doi: 10.3969/j.issn.0490-6756.2019.01.009

基于扩展的情感词典和卡方模型的中文情感特征选择方法

胡思才[1,2], 孙界平[1], 琚生根[1], 王 霞[1], 龙 彬[1,3], 廖 强[4]
(1. 四川大学计算机学院, 成都 610065; 2. 解放军 61920 部队, 成都 610505;
3. 解放军 78179 部队, 成都 611130; 4. 四川大学外国语学院, 成都 610065)

摘 要: 根据经典的特征选择方法在中文情感评论文本中应用的缺陷和不足,提出了一种改进的中文情感特征选择方法. 目前,现有的情感特征选择方法普遍只利用了特征项在褒贬类中的统计信息,忽略了情感极性值对特征选择的影响;同时情感文本中否定词会带来特征项情感极性反转的情况,为特征选择带来较大的负面影响. 针对这些问题,首先对情感文本中的否定词进行了检测和判定,对否定词界定范围内的情感特征词进行反义变换处理,有效的解决了情感文本中极性反转的问题. 同时还将特征项的情感极性值和其在类中的频率特点两个因素融入到卡方特征选择模型(CHI)中,从而提升了卡方模型在文本情感特征选择的效果. 实验结果表明,本文算法较其他算法在多个领域数据集上的情感分类准确率提高了 1.5%左右.

关键词: 情感词典; 卡方模型; 特征选择; 知网; 否定词
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2019)01-0037-08

Chinese emotion feature selection method based on the extended emotion dictionary and the chi-square model

HU Si-Cai[1,2], *SUN Jie-Ping*[1], *JU Sheng-Gen*[1], *WANG Xia*[1], *LONG Bin*[1,3], *LIAO Qiang*[4]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. Troops 61920 of PLA, Chengdu 610505, China; 3. Troops 78179 of PLA, Chengdu 611130, China;
4. College of Foreign Languages and cultures, Sichuan University, Chengdu 610065, China)

Abstract: According to the defects and deficiencies of the classical feature selection method in Chinese comment text, an improved method is proposed for Chinese emotion feature selection. The current existing emotion feature selection method generally only used the statistical information of the feature items in the classes, ignoring the influence of the emotional polarity value on feature selection. Meanwhile the negative words in the sentiment text can cause the reversal of the emotional polarity of the characteristics, which would bring great negative effects on the feature selection. To tackle these problems, antisense transformation processing of the emotion characteristic words is performed in the range of the negative words, which effectively solves the emotional polarity reverasl in the sentiment text. The paper also introduces the emotional polarity values and its frequency into the chi-square model (CHI) to improve the effect of CHI on the emotion feature selection. The experimental results show that the proposed

收稿日期: 2018-05-24
基金项目: 四川省重点研发项目(2018GZ0182)
作者简介: 胡思才(1987-), 男, 硕士生, 研究方向为数据科学.
通讯作者: 琚生根. E-mail: jsg@scu.edu.cn

method can improve the accuracy of emotion classification by about 1.5% in multiply domain data sets, compared with other algorithms.

Keywords: Emotional dictionary; Chi-square model; Feature selection; Hownet; Negative word

1 引 言

近几年,情感分析已成为自然语言处理中的一个热点问题,其在市场预测分析、民意调查、智能导购和大众评论等诸多领域都有着广阔的应用空间和发展前景. 情感评论文本经过特征向量化后,可能会产生多维度灾难,不利于模型训练,因此,文本特征选择尤为重要. 情感文本特征选择一般分为两类,一类是利用情感词典直接通过查表法提取情感特征词[1],优点是简单高效,缺点是未考虑情感特征对模型的影响程度,同时,忽略了情感词典外的特征;另一类是基于统计的特征选择方法,如卡方模型、IG 方法、WLLR 方法、MI 方法和 TF-IDF 方法等,利用统计方法对训练文本特征项词频和文档频率进行计算,从而求出每个特征项的权重系数,将系数高的特征项作为最终选择的特征项. 目前,第二种方法比较常用,前人经过实验研究得出卡方模型和 IG 方法是目前最有效的特征选择算法之一的结论,特别是在类别分布相对均衡时明显优于其他方法. 因此探讨并修正这两种方法的缺陷和不足,提高特征选择的效率具有非常重要的实际意义.

近年来,一些学者针对 IG 算法和卡方模型算法的不足作了一些改进工作[2-4]. TFIG 算法[2]利用特征项出现频率(包括特征项未出现、出现一次以及出现多次三种情况)进行信息增益的计算,该方法对长文本分类效果不错,但应用在短文本情感分析中,则效果一般. CHI_LF 算法[3]将特征项的类内频数、类内位置、类间频数分布等信息以及特征项与类别间正负相关度融入卡方模型中,在类偏斜条件下效果较好,但是应用在类别分布相对均衡的短文本中性能提升有限.

情感评论文本还有一个显著特点是:评论文本中普遍存在否定词. 否定词的出现使其界定范围内的情感极性发生反转,会给基于词袋的监督学习分类算法带来一定的负面影响. 因此在进行特征选择之前需对否定词进行处理. Xia[5]等人利用否定词检测和特征词统计方法对文本的情感特征词转换来获取对应的转换文本,充分利用扩展后的文本进行成对的训练和测试. 该方法减少了对额外的语料数据的依赖,对各领域的情感文本适应性较强,但该方法因为要对否定词界定范围外的大部分情感特征词进行反转,存在一些特征噪声,如果对中文分词效果不好的情感文本,会有一定的负面影响.

针对目前研究难点和不足,本文提出了基于扩展的情感词典和卡方模型的中文情感特征选择方法. 首先结合知网[6]和改进的基于字频的极性值计算方法对词典中每个情感词进行极性值计算,建立带有情感极性值的词典;然后通过对评论文本子句中否定词及其出现个数进行检测,并对否定词界定范围内的情感词进行处理[7,8],从而有效地限制了否定词带来的负面影响;最后将特征项情感极性值和相关的特征类间词频与卡方模型进行融合,基本上修正了卡方模型的缺陷,使得改进后的模型具有更好的情感特征选择性能,能够有效地提高文本情感分类效果.

2 建立带情感极性值的词典

情感特征项的极性值计算主要有以下几种方法:一种是基于语义的方法,利用知网义原树来计算情感特征项与褒贬义基准词之间的相似度,将该相似度作为该情感特征项的极性值[9,10],基于知网的情感极性值计算依赖于基准词,同时对于一些未登录词无法计算其情感极性值;一种是基于统计的方法[11],通过统计情感特征项与基准词之间的共现信息来得到其极性值,这种方法计算简单,但是需要大语料库的支持,同时也依赖于基准词,时间复杂度高;还有一种是基于字频的情感计算方法[12],利用情感特征项的单字在褒贬义情感词典中出现的字频信息来获得其情感极性值,该方法不依赖基准词,同时不需要大量的语料库来进行共现搜索,计算简单,时间复杂度低,但其比较依赖于情感词典的词汇量.

本文结合知网和基于字频的情感计算方法,并对其进行了相关改进,获得最终的情感极性值.

2.1 词典扩展及预处理

(1) 本文结合知网发布的情感词典和台湾大学自然语言处理实验室提供的简体中文情感词典,对该词典进行预处理(包括情感词的合并和删除),得到扩展的情感词典.

（2）将情感词典中每个词分解成单字，然后统计所有字在褒贬义情感词典中出现的数量，形成带有字频的字典.

2.2　基于知网的相似度计算

知网是一部比较完善的、以汉语和英语所代表的概念为对象的、以揭示概念与概念之间的属性关系为基础的常识知识库.

知网中的汉语词汇描述都是基于“义原”这一基本概念. 由于汉语中一个词语在不同的语境中可能会表达出不同的概念，因此，知网将词语整理成若干个概念的集合，每个概念都由一组描述义原所表示. 如表 1 所示，词汇“讨厌”有三个概念，其中两个是形容词，一个是动词，每一个概念由多个义原表示，如“aValue | 属性值”，“easiness | 难易”. 刘群[6]等人详细描述了两个词语基于知网的语义相似度的计算方法.

表 1　词汇的义原表示形式

Tab. 1　The primitive expression of the concept

概念	词性	描述义原
讨厌	ADJ	aValue|属性值，easiness|难易，difficult|难，undesired|莠
讨厌	ADJ	aValue|属性值，impression|印象，bad|坏，undesired|莠
讨厌	V	disgust|厌恶
一诺千金	N	text|语文，* MakeAppointment|约定，$ obey|遵循
安检	N	fact|事情，check|查，# tour|旅游，# safe|安
昂	V	CausePartMove|部件他移，direction=upper|上

情感词基于知网的极性值计算方法是一种基于情感词典的方法，它是选取一些褒贬义基准词，然后根据情感词与褒贬义基准词的紧密程度对情感词进行计算. 本文采用 40 对褒贬义基准词[9]，其中褒义基准词表示为 p_i（i 取 1～40），贬义基准词表示为 n_j（j 取 1～40）. 具体计算公式如下.

$$e_{\text{hownet}}(w) = \sum_{i=1}^{40}\text{sim}(p_i, w) - \sum_{j=1}^{40}\text{sim}(n_j, w) \tag{1}$$

其中，$\text{sim}(p_i, w)$ 表示该情感词与褒义基准词 p_i 的相似度；$\text{sim}(n_j, w)$ 表示该情感词与贬义基准词 n_j 的相似度；$e_{\text{hownet}}(w)$ 表示该情感词的极性值，当 $e_{\text{hownet}}(w) > 0$，则将该情感词归为褒义词，$e_{\text{hownet}}(w) < 0$，则将该情感词归为贬义词，数值大小代表情感词的情感强烈程度.

再对 $e_{\text{hownet}}(w)$ 进行最大最小归一化处理，即为该情感词的基于知网的极性值.

2.3　基于字频的相似度计算

汉语词语的意义是通过组成该词的汉字来表达的，因此可将情感词的极性值用组成该词的所有汉字的倾向度函数来表示. 而汉字的倾向度，通过计算该汉字在情感词典中出现的频率来获得[12].

设情感词 W 由 k 个汉字组成，表示成 $W = w_1w_2\cdots w_k$，汉字 w_i（i 取 1～k）的情感倾向度为 T_i，则情感词的极性值 $e_{tf}(w)$ 的计算如式（2）所示.

$$e_{tf}(w) = \sum_{i=1}^{k}\lambda_i T_i \tag{2}$$

其中，λ_i 表示组成汉字 w_i 的权值，$\sum_{i=1}^{k}\lambda_i = 1$，情感词的意义是由各组成汉字共同决定的，这里可将每个汉字设置为相等的权值. T_i 表示汉字 w_i 的情感倾向度，可由式(3)～式(5)共同得出.

$$\bar{T}_i = \frac{1}{\sigma}\left(\frac{f_{P_i}}{\sum_{j=1}^{n} f_{P_j}} - \frac{f_{N_i}}{\sum_{j=1}^{m} f_{N_j}}\right) \tag{3}$$

其中，$\bar{T}_i$ 表示汉字的初始倾向度；f_{P_i}、f_{N_i} 分别表示汉字 w_i 在褒义词典和贬义词典中出现的频率；n、m 分别表示在褒义词典和贬义词典中不同汉字的数量. σ 为归一化处理因子，其公式为

$$\sigma = \left(\frac{f_{P_i}}{\sum_{j=1}^{n} f_{P_j}} + \frac{f_{N_i}}{\sum_{j=1}^{m} f_{N_j}}\right) \tag{4}$$

根据式(3)的计算方法，如果该汉字在褒义词典中出现但次数不定，且在贬义词典中未出现，则算出的情感倾向度都为 1，这明显与实际不相符，汉字情感倾向度通常是随着褒贬义词典中的出现频率差值的增大而增大，因此需对式(3)中的 $\bar{T}_i$ 进行加权修正，如式(5)所示.

$$T_i = \left(1 - \frac{1}{\lambda + 1}\right)\bar{T}_i \tag{5}$$

其中，λ 表示该汉字在褒贬义词典中出现的频率差值的绝对值.

根据上述公式，可得到情感词基于字频的极性值. 对于任何一个情感词 W，其倾向度 $e_{tf}(w)$ 的取值范围为 -1～1，$e_{tf}(w)$ 大于 0 表示该词表达正面情感，小于 0 则表达的是负面情感，$e_{tf}(w)$ 的绝对值越大，说明该词表达的情感越强烈.

2.4　建立带有情感极性值的词典

本文结合知网和基于字频的情感值计算方法，并通过线性回归模型获取这两种方法的最佳权重

值,获得最终的带有情感极性值的词典,具体建立过程如图 1 所示.

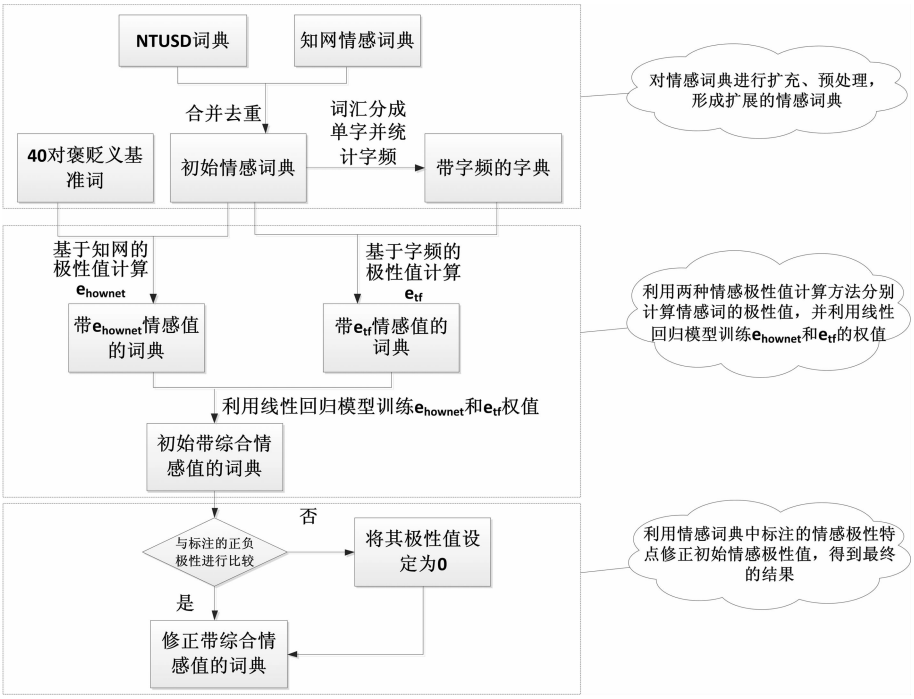


图 1 带有情感极性值的词典建立流程图

Fig. 1 The flowchart of creating the dictionary with emotional polarity value

其中,本文特征词极性值由式(1)和式(2)加权可得. 计算公式如式(6)所示.

$$e=\alpha\times e_{\text{hownet}}+\beta\times e_{tf} \tag{6}$$

其中,α 和 β 的取值范围均为 0~1,且 $\alpha+\beta=1$. 本文利用线性回归模型对情感词典中已知极性的所有登录词进行实验,得到其情感极性判断准确率随 α 变化的曲线图如图 2 所示. 当 $\alpha=0$ 时,情感词的极性完全由 e_{tf} 决定,当 $\alpha=1$ 时,情感词的极性完全由 e_{hownet} 决定,极性判断准确率较低. 当 $\alpha=0.451$, $\beta=0.549$ 时,准确率达到最高值.

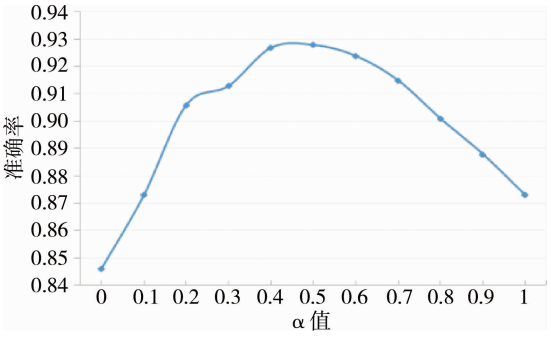


图 2 情感极性判断准确率曲线图

Fig. 2 The accuracy of the emotional polarity judgment

利用式(6)计算出情感词典中的所有特征词极性值,如果特征词极性判断错误,则将其极性值设定为 0,如果极性判断正确,则计算结果即为该词的情感极性值.

3 对否定词的处理

在对情感评论文本的机器学习分类中,词袋模型是目前应用最广泛的特征向量化方式,但是词袋模型忽视了句中单词的顺序,破坏了句法结构,不能对否定词进行处置. 如果评论文本中出现否定词,会将其界定范围内的句子情感极性转换为其相反的极性,如:“我/喜欢/这本/书”和“我/不/喜欢/这本/书”,第二个评论句中出现了否定词,将该评论句的极性进行反转,而在词袋中,这种现象体现不出来,因而通过机器学习分类,极易将第二句错误地归为褒义类. 为了减少这种情况发生,本文在进行特征选择之前,首先对否定词及其界定范围内的文本进行处理.

3.1 建立反义词典

在情感评论文本中,一般用动词、形容词、副词来形容对事物的情感描述. 因此,大部分动词、形容词、副词的情感极性相对其他词性来说,比较明显. 基于这个特点,本文对训练集中所有的情感词进行词性判断,将属于动词、形容词、副词的情感词提取出来纳入词典中[13],通过式(6)对词典中的情感词极性进行计算,然后将获得的褒义词集和贬义词集

按照情感强度从大到小的顺序进行排序. 设 P_i 为褒义极性值排名第 i 位的褒义情感词，N_i 为贬义极性值排名第 i 位的贬义情感词，即可形成褒贬情感词一一对应关系（P_i，N_i），其中 $i \leqslant \min(len(P_i), len(N_i))$，$len(P_i)$ 表示褒义词集的数量，$len(N_i)$ 表示贬义词集的数量. 对于没有形成对应关系的情感词，由于这些词极性值很小，基本上可以忽略其极性值，所以在反义词典中直接将这一部分情感词删除.

3.2　检测和处理否定词

利用查表法对情感评论文本训练集和测试集的所有否定词进行检测，如表 2 所示. 确定否定词出现的位置.

表 2　否定词表

Tab. 2　Negative word list

否定词
不、不能、不应该、不让、不许、不允许、不是、不太、不会、不曾、不要、不必、不用、毫不、绝不、永不、从不、未、未必、从未、勿、弗、无、毫无、木有、禁、禁忌、禁止、没、没有、否、非、莫、忌、防止、拒绝、杜绝

针对否定词界定的范围，本文直接将否定词开始到子句（情感评论文本由断句标点分隔成多个子句）结尾词之间的范围作为否定词的界定范围[5]. 在否定词界定的范围内，将出现在反义词典中的情感词全部转换为其对应的反义词，其他情感词保持不变，并将否定词去掉. 如："我/不/喜欢/这本/书"转换后变成"我/讨厌/这本/书".

如果子句中出现偶数个否定词，如："我/不/认为/这个/酒店/不/干净"，"没有/一个/人/不/满意"等，这类子句是双重否定句，否定词界定范围内的情感词极性并没有被反转，因此对出现偶数个否定词的子句，直接将否定词去掉，否定词界定范围内的情感词保持不变；对出现奇数个否定词的子句，将否定词界定范围内的情感词进行反转，同时将否定词去掉.

将否定词处理结束后，再对所有的情感特征词进行特征选择.

4　改进的特征选择方法

4.1　卡方特征选择模型

卡方模型利用了统计学中的"假设检验"的基本思想：首先假设特征词与类别直接是不相关的，如果利用卡方分布计算出的检验值偏离阈值越大，那么更有信心否定原假设，接受原假设的备则假设：特征词与类别有着很高的关联度. 卡方模型的计算公式如下.

$$\chi(t_i, c_j) = \frac{N(A_{ij}D_{ij} - C_{ij}B_{ij})^2}{(A_{ij}+C_{ij})(B_{ij}+D_{ij})(A_{ij}+B_{ij})(C_{ij}+D_{ij})} \tag{7}$$

其中，A_{ij} 表示包含特征项 t_i 且属于类 c_j 的文档数量；B_{ij} 表示包含特征项 t_i 但不属于类 c_j 的文档数量；C_{ij} 表示不包含特征项 t_i 但属于类 c_j 的文档数量；D_{ij} 表示不包含特征项 t_i 且不属于类 c_j 的文档数量；N 表示文档总数量.

4.2　将情感特征项的极性值融入卡方模型

将情感评论文本中的情感特征项逐一与带有情感极性值的词典比对，如果该情感特征项在词典中，则将词典中该特征项的值作为该特征项的情感极性值，例如"优秀"在词典中对应的值为 0.919，则将 0.919 作为"优秀"的情感极性值. 如果该特征项未出现在词典中，则利用式(6)进行计算，如"色彩鲜艳"，通过公式计算其情感极性值为 0.705. 最终获得文本中所有情感特征项的极性值.

从卡方模型定义能够发现，矩阵 $\boldsymbol{A}$ 表示包含特征项的文档属于各个类的数量，是特征选择中最重要的一项，它的值直接决定了特征选择的终值. 在其他情况都相同的情况下，矩阵 $\boldsymbol{A}$ 中包含特征项 t_i 且属于类 c_j 的值 A_{ij} 越大，则特征项 t_i 对类 c_j 的贡献值就越高，特征项 t_i 与类别 c_j 就具有越深的关联度，此时将特征项 t_i 选择出来能够提高文本分类的准确率. 同时，其他矩阵 $\boldsymbol{B}$、$\boldsymbol{C}$、$\boldsymbol{D}$ 都由矩阵 $\boldsymbol{A}$ 计算得到的.

但矩阵 $\boldsymbol{A}$ 只考虑了文档数量，对特征项自身的情感极性未加考虑，而情感极性值高的特征项在情感分类中具有较大的作用. 本文考虑到这两部分因素对分类的影响，对情感极性值与矩阵 $\boldsymbol{A}$ 进行了融合，将包含特征项的文档数量附上该特征项的权重（情感极性值），矩阵 A 演变为 $A'_{ij} = A_{ij} \times E_{ij}$. 例如在其他情况都相似的情况下，特征项"调节"，"喜欢"在褒义类中出现的文档数分别为 5，5，原矩阵 $\boldsymbol{A}$ 只是通过文档频率来判断这两个词对褒义类的重要程度，这种情况则很难区分哪个词对褒义类更重要. 而针对修正的矩阵 $\boldsymbol{A}'$，由于这两个词的极性值分别为 0.1，0.9，则其在矩阵 $\boldsymbol{A}'$ 中对应的值为 5.5，9.5，相当于为情感极性值高的特征项提供一个增量，在特征选择阶段能够较大可能地将在训

练语料中出现较少但情感极性值较高的特征给选择出来. 其中,E_{ij} 是一个情感强度矩阵,行表示情感特征项,列表示类,第一列表示贬义类,第二列表示褒义类,e_i 是利用情感词典和式(6)得到的特征项情感值,其中当 $e_i \geqslant 0$ 时,该特征项为褒义词,则 E_{ij} 中两类中的值分别为

$$\begin{cases} E_{i0} = 1 \\ E_{i1} = 1 + e_i \end{cases} \tag{8}$$

从式(8)可以看出,当该特征项为褒义词时,A'_{ij} 相当于给矩阵 $\boldsymbol{A}$ 中包含特征项 t_i 且属于类 c_1(褒义类)的值 A_{i1} 乘上了一个大的权重$(1+e_i)$,矩阵 $\boldsymbol{A}$ 中包含特征项 t_i 且属于类 c_0(贬义类)的值 A_{i0} 保持不变.

当 $e_i < 0$ 时,该特征项为贬义词,则 E_{ij} 的两类中值分别为

$$\begin{cases} E_{i0} = 1 + |e_i| \\ E_{i1} = 1 \end{cases} \tag{9}$$

从式(9)可以看出,当该特征项为贬义词时,A'_{ij} 相当于给矩阵 $\boldsymbol{A}$ 中包含特征项 t_i 且属于类 c_0(贬义类)的值 A_{i0} 乘上了一个大的权重$(1+|e_i|)$,矩阵 A 中包含特征项 t_i 且属于类 c_1(褒义类)的值 A_{i1} 保持不变.

按照上述公式求出 $\boldsymbol{A}'$ 后,再结合特征数量 N,求出 $\boldsymbol{B}'$,$\boldsymbol{C}'$,$\boldsymbol{D}'$,最后得出新的特征值.

$$\bar{\chi}(t_i, c_j) = \frac{N(A'_{ij}D'_{ij} - C'_{ij}B'_{ij})^2}{(A'_{ij}+C'_{ij})(B'_{ij}+D'_{ij})(A'_{ij}+B'_{ij})(C'_{ij}+D'_{ij})} \tag{10}$$

4.3 特征项类间频率信息与卡方模型结合

卡方模型考虑的总是包含特征项或者不含特征项的文档数目,未考虑到特征项在类中出现的频率,但特征频率也是特征选择的一个重要因素. 比如特征项 t_i 和 t_k 由卡方模型计算出的值比较接近,但是假如特征项 t_i 在类 c_j 中出现的频数很大,而 t_k 出现的频数很少,则 t_i 在类 c_j 的表现能力明显要比 t_j 强,但卡方模型却反映不出这方面的差异. 本文中利用 $TF(t_i, c_j)$ 表示特征项 t_i 在类别 c_j 中的词频数,公式为

$$TF(t_i, c_j) = \sum_{k=1}^{|c_j|} \frac{tf_k(t_i, c_j) - tf(t_i, c_j)_{\min}}{tf(t_i, c_j)_{\max} - tf(t_i, c_j)_{\min}} \tag{11}$$

其中,$tf_k(t_i, c_j)$ 表示特征 t_i 在属于类 c_j 的文档 d_k 中的词频数;$tf(t_i, c_j)_{\min}$ 表示特征 t_i 在属于类 c_j 的单文档中出现的最小词频数;$f(t_i, c_j)_{\max}$ 表示特征 t_i 在属于类 c_j 的单文档中出现的最大词频数;$|c_j|$ 表示属于类 c_j 的文档数量.

考虑到不同类别之间文档数目对其的影响,现对其进行归一化处理.

$$\overline{TF}(t_i, c_j) = \frac{TF(t_i, c_j)}{\sqrt{\sum_{j=1}^{|c|} TF(t_i, c_j)^2}} \tag{12}$$

$\overline{TF}(t_i, c_j)$ 即为特征项在各个类中归一化词频.

4.4 本文情感特征选择公式

本文将情感特征项的极性值和特征项的类间频率融入卡方模型中,最终的公式如下所示.

$$\chi_{TF}(t_i, c_j) = \overline{TF}(t_i, c_j) \times \bar{\chi}(t_i, c_j) \tag{13}$$

根据式(13),即可求得每个情感特征的修正卡方值,对所有的值按由大到小的顺序排列,取出特定数量的最大修正卡方值所对应的情感特征作为最终的情感特征,然后根据所选的情感特征对情感文本进行特征向量化,再利用文本分类器进行分类.

5 实验说明

5.1 实验数据

本文采用谭松波采集的酒店、书籍、笔记本三个领域的中文评论数据,每个领域各包含 2000 条正面和 2000 条负面评论. 数据在网站 http://www.searchforum.org.cn/tansongbo/corpus/ 下载.

表 3 实验数据

Tab. 3 Experimental data list

评论领域	褒义评论数目	贬义评论数目
酒店	2000	2000
书籍	2000	2000
笔记本	2000	2000

5.2 算法实验

本文主要对比前人算法文献[3]、文献[5]以及基础的 IG 特征选择算法(base 算法),来验证本文算法的有效性.

对每个数据集进行五折交叉验证:将数据集随机划分为 5 等份,其中,4 份用作训练数据,另外一份用作测试数据,按这个方式对数据集进行训练和测试,并利用朴素贝叶斯分类和支持向量机两种分类方法进行分类. 将得出的 5 个数值取平均值获得最终的结果. 实验结果如下所示.

(1) 对酒店中文评论数据进行朴素贝叶斯和

支持向量机分类的结果如图 3 所示.

图 3(a)是利用朴素贝叶斯方法对酒店中文评论数据进行分类，图 3(b)是利用支持向量机对酒店数据进行分类. 从图 3 可以看出，本文算法在特征数量为 200～3000 期间皆优于其他算法.

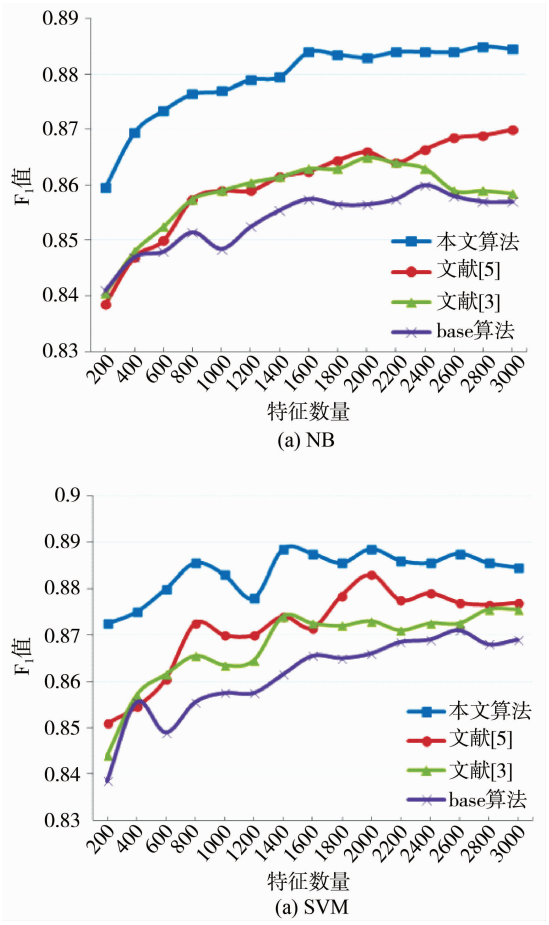
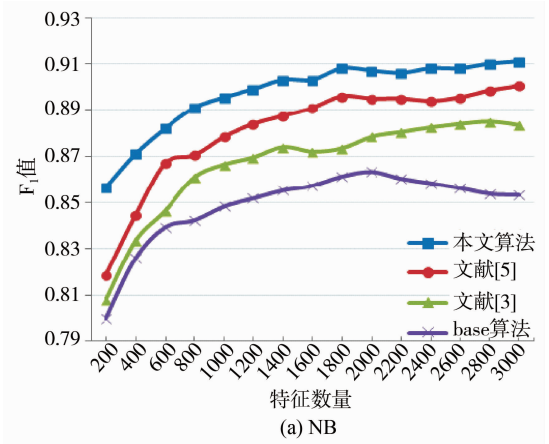


图 3　酒店数据分类精度曲线图

Fig. 3　The classification accuracy of the hotel data

(2) 对书籍中文评论数据进行朴素贝叶斯和支持向量机分类的结果如图 4 所示.

从图 4 可以看出，本文算法利用朴素贝叶斯分



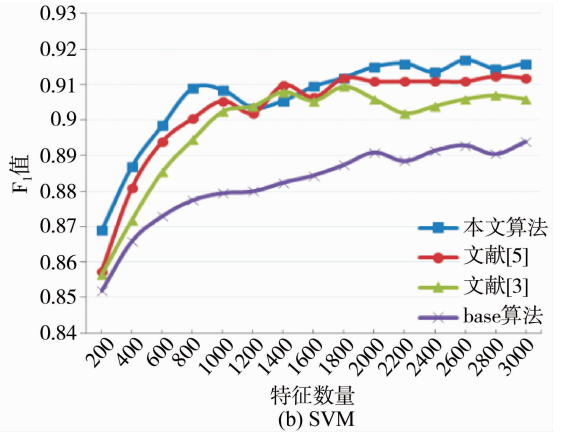


图 4　书籍数据分类精度曲线图

Fig. 4　The classification accuracy of the book data

类器时效果要好于其他算法，利用支持向量机分类器时，本文算法在特征数量为 2600 时，达到 F 最大值 0.917，而文献[5]、文献[3]、base 算法在特征数量为 2600 时，F 值分别是 0.911，0.906，0.893. 本文算法比其他算法要好.

(3) 对笔记本中文评论数据进行朴素贝叶斯和支持向量机分类的结果如图 5 所示.

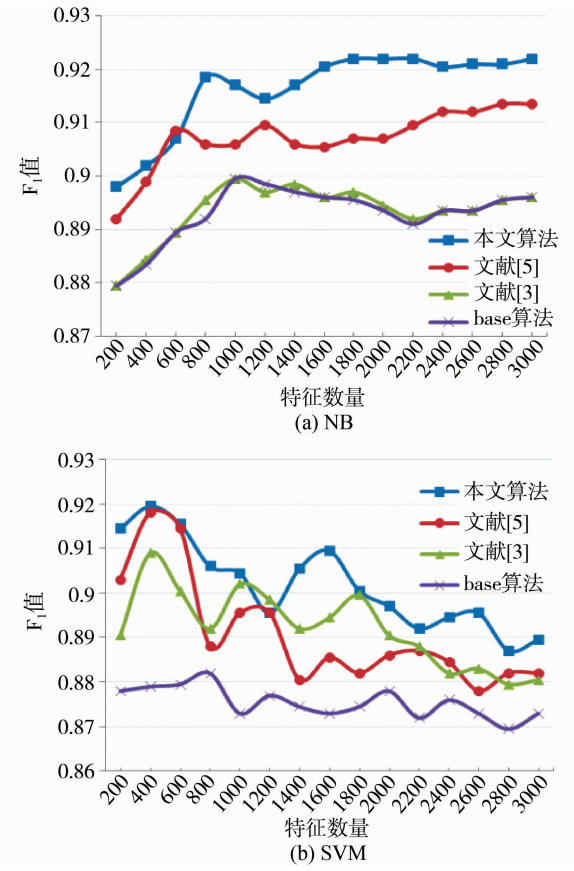


图 5　笔记本数据分类精度曲线图

Fig. 5　The classification accuracy of the notebook data

从图 5 可以看出,本文算法在特征数量为 200～3000 期间基本上优于其他算法. 本文算法利用朴素贝叶斯分类时,在特征数量为 1800 时,达到 F 最大值 0.922,文献[5]、文献[3]、base 算法在特征数量为 1800 时为 0.907, 0.897, 0.896,本文算法明显要好于其他算法. 在利用支持向量机分类时,文献[5]算法波动太大,文献[3]、base 算法虽然波动小,但最大 F 值较小,分别为 0.909, 0.882,本文算法基本上处于所有算法的最高值,F 值最大为 0.92,要好于其他算法.

本文通过对酒店、书籍、笔记本三个领域的中文评论数据进行实验来验证本文算法. 从图 3～5 中可以看出,本文对各个数据集进行情感特征选择后利用朴素贝叶斯分类时,相对于基础算法,F 值提高了 4%左右,相对于已有改进算法,F 值提高了 2%左右;利用支持向量机分类时,相对于基础算法,F 值提高了 3%左右,相对于已有改进算法,F 值提高了 1%左右. 说明本文算法相较于已有算法进一步提升了文本情感特征选择的效果.

6 结 论

本文在前人研究的基础上,将情感极性值和词频信息融入到卡方模型中进行情感特征项的选取,对特征选择方法进行了优化,该方法同时还考虑到否定词对文本情感极性的影响,对否定词进行检测和判断,并对否定词界定范围内的情感词进行处理,进一步提高了情感文本分类的准确性. 未来可以继续探索情感极性值与分类器中特征向量权重的关联,利用情感极性值优化情感文本分类器,提高分类器针对情感文本的分类精度.

参考文献:

[1] 杨奎, 段琼瑾. 基于情感词典方法的情感倾向性分析 [J]. 计算机时代, 2017, 3: 10.

[2] Xu Y, Chen L. Term-frequency based feature selection methods for text categorization [C]// Proceedings of the Fourth International Conference on Genetic and Evolutionary Computing(ICGEC). Shenzhen: IEEE Computer Society, 2010: 280.

[3] 宋阿羚, 刘海峰, 刘守生. 基于位置及词频信息的优化 CHI 文本特征选择方法 [J]. 计算机科学与应用, 2015, 5: 322.

[4] 石慧, 贾代平, 苗培. 基于词频信息的改进信息增益文本特征选择算法 [J]. 计算机应用, 2014, 34: 3279.

[5] Xia R, Xu F, Zong C, *et al*. Dual sentiment analysis: considering two sides of one review [J]. IEEE Trans Knowl Data En, 2015, 27: 2120.

[6] 刘群, 李素建. 基于《知网》的词汇语义相似度计算 [C]. 第三届汉语词汇语义学研讨会. 台北: [s. n.], 2002.

[7] Li S S, Lee S Y M, Chen Y, *et al*. Sentiment classification and polarity shifting [C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, 2010.

[8] 刘玉娇, 琚生根, 伍少梅, 等. 基于情感字典与连词结合的中文文本情感分类 [J]. 四川大学学报: 自然科学版, 2015, 52: 57.

[9] 王振宇, 吴泽衡, 胡方涛. 基于 HowNet 和 PMI 的词语情感极性计算 [J]. 计算机工程, 2012, 38: 187.

[10] 范弘屹, 张仰森. 一种基于 HowNet 的词语语义相似度计算方法 [J]. 北京信息科技大学学报: 自然科学版, 2014, 29: 42.

[11] 吴金源, 冀俊忠, 赵学武, 等. 基于特征选择技术的情感词权重计算 [J]. 北京工业大学学报, 2016, 42: 142.

[12] 徐晓丹, 段正杰, 陈中育. 基于扩展情感词典及特征加权的情感挖掘方法 [J]. 山东大学学报: 工学版, 2014, 44: 15.

[13] 张磊, 李梦诗, 陈黎, 等. 基于双层 HHMM 的产品评论特征和情感分类 [J]. 四川大学学报: 工程科学版, 2013, 45: 94.

引用本文格式:

中 文: 胡思才, 孙界平, 琚生根, 等. 基于扩展的情感词典和卡方模型的中文情感特征选择方法 [J]. 四川大学学报: 自然科学版, 2019, 56: 37.

英 文: Hu S C, Sun J P, Ju S G, *et al*. Chinese emotion feature selection method based on the extended emotional dictionary and the chi-square model[J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 37.