# 基于区间直觉模糊的情感分类模型①

赵 纯, 高俊波

(上海海事大学 信息工程学院, 上海 201306)

**摘 要**: 随着电子商务, 个人博客, 社交网站和微博的蓬勃发展, 互联网进入了一个崭新的时代, 而在线评论的情感分类关系到个人决策、企业管理甚至社会安全. 提出了一种基于区间直觉模糊的情感分类模型, 采用了区间直觉模糊算子来计算特征词的区间直觉模糊数, 利用区间直觉模糊集的隶属度、非隶属度和犹豫度分别定量地描述特征词, 通过情感合成确定文本的情感倾向, 从而获得准确率较高的情感倾向性分析结果. 最后通过相同语料库的比较实验证明该分类模型的可行性、正确性和较高的分类性能.

**关键词**: 区间直觉模糊; 合成; 情感分类; 情感倾向; 在线评论

## Sentiment Classification Model Based on Interval-Valued Intuitionistic Fuzzy Sets Model

ZHAO Chun, GAO Jun-Bo

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

**Abstract**: With the development of e-commerce, blog, social networking sites and micro-blog become much more flourishing. The Internet has entered a new era, and the sentiment classification of online comments related to individual decisions, business management and also social security. A model of sentiment classification based on interval-valued intuitionistic fuzzy sets and the interval-valued intuitionistic fuzzy operator was proposed to calculate the interval-valued intuitionistic fuzzy numbers of feature words. Meanwhile, by using membership, non-membership and hesitation to quantitatively describe the feature words. It can get the sentiment tendency through the sentiment synthesis, in order to take more accurate analysis results of sentiment tredency. Finally, throughing the comparative experiment based on same corpus, it proves that this model has high feasibility, correctness and classification performance.

**Key words**: interval-valued intuitionistic fuzzy sets(IVIFS); synthetize; sentiment classification; sentiment tendency; online comments.

随着因特网的迅猛发展, 互联信息量正在快速地增长, 互联网的普及以及多种新型网络媒体的出现不仅给人们带来了海量的信息, 同时也给人们提供了各种表达自己情感的舞台, 例如微博、博客、评论等在线评论平台. 这些反馈信息中隐藏着用户对某些事件或者产品的情感倾向, 挖掘反馈信息中隐藏的情感倾向能更好地了解用户的消费习惯、分析热点事件的舆情, 目前这方面的研究[1]已成为国内外众多学者关注的热点.

情感分类(又称情感倾向性分析、情感分析)就是指通过挖掘和分析文本中的立场、观点、看法、情绪等主观信息, 并对文本的情感倾向做出类别判断. 它可以广泛的应用于社会舆情分析、产品在线跟踪与质量评价、影视评价等方面. 目前文本的情感分类方法可以分为: 无监督学习、半监督学习和监督学习 3 类. 由于很难找到适用于无监督学习的情感词典, 而半监督学习方法需要较强的假设性, 这增加了个人的主观性作用. 以上的这些局限性限制了非监督学习和半监督学习方法在情感分类任务中的应用, 所以目前情感分类大都采用监督学习的方法.

利用监督学习进行情感分类目前有两种情况: 一种是借鉴传统文本分类的方法, 利用特征表示文本, 利

①收稿时间:2014-02-18;收到修改稿时间:2014-03-17

用机器学习的方法预测情感倾向[2]. 但由于没有考虑特征之间的关系(如位置关系), 分类性能不够好. 另一种通过分别累计文本中的积极情感和消极情感来判断文本的整体情感倾向, 这类方法更为适合情感分类问题, 文献[3]的研究也表明定量描述特征的模糊性可以提高分类效果. 然而第二种分类方法只考虑了特征对文本属于某类别的支持程度, 忽视了特征对文本不属于某类别的支持程度, 因此并没有完全的利用从语料库中提取的相关信息.

针对情感分类问题, 本文提出了一种基于区间直觉模糊的情感分类模型, 采用了区间直觉模糊算子来计算特征词的区间直觉模糊数, 进而构建了特征词的区间直觉模糊集; 同时对程度副词、转折词、否定词进行定性处理, 从而获得较高的分类精度; 最后通过情感合成确定文本的情感倾向, 以期达到更为准确的情感分类.

## 1　基于区间直觉模糊的情感分类模型

1983 年, Atanassov 在 Zadeh 的模糊集合的研究基础上, 拓展了模糊理论, 提出同时考虑了隶属度、非隶属度、犹豫度三个方面的信息的直觉模糊集的概念. Atanassov 和 Gargov[4]进一步分别把隶属度和非隶属度的值由[0,1]之间的某个数字拓展为一个[0,1]的子区间. 区间直觉模糊比直觉模糊的进步在于犹豫度不再是一个数而是一个范围,区间直觉模糊集在处理不确定信息方面有更强的适用性.

本文基于区间直觉模糊, 提出了一种新的情感分类模型, 如图 1 所示. 首先在情感倾向性分析中引入预区间直觉模糊机制, 利用区间直觉模糊集的隶属度、非隶属度和犹豫度分别定量地描述情感特征词; 同时根据区间直觉模糊数和区间直觉模糊集进行情感分类, 利用区间直觉模糊算术集结算子合成句子级和文本级的情感倾向, 从而获得准确率较高的情感分类模型和情感倾向性分析结果.

本模型首先对语料进行预处理, 通过预处理得到情感特征词和修饰情感特征词的程度副词、转折词、否定词等; 然后得出情感特征词所对应的区间直觉模糊数, 并确定程度副词、转折词的权重和“否定词+特征词”的区间直觉模糊数; 再进一步进行词组级、句子级、文本级的情感合成, 最后得到语料最终的情感倾向.

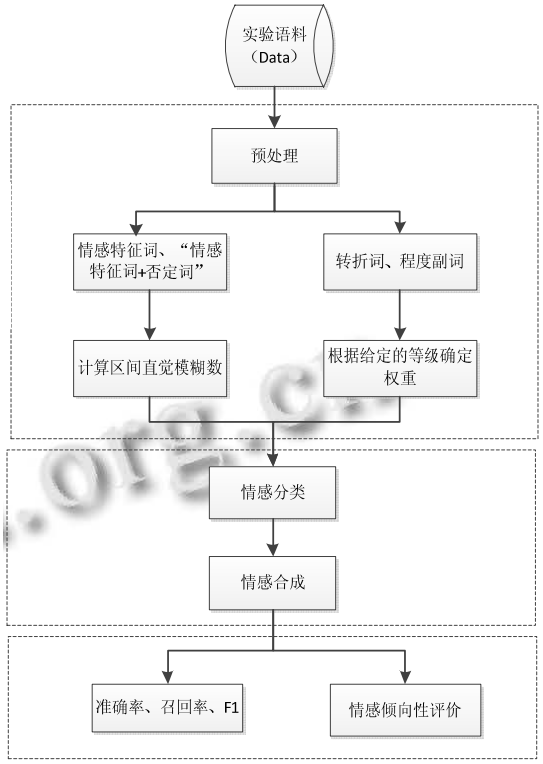


图 1　基于区间直觉模糊的情感分类模型

## 2　情感分类模型的步骤

### 2.1　文本预处理

#### 2.1.1　特征词提取

特征提取的 N-gram 算法是以短句为单位进行特征提取的. 当 $N=2$ 时即为 Bigram 特征提取算法. Zhongwu Zhai 等[5]的实验表明基于 Bigram 的特征提取算法是 N-gram 特征提取算法中效果最佳的算法. 本文采用斯坦福大学基于 Bigram 的特征提取软件 Stanford Parser[6]进行情感特征词的提取.

#### 2.1.2　程度副词、转折词和否定词的确定

修饰情感特征词的程度副词、转折词、否定词在表达感情倾向时也起着很重要的作用. 比如程度副词中 “有点”和 “相当”所表达的情感色彩有很大的区别. 我们结合认知心理学和语言学的知识, 采用朴闰柱[7]给出的程度副词词典和其中确定的从 1 到 4 之间的程度等级来表示程度副词加强的程度(如下表 1 所示). 本文规定程度副词的作用范围为句中其后的首个特征词.

转折词通常指明了文本的中心思想所在, 对这类词本文事先进行了整理. 本文采用的转折词表参考金允经[8]等给出的转折词系列. 并采用从 1/9 到 9 之间

的数字来表示其对情感的程度(增强、减弱)作用, 即如果一个转折词的对应的权重系数小于 1, 说明其对特征的情感倾向程度有减弱的作用, 相反, 如果一个转折词的对应的权重系数大于 1, 说明对特征的情感倾向程度有增强的作用. 本文规定转折词的作用范围为转折词出现位置至下一个转折词出现之前或文本末尾.

在情感倾向分类时, 否定词不能直接的忽略掉. 本文会综合一定的规则[9]来确定 “否定词+特征词”词组的区间直觉模糊集. 并设定词组 “否定词+特征词”对应的区间直觉模糊集为 $< f_i, 0.75u, 0.75v >$.

表 1 程度副词表及其权重等级

| 等级 | 程度副词 |
|---|---|
| 1 | 略微、略略、略为、稍稍、稍许、稍为、稍、略、微 微、多少、不大、不小、不很、不甚、有些、有点 |
| 2 | 较、比较、较为、还、颇、很、挺、甚、大、好、多、还、老、够、特 |
| 3 | 尤其、尤为、格外、更、更加、更为、还、越发、越加、益发、愈加、愈为 |
| 4 | 最、顶、最为、绝顶、绝伦、无比、太、过、过于、过分、万分、分外 |

## 2.2 情感特征词区间直觉模糊数的确定

假设 $u_A(x)$ 表示 $x$ 对 $A$ 的隶属程度, $v_A(x)$ 表示 $x$ 对 $A$ 的非隶属程度, 它们组成的有序对称为区间直觉模糊数 $\alpha = (u_A(x), v_A(x))$. 简记为 $\alpha = (u, v)$. 称 $\pi(x) = 1 - u_A(x) - v_A(x)$ 为 $x$ 对 $A$ 的犹豫程度, 等于区间直觉模糊数的不确定区间 $(u_A(x), 1 - u_A(x))$ 的长度. 以此形式来量化不确定性, 正是区间直觉模糊集比传统模糊集的优越之处.

我们综合王海[9]和属性区间直觉模糊数的计算公式, 对区间直觉模糊数的隶属度 $u_i$ 和非隶属度 $v_i$ 分别做如公式(1)、(2)定义:

$$u_i = p(pos / f_i) = \frac{p(pos, f_i)}{p(f_i)} \tag{1}$$

$$v_i = p(neg / f_i) = \frac{p(neg, f_i)}{p(f_i)} \tag{2}$$

其中 $p(pos, f_i)$ 表示特征 $f_i$ 出现时文本属于 $pos$ (积极)的概率; $p(neg, f_i)$ 表示特征 $f_i$ 出现时文本属于 $neg$ (消极)的概率. 显然, 由于 $0 \le p(pos, f_i) + p(neg, f_i) \le 1$, 所以有 $0 \le u_i + v_i \le 1$. 最后根据公式(1)、(2)的定义得到情感特征词所对应的区间直觉模糊数, 其中部分特征词的区间直觉模糊数如下表 2 所示:

表 2 部分情感特征词的区间直觉模糊数

| 情感词 | 区间直觉模糊数 | 情感词 | 区间直觉模糊数 |
|---|---|---|---|
| 好 | (0.748, 0.243) | 失望 | (0.185, 0.812) |
| 糟糕 | (0.035, 0.956) | 郁闷 | (0.225, 0.759) |
| 美 | (0.779, 0.213) | 愉快 | (0.613, 0.382) |
| 温馨 | (0.893, 0.098) | 一般 | (0.429, 0.562) |
| 清晰 | (0.689, 0.301) | 模糊 | (0.073, 0.926) |
| 魅力 | (0.855, 0.141) | 经典 | (0.826, 0.172) |
| …… | | | |

文本情感分类最重要的是获得文本作者所表达的情感倾向, 对于一篇具体的文本来说, 除了 $pos$ 、$neg$ 之外, 还可以是中立的. 对于特征 $f_i (i = 1, 2, \cdots, n)$, 区间直觉模糊集 $< f_i, u, v >$ 表示特征支持文本属于 $pos$ 或者 $neg$ 类的概率. 例如: $<$清晰, 0. 689, 0. 301$>$ 表示 “清晰”对文本属于 $pos$ 的支持度为 68.9%, 对于文本属于 $neg$ 的支持度为 30.1%, 还有 1%的中性倾向.

## 2.3 情感合成

以往情感分类算法中对于句子级和段落的情感合成仅仅采用简单相加(例如算术求和等)的方法, 这种方法[10]明显具有很大的缺陷和不合理性. 本文根据区间直觉模糊理论定义用于融合一组区间直觉模糊信息的集结算子来解决句子级和段落的情感合成问题. 区间直觉模糊加权平均算子的定义如下:

设 $\alpha = ([a_j, b_j], [c_j, d_j])(j = 1, 2, \cdots, n)$ 为一组区间直觉模糊数,令:

$I - IFWAA : Q^n \to Q$, 若设

$$I - IFWAA_\omega(\alpha_1, \alpha_2, \cdots, \alpha_n) = \sum_{j=1}^{n} \omega_j \alpha_j$$

$$= ([1 - \prod_{j=1}^{n} (1 - a_j)^{\omega_j}, 1 - \prod_{j=1}^{n} (1 - b_j)^{\omega_j}], [\prod_{j=1}^{n} c_j^{\omega_j}, \prod_{j=1}^{n} d_j^{\omega_j}])$$

其中, $\omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$ 为 $\alpha$ 的属性权重, 满足 $\omega_j \in [0, 1]$ 和 $\sum_{j=1}^{n} \omega_j = 1$. 则称函数 $I - IFWAA$ 为 $n$ 维区间直觉模糊数加权算术平均( $I - IFWAA$ )算子.

根据区间直觉模糊加权平均算子对词组级、句子级和段落进行情感合成. 设特征 $x_{ki}$ 对应的区间直觉模

糊集为 $< f_x, u_x, v_x >$，修饰特征 $x_{ki}$ 的词对应的 $N_{ki}$ 个权重系数为 $W_j$，其中 $j = 1, 2, \cdots, N_{ki}$. 则第 $i$ 个词组的积极( $pos$ )倾向可以由区间直觉模糊数 $(u_{ki}, v_{ki})$ 表示:

$$(u_{ki}, v_{ki}) = \sum_{j-1}^{N_{ki}} w_j (u_x, v_x) \tag{3}$$

对于句子级的情感倾向合成，由于特征词 $x_{ki}$ 将句子分为词组的集合，假设第 $k$ 个句子共包含 $M_k$ 个词组，则利用区间直觉模糊算术平均集结算子，第 $k$ 个句子的积极( $pos$ )倾向用 $(u_k, v_k)$ 可以表示为(其中 $k = 1, 2, \cdots, K$ ):

$$(u_k, v_k) = \frac{1}{m_k} \sum_{i=1}^{m_k} (u_{ki}, v_{ki}) \tag{4}$$

而对于整段文本的情感倾向合成，采用如下方法:

假设文本的情感倾向用 $(u, v)$ 表示，而且整段文本由 $S_k$ 个句子合成，并将 $\omega_k$ 作为修饰第 $k$ 个句子的转折词的权重，则利用加权平均集结算子文本的积极 $(pos)$ 倾向合成方法为(其中 $k = 1, 2, \cdots, K$ ):

$$(u_{pos}, v_{pos}) = \frac{1}{S_k} \sum_{i=1}^{m_k} \omega_k (u_k, v_k) \tag{5}$$

文本的消极 $(neg)$ 倾向合成方法为: 在式子 3 中，交换 $u_x$ 与 $v_x$，则可得情感特征词消极 $(neg)$ 倾向的区间直觉模糊数; 然后重新执行句子级的情感合成公式(即公式 4); 再利用算术平均集结算子得到文本的 $neg$ 倾向，记为 $(u_{neg}, v_{neg})$.

### 2.4 情感倾向分类

如果最终结果为 $(u_{pos}, v_{pos}) > (u_{neg}, v_{neg})$，则评论文本最终表现出的情感倾向属于积极 $(pos)$ 类; 反之假如 $(u_{pos}, v_{pos}) < (u_{neg}, v_{neg})$，则评论文本最终表现出的情感倾向属于消极 $(neg)$ 类; 若上面两种情况都不符合，则评论文本最终表现出的情感倾向属于中性(即无法判断其情感倾向).

## 3 实验结果

本文选取 IMDB 中文网的 20 个电影的相关评论共 2000 条作为实验语料，其中 1600 条作为分类器训练语料，并对剩余语料利用区间直觉模糊模型分类实验，将得出的电影评分与 IMDB 中文网的评分进行比较(如下图 2 所示)来体现基于区间直觉模糊模型的情感分类方法的性能和准确率.

通过对上图数据进行相关性分析，得出两组数据的相关系数约为 0.91-0.92，因此可以看出基于区间直觉模糊模型的情感分类方法的准确率还是比较高的. 例如: 电影《十二怒汉》的评论 “很精彩. 本以为黑白电影的魅力并不如此大，但看完这部话剧般的电影后实在触动很大，感叹美国公民维护法律的意识. “利用区间直觉模糊进行情感分类，使得评论者使用的转折语气和情感词在整篇评论的情感倾向分析中更为准确的表达出来，最终得出的整篇评论的倾向性为 91%(而 IMDB 中电影《十二怒汉》的评分为 8.9). 再比如: 电影《教父》的评论“很早就看过了这部经典的电影，在电影的结尾才真正了解了‘教父’的威望，很值得一看” 最终表达的积极的情感倾向为 92.8%(IMDB 中电影《教父》的评分为 9.2)，所以就分类结果而言，目前基于区间直觉模糊模型的情感分类方法是可行的，并且与人工分类的结果吻合度较高.

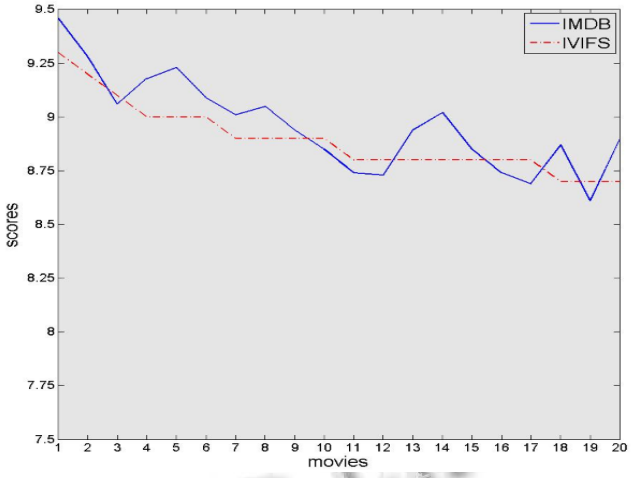


图 2　IVIFS 分类的评价结果和 IMDB 评分比较结果

我们还将此模型的分类结果和当前分类结果较好的方法[11](基于 SVM 的电影评论情感分类方法)对电影评论进行情感分类比较实验，数据仍然采用 IMDB 中的评论语料. 由于 $\beta$ 表示对 $P$ (准确率)和 $R$ (召回率)的侧重程度，而本文对二者一样重视，因此 $\beta$ 取值为 1. 最终实验结果如下表 3 所示:

表 3　实验结果

| 分类方法 | 类别 | $P$/% | $R$/% | $F1$/% |
|---|---|---|---|---|
| SVM | pos | 92.51 | 86.04 | 89.16 |
| | neg | 86.79 | 92.97 | 89.77 |
| IVIFS | pos | 93.19 | 91.13 | 92.15 |
| | neg | 91.25 | 93.69 | 92.45 |

通过实验证明基于区间直觉模糊模型的情感分

类方法准确率更高, 在 *pos* 类倾向的语料中召回率较高, 而在 *pos* 类和 *neg* 类的语料中都有很高的 F1 值. 所以就分类结果而言, 目前基于区间直觉模糊模型的情感分类方法还是可行的.

## 4 结语

区间直觉模糊集利用区间数来表示直觉模糊集中的隶属度、非隶属度和犹豫度, 从而能更加灵活地表达不确定信息.利用区间直觉模糊理论合成特征的情感倾向, 并在对电影评论情感倾向分类的实验中获得了较高的精度, 从而验证了方法的正确性与实用性. 这对从整体把握网络信息、为热点问题或突发事件提供决策支持都是很有价值的. 而如何调整模型以进一步提高分类准确度, 以及如何利用本文的分类方法挖掘隐含信息(比如在实验的背景下, 对影院引进影片的模糊综合评价作决策支持)获取更为准确的情感倾向, 则是将来需要进一步研究的方面.

## 参考文献


1 郝媛媛.在线评论对消费者感知与购买行为影响的实证研究[学位论文].哈尔滨:哈尔滨工业大学,2010.

2 徐军,丁宇新,王晓龙.使用机器学习方法进行新闻的情感自动分类.中文信息学报,2007,21(6):95–100.

3 付雪峰,刘邱云.不确定性推理在文本分类上的应用研究.江西师范大学学报:自然科学版,2007,31(4):383–386.

4 Atanassov K. Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 1986, 20: 87–96.

5 Zhai ZW, Xu H, Kang BD, Jia PF. Exploiting effective features for Chinese sentiment classification. Expert Systems with Applications, 2011, 38(8): 9139–9146.

6 Stanford Parser. http://nlp.stanford.edu/software/index.shtml.

7 朴闰柱.现代汉、韩程度副词的比较[学位论文].北京:清华大学,2004.

8 金允经,金昌吉.现代汉语转折连词组的同异研究.汉语学习,2001(2):34–40.

9 王海,冯向前,钱钢.网页在线评论情感倾向的直觉模糊分类.计算机工程与应用,2013,49(1),148–151.

10 Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proc. of the Conf on Empirical Methods in Natural Language Processing(EMNLP). 2002. 79–86.

11 Liu B. Sentiment analysis: A multifaceted problem. IEEE Intelligent Systems, May/June 2010, 25(3): 76–80.

12 余永红,向小军,商琳.并行化的情感分类算法的研究.计算机科学,2013,40(6),206–210.

13 徐泽水.区间直觉模糊信息的集成方法及其在决策中的应用.控制与决策,2007,22(2):215–219.

14 彭振文.区间直觉模糊集的聚类算法研究[学位论文].厦门:厦门大学,2009.