin, born in 1993, master. Her main research interests include natural language processing and so on.



CHAI Yu-mei, born in 1964, master, professor. Her main research interests include machine learning, data mining and natural language processing.