



基于改进特征选择方法的文本情感分类研究

刘洛辛^{1,2}, 陈晶^{1,3}, 王麒媛¹

- (1. 燕山大学信息科学与工程学院, 河北 秦皇岛 066004;
2. 河北省信息传输与信号处理重点实验室, 河北 秦皇岛 066004;
3. 河北省计算机虚拟技术实验室, 河北 秦皇岛 066004)

摘要: 提出了结合情感词典的改进信息增益特征选择方法。首先, 针对现有的信息增益特征选择存在注重特征词的文档频率而忽视语料均衡等问题, 提出了改进方法。其次, 考虑情感词对文本分类的影响, 提出了基于情感词典的特征选择 (information gain combining sentiment classification, IGSC) 算法进行文本分类。该算法通过对文本情感词进行匹配并结合情感词赋权重, 实现了特征降维并解决了文本数据稀疏影响分类性能的问题; 最后, 针对旅游评论数据集对所提出的特征选择方法进行了实验验证及分析。实验结果表明, 本文提出的改进文本情感分类特征选择方法在分类准确率、召回率和 F 值方面均得到了提升, 并且具有较好的分类稳定性。

关键词: 信息增益; 情感词典; 特征选择; 情感分类

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018250

Research on text sentiment classification based on improved feature selection method

LIU Mingxin^{1,2}, CHEN Jing^{1,3}, WANG Qiyan¹

1. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
2. Hebei Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao 066004, China
3. Computer Virtual Technology Laboratory in Hebei Province, Qinhuangdao 066004, China

Abstract: An improved information gain feature selection method based on sentiment dictionary was proposed. Firstly, aiming at the existing problems of information gain feature selection, such as paying attention to the frequency of feature word and ignoring the balance of corpus, an improved method was proposed. Secondly, considering the influence of sentiment words in text classification, a feature selection method IGSC (information gain combining sentiment classification) based on sentiment dictionary was proposed for text classification. By matching the text emotion words and combining the weight of emotion words, the feature dimension reduction was realized and the problem of text data sparseness affecting classification performance was solved. Finally, according to the proposed feature selection method of travel review data set for experimental verification and analysis, the experimental results show that the improved text sentiment classification feature selection method has been improved in terms of classification accuracy, recall and F value, and classification has better stability.

Key words: information gain, sentiment dictionary, feature selection, sentiment classification

收稿日期: 2018-03-06; 修回日期: 2018-09-09

通信作者: 陈晶, xychenjing@ysu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61602401, No.61472340)

Foundation Items: The National Natural Science Foundation of China (No.61602401, No.61472340)



1 引言

随着网络的飞速发展,人们不仅可以便捷地获取信息,而且也可以分享见闻、讨论热点话题或在电商平台上评价商品。目前众多的用户已参与到互联网中,并将其作为展现自我的情感平台,其结果会导致网络信息量爆炸式增长。面对海量数据,如何有效地进行文本信息挖掘与分析、准确地对文本进行情感分类具有重要的意义。因此,文本情感分析应运而生。

文本情感分析是指用自然语言处理、文本挖掘、机器学习和计算机语言学等对主观性文本中的信息进行一系列的分析处理过程,并归纳概括出其情感倾向性^[1]。文本情感分析通过对多元、海量和复杂的文本所表达的情感信息进行分类,其研究具有很多现实意义和应用价值。在舆情监测方面,通过人们在网络上对热点事件、社会公共事件以及政策评论等信息进行文本情感分析,可以帮助政府了解民意,并有效地抑制谣言传播。在金融方面,利用情感分析可以对金融评论文本的情感进行有效挖掘,从而预测金融的走势。在电商平台上通过分析用户对商品及服务评价的情感倾向,不仅可以了解用户的喜好和感受,而且可以进行有针对性的个性化商品推荐。在旅游评论方面,通过在游客对景点和住宿等方面的诸多评论中进行情感分析,对于制定出行计划和快速了解景区优劣具有较高的参考价值。

本文以现有的信息增益特征选择算法为基础,考虑特征词在每个类别的词频分布及正负情感文本数目,将情感词典和特征选择相结合:首先,将文本与情感词典相匹配,仅对文本中的情感词进行特征选择;其次,对文本中的情感词赋权重后,将文本中所有的词进行特征选择。本文的主要贡献是通过将情感词赋权重,并与改进的信息增益特征选择算法相结合,相比于传统的信息增益特征选择方法,提高了分类准确率,并具

有较稳定的性能。

2 相关工作

情感分析源于 Das 等人^[2]对股票市场留言板中的文本进行分析,并把情感定义为文本中所蕴含的积极和消极意见。文本情感分类属于文本分类的一个特例。目前,文本情感分类研究方法主要有:基于无监督方法、基于有监督方法和基于半监督方法 3 类,而文本情感分类结果通常包括正情感和负情感。典型的无监督学习的文本情感分类方法是由 Turney^[3]提出的 SO-PMI 算法,该方法选择“excellent”和“poor”作为情感基准词表示正面评价和负面评价,并通过计算评价词和基准词之间的距离进行情感分类。常见的基于有监督的文本情感分类方法包括支持向量机(SVM)、朴素贝叶斯(NB)和最大熵等,其分析方法是在一定的先验知识或训练样本集的基础上进行情感知识学习,而后构建情感分类器,以实现目标文本进行情感分类的目的^[4-7]。

近年来,不少学者在上述研究的基础上又提出了许多新的文本情感分类方法。高伟等人^[8]基于半监督学习的情感分类方法,在小规模标注样本的基础上,借助于非标注样本提出了基于一致性标签的集成方法,并提高了情感分类性能。赵传君等人^[9]针对目标领域带标签数据偏少的问题,将半监督学习、AdaBoost、集成学习、BootStrapping 和数据分组等策略与技术相结合,提出了一种基于分组提升集成的跨领域文本情感分类方法,并将该方法在亚马逊购物网站的情感数据集上进行了实验验证。张越兵等人^[10]针对情感分类方法忽略了训练和分类过程的时间代价问题,将粒计算的思想运用于文本数据的三层粒度结构(词一句一篇文章),并提出了文本情感分类特征(sentence-level sentiment strength, SSS)。通过实验对 SSS 特征进行了验证,其结果表明,提出的三支决策方法有效地减少分类过程中所耗费的时

间。于海燕等人^[11]针对传统情感分类算法存在的参数学习困难、分类性能较低的问题,提出了一种基于核超限学习机的中文文本情感分类方法。该方法通过信息增益对训练数据进行特征选择,构建了基于小波核超限学习机的分类器,实现对中文文本的情感分类。朱宪莹等人^[12]针对传统特征选择方法存在的语义缺陷,采用神经网络模型,提出了基于概率模型的词向量权重分配深层特征表示的方法。经过实验验证,采用特征融合的层次结构情感分类方法有效地弥补了传统特征选择方法语义的缺陷,并提高了微博情感分类的准确率。Catal 等人^[13]基于多分类器构建了文本情感分类模型,该模型将投票算法与 Naive Bayes、SVM 和 Bagging 3 个分类器相集合,并对支持向量机作为个体分类器进行了参数优化。经实验验证,多分类器系统能获得较好的情感分类结果。Onan 等人^[14]基于差分进化算法提出了多目标加权投票集成分类器用于文本情感分类。该分类器以提高预测性能为目标,采用优化的加权投票方案对分类器和分类算法的每个输出级分配适当的权重值。杜慧等人^[15]提出了一种基于情感词向量的情感分类方法。该方法能从大规模语料中通过上下文信息挖掘出潜藏的词词语间语义关联,并以此为基础通过对其进行情感调整,得到的情感分类结果准确率更高、稳定性更好。Liu 等人^[16]以改进的一对一策略和 SVM 算法为基础,提出了多情感分类方法。该方法将训练文本转换为特征向量,利用信息增益算法选择多类情感分类的重要特征,并对每对情感类的训练特征向量进行二值 SVM 分类器训练,基于多个 SVM 分类器的结果,构造了多个 SVM 分类器的置信度矩阵。李然等人^[17]从情绪分类的角度整理归纳了文本情绪分析的主流方法,并阐述了文本情绪分析存在的数据稀缺性、类别不平衡、领域依赖性、语言不平衡等问题。Chen 等人^[18]首次提出了基于多标签分类的情感分类方法,并构建了情感分析的原型系统。该原

型系统由文本分割、特征提取和多标签分类 3 个主要部分组成。该原型系统在原分段词和情感特点的基础上,对 3 种不同的情感词典进行了实证研究和分类性能比较。由于文本情感分析问题会涉及文本语料中的多个属性和特性,Zhang 等人^[19]结合了社会领域知识和情绪指标,提出了一种基于因子图算法的情绪分类模型。该方法通过观察带注释的 Twitter 数据集,并归纳出影响用户情感的特征。所提出的算法相比于决策树和 SVM 有更好的分类效果。而针对现有语料库中情感类别不平衡的特征,Li 等人^[20]提出了一种基于最大熵的短文本情绪分布检测方法——MME。该方法利用最大熵原理估计词与社会情绪之间的关系,提高了在新闻、微博等多种规模语料库中的分类精度和预测能力。

3 融合语义情感的特征选择方法研究

由于文本情感分类框架是影响文本情感分类性能的重要因素,也是分类和特征选择的基础。本文给出了如图 1 所示的文本情感分类框架。

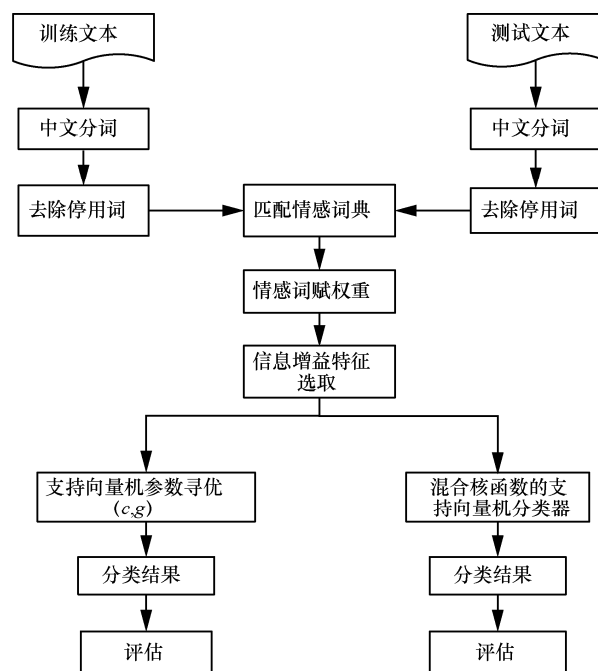


图1 文本情感分类框架



文本情感分类框图实现了对文本进行向量空间表示、特征处理和分类器分类的过程。本文通过各种改进措施,使得文本情感分类效果得到提升。该框架由3部分组成:文本预处理、融合语义情感的信息增益特征选择和优化支持向量机分类器。

其中,融合情感语义的信息增益特征选择部分,针对现有信息增益特征选择算法中的问题,考虑特征词的词频在每个类内的分布程度和整个语料的不平衡因素,分别进行了系数添加,通过结合情感特征,提出了改进的特征选择方法。在经过预处理后,采用SVM对文本进行表示。

由于将文本中的词作为特征,当语料库中的数据量很大时,进行文本表示会导致特征向量维度变大且文本向量存在稀疏性等问题,由此带来的分类器负担会对分类结果产生影响。因此,对向量空间进行有效的特征选择以降低向量空间维度是非常重要的。

3.1 改进的信息增益算法

信息增益是一种基于特征评估方法的信息熵,它表征了文本情感分类系统的信息量,携带的信息量越多,该特征越重要。信息增益如式(1)所示:

$$\begin{aligned} IG(t_j) = & -\sum_{i=1}^n P(c_i) \lg P(c_i) + \\ & P(t_j) \sum_{i=1}^n P(c_i|t_j) \lg P(c_i|t_j) + \\ & P(\bar{t}_j) \sum_{i=1}^n P(c_i|\bar{t}_j) \lg P(c_i|\bar{t}_j) \end{aligned} \quad (1)$$

其中, n 为评论文本类别总数; $P(c_i)$ 为类 c_i 的概率; $P(t_j)$ 为特征 t_j 在文本集中出现的概率; $P(\bar{t}_j)$ 为特征 t_j 不在文本集中出现的概率; $P(c_i|t_j)$ 为出现特征 t_j 的文本属于类 c_i 的概率; $P(c_i|\bar{t}_j)$ 为不出现特征 t_j 的文本属于类 c_i 的概率。

由式(1)可知,若过多地关注特征词出现的频率而忽视特征词的词频,会降低特征预期的功

能,因此,原始的信息增益方法仅仅考虑了特征词和类别之间的相关性,而忽略了类内的分布程度和语料的均衡问题。因此,本文对信息增益算法进行了如下改进。

(1) 类内词频分布

若某个特征词在多个类别中反复出现,则该特征分类能力弱,应赋予较小的权重;若某个特征词在一个类别中多次出现,且均匀分布在该类别的每个文本中,则该特征分类能力强,应赋予较大的权重。因此,类内词频分布如式(2)、式(3)所示:

$$\alpha_i = 1 - \frac{\alpha_j}{\sqrt{\sum_{j=1}^m \alpha_j^2}} \quad (2)$$

$$\alpha_{ij} = \sqrt{\frac{1}{N_i - 1} \sum_{k=1}^{N_i} [\text{tf}_{ik}(t_j) - \frac{1}{N_i - 1} \sum_{k=1}^{N_i} \text{tf}_{ik}(t_j)]^2} \quad (3)$$

其中,特征集 $T = \{t_1, t_2, \dots, t_m\}$, $d_{ik} (1 \leq k \leq N_i)$ 为类别 C_i 中的文本, $\text{tf}_{ik}(t_j)$ 代表特征 t_j 出现在类别 c_i 的文本 d_{ik} 中出现的频数, α_i 为归一化值,特征项在某一类别中的每个文本中出现的频数越均匀,则权重值 β_i 越大,对应的分类能力越强。

(2) 不平衡数据集

本文选择了旅游评论数据集,由于该数据集中负评论相对较少,经人工筛选后,选择了1 000条正评论、700条负评论作为测试样本。由于旅游数据集两个类别中的文本数目会存在不平衡现象,应用传统信息增益进行计算,会导致某一特征词在具有少量文本的类别中信息增益值过小,而该特征词对分类贡献却较大的情况,会对最终分类结果造成影响。因此,本文在传统信息增益公式的基础上,首先依据每个类别的文本数,计算该类别在文本集中出现的概率;然后用特征词在每个类别中出现的概率除以该类别出现的概率以改善不平衡数据集对分类结果的影响;最后,对文本数量较少的类别中出现的特征词的信息增益进行相应调整,以达到提高分类效果的目的。综

合考虑上述两个因素, 并将其融入信息增益的计算过程中, 则信息增益如式 (4) 所示:

$$\begin{aligned} IG(t_j) = & -\sum_{i=1}^n P(c_i) \lg P(c_i) + \\ & P(t_j) \sum_{i=1}^n P(c_i|t_j) \lg \frac{P(c_i|t_j)}{P(c_i)} + \\ & P(\bar{t}_j) \sum_{i=1}^n P(c_i|\bar{t}_j) \lg \frac{P(c_i|\bar{t}_j)}{P(c_i)} \end{aligned} \quad (4)$$

3.2 融合语义情感的文本情感分类特征选择

3.2.1 融合语义情感的特征选择方法处理流程

信息增益特征选择算法是通过特征词在文本中出现的次数等信息来分析该特征词对文本情感分类的作用。因此, 选取对文本分类贡献大的特征词既减少了无用信息、降低了文本空间向量维度, 又可以消除部分噪音, 提高分类准确率。传统的特征选择方法是对文本进行无关语义分析, 仅通过计算增益值来判断每个特征词对分类的贡献。而本文考虑特征词的语义情感信息, 利用情感词典和情感词赋权重的方式, 将分词后的词语与情感词典相匹配, 将每个文本中与情感词典匹配的情感词作为初始词集, 利用信息增益计算, 并进行特征选择。以此为基础, 选取增益值较高的 k 个词作为特征词集, 进行向量空间表示后作为分类器的输入。融合语义情感的特征选择方法流程如图 2 所示。

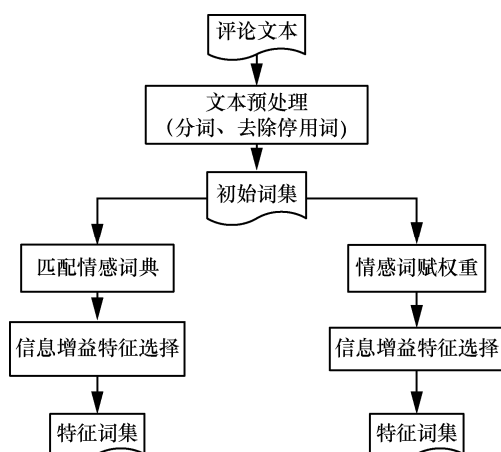


图 2 融合语义情感的特征选择方法处理流程

3.2.2 结合情感词典的特征选择方法

情感词典是包含了正、负面情感词的词语集合。目前常用的情感词典有知网情感词典、NTUSD 中文情感词典以及大连理工大学的中文情感词汇本体库。由于在文本情感分类的过程中, 情感词语本身对分类有一定的贡献, 对于评论文本而言, 语句会包含情感词进行情感的表达。若情感词出现的次数较少, 则其他对景点或者景区的描述相对较多。由于情感词是表达语句情感的重要因素, 选择只用情感词作为初始词集, 其维度比原始词语的维度小得多, 很大程度地降低了维度。因此, 以改进的信息增益公式计算过程为基础, 对所有的情感词语进行信息增益计算, 选择增益值较大的情感词作为特征词集, 如式 (5) 所示:

$$\begin{cases} IG(t_j) = -\sum_{i=1}^n P(c_i) \lg P(c_i) + P(t_j) \sum_{i=1}^n \alpha_i P(c_i|t_j) \lg \frac{P(c_i|t_j)}{P(c_i)} \\ \quad + P(\bar{t}_j) \sum_{i=1}^n P(c_i|\bar{t}_j) \lg \frac{P(c_i|\bar{t}_j)}{P(c_i)}, & t_j \text{ 为情感词} \\ IG(t_j) = 0, & t_j \text{ 为非情感词} \end{cases} \quad (5)$$

如式(5)所示, 当 t_j 为情感词时, 进行信息增益计算; 当 t_j 为非情感词时, 信息增益值记为 0, 即不进行增益值计算。

本文选用台湾大学的 NTUSD 中文情感词典进行文本情感词匹配, 该词典包含 11 086 个词语, 其中 2 810 个积极属性词语即正面情感词语, 8 276 个消极属性词语即负面情感词语, 将正、负面情感词典合为一个词表, 称为情感词表, 应用包含两种情感极性的词表去匹配文本, 从文本中找到情感词。实现方法为, 在文本中识别情感词, 选用增益值较高的情感词作为特征词集, 进行向量空间表示之后分类。本文匹配情感词典的目的是将情感词在文本相应位置留下, 将非情感词删除, 并输出只包含情感词的文本。结合情感词典的 IGSC_Pre 算法描述如下。

输出：情感词匹配后的文本 SText。

{bool key; //key 为 map 容器中的情感词, 若情感词, 返回值为 True, 否则为 False;

逐行读取情感词和句子到字符数组中:

将 Text 转换成字符串存入 vector 容器:

Calculating by formula (5);

key=true:

else

end if }

}

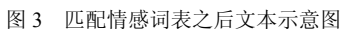
end

利用上述算法对任意 Text 文档分词后进行情感词匹配的测试结果如图 3 所示。从图 3 可知, 经情感词匹配后的输出文本, 每条评论文本仅由情感词进行表示, 大幅度地降低了初始特征向量的维度。

在情感分类中仅通过信息增益计算,会造成较低的信息增益值在特征选择时被筛选掉,影响到分类的准确率。因此本文采用为情感词赋予权重的方式,选出增益值较高的特征词。其目的是可以保证对分类意义较大的情感词和其他词语都被有效地选择出来。特征选择方法如式(6)所示。

当 t_j 为情感词时,进行赋权重计算, β 即权重值, β 的取值通过多次实验测定。用结合赋权重的方法改进信息增益的计算过程,当 t_j 为非情感词时,采用式(4)进行计算。

综上所述,通过改进的信息增益计算方法,有效地结合情感词典和为情感词赋权重的方式,本文提出的基于情感词典的特征选择方法进行文本分类算法 IGSC 描述如下。



$$\begin{cases}
 IG(t_j) = \\
 -\sum_{i=1}^n P(c_i) \lg P(c_i) + P(t_j) \sum_{i=1}^n \alpha_{i1} P(c_i|t_j) \lg \frac{P(c_i|t_j)}{P(c_i)} + \\
 P(\bar{t}_j) \sum_{i=1}^n P(c_i|\bar{t}_j) \lg \frac{P(c_i|\bar{t}_j)}{P(c_i)}, t_j \text{ 为情感词} \\
 IG(t_j) = \\
 -\sum_{i=1}^n P(c_i) \lg P(c_i) + P(t_j) \sum_{i=1}^n P(c_i|t_j) \lg P(c_i|t_j) + \\
 P(\bar{t}_j) \sum_{i=1}^n P(c_i|\bar{t}_j) \lg P(c_i|\bar{t}_j), t_j \text{ 为非情感词}
 \end{cases} \quad (6)$$

$$\alpha_{ij} = \sqrt{\frac{1}{N_i - 1} \sum_{k=1}^{N_i} [\beta t_{f_{ik}}(t_j) - \frac{1}{N_i - 1} \sum_{k=1}^{N_i} \beta t_{f_{ik}}(t_j)]^2} \quad (7)$$

输入：词—词频矩阵文件 w-wfext，情感词—词频矩阵文件 s-wfext；

输出：文本—赋权重词频文件 t-Wwfext；

begin

{ define key, value; // key 和 value 分别为 map 容器中的词和词频；

open the files w-wfext and s-wfext;

while (files != EOF) {

将 w-wfext 文件中的数据逐行读取到字符数组 A 中；

将 s-wfext 文件中的数据逐行读取到字符数组 B 中；

将词、词频、情感词词频分别存入 map、vector1 和 vector2 容器中；

将 map 容器中的 key 与 vector2 容器中的情感词进行匹配；

遍历 map 容器中匹配到的 key 值词频 value，赋权重并将结果存入 vector3 容器中；

将 map 容器中的 key 重新映射到 vector3 容器中，并更新 value; }

output Map container;

output t-Wwfext file;

}

end

4 实验与结果分析

4.1 实验环境设置

本文采用 MATLAB 语言对文本—词频多维矩阵进行向量空间计算、信息增益计算，选取特征并表示成分类器所需的向量空间模型；选择支持向量机分类器；应用 SVM 模式包 LIBSVM，将 LIBSVM 配置到 MATLAB 软件中进行编译；选取默认的支持向量机参数 C 和径向基核函数参数 g 。将训练集输入特征向量表示样本设为 train，情感词极性 train_label 表示为 1 或者 -1，测试集输入特征向量表示样本设为 test，情感词极性 test_label 表示为 1 或者 -1，用 svmtrain 进行训练，svmpredict 为预测。

4.2 算法对比与实验分析

(1) 改进信息增益算法与原算法实验对比

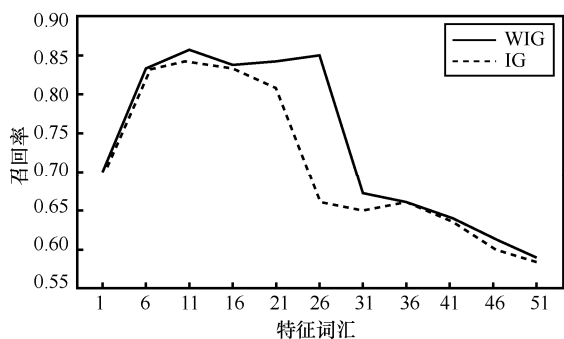
将考虑词频分布信息和不平衡数据集因素的改进增益算法与原始的信息增益算法进行对比，将准确率、召回率和 F 值作为情感分类评估标准参数。

进行信息增益计算后，将信息增益值进行降序排列。实验中特征词的数量范围为 1~51，步长为 5，因此共 11 个特征词数取值。图 4 描述了改进信息增益算法 WIG 与原始算法 IG 在分类准确率、召回率和 F 值上的实验对比。

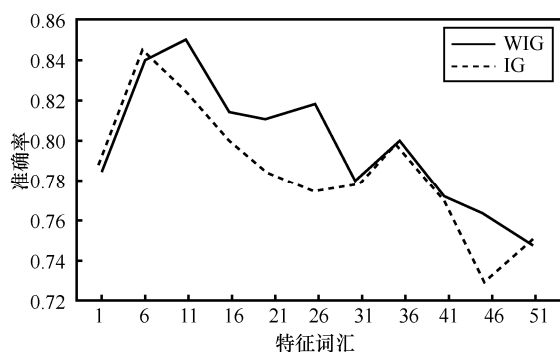
从图 4 可知，改进后的算法 WIG 在分类准确率、召回率和 F 值方面均得到了一定的提升且优于原始的 IG 算法。因此，增加了分布因子和权衡概率的改进算法，在特征词数选取较少时，有较好的分类性能。其原因是由于特征词数选取得越少，空间向量维度越小，在减少空间复杂度和时间复杂度的同时，提高了分类的精度。

(2) 结合情感词典的特征选择方法实验对比

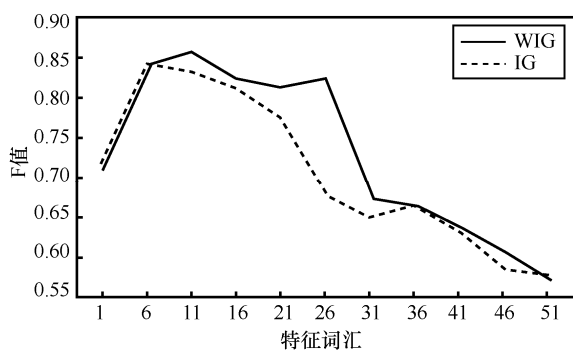
将结合情感词典的信息增益特征选择方法 IGSC_Pre 与原本信息增益算法 IG 进行比较，实验中特征词的数量范围为 1~101，步长为 5，因此共 21 个特征词数取值。图 5 描述了改进信息增益



(a) 分类准确率

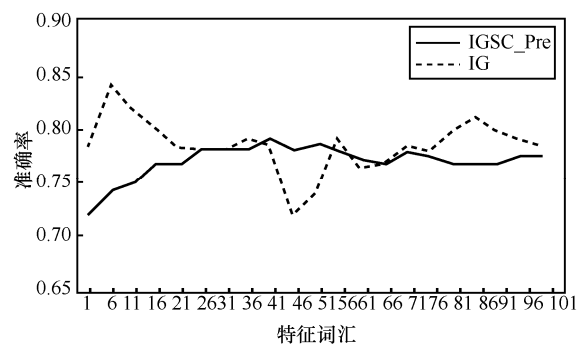


(b) 分类召回率

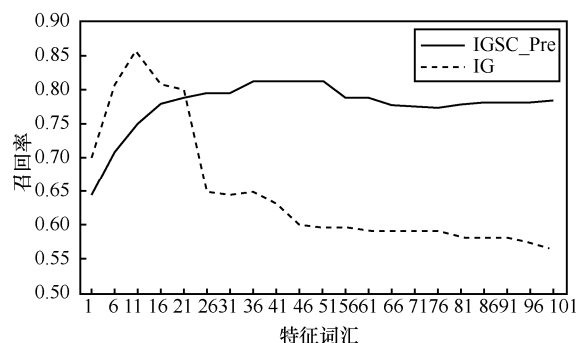


(c) 分类F值

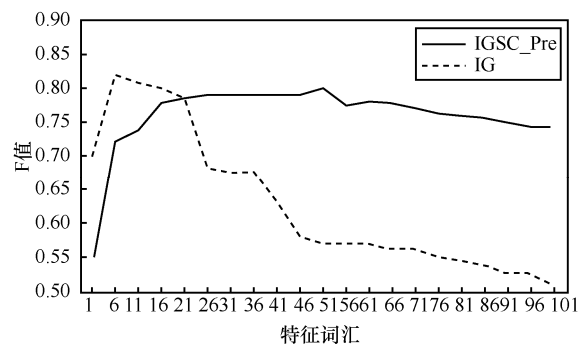
图4 改进信息增益算法WIG与原始算法实验对比



(a) 分类准确率



(b) 分类召回率



(c) 分类F值

图5 改进信息增益算法IGSC-Pre与原始算法IG实验对比

算法 IGSC_Pre 与原始算法 IG 在分类准确率、召回率和 F 值上的实验对比。

从图 5 可知, IGSC_Pre 算法与 IG 算法相比, 在获得了较高的分类准确率、召回率和 F 值的同时, 整体的分类性能也较为稳定。本文选择了旅游评论文本数据集, 文本语料长短不一, 有些评论文本仅由两个词语组成, 会导致数据过于稀疏的问题。由于在进行情感词匹配后, 整个文本可能会被剔除掉而产生空文本。而后, 在分类器进行学习时会产生误差, 进而不能够准确判断

该文本属于正情感还是负情感, 并使分类性能变差。因此, 改进的算法 IGSC_Pre 会产生部分实验结果在分类准确率、召回率、F 值在峰值略低于原始的特征选择方法。

由于进行特征选择时, 特征数目 k 不确定, 需要依据实验来确定, 若想获得较高的峰值, 需要不断地进行参数调整, 因此, 在较大范围内保持较高的分类性能是非常重要的。而本文提出的结合情感词典的改进算法在分类性能的稳定性方面具有较好的优势。

(3) 结合情感词赋权重的特征选择方法实验对比

针对语料数据稀疏的问题,本文对所有词语进行增益计算后再进行特征选择,将情感词部分识别出来,并对其赋权重,进而增加了情感词被留下作为最终特征的概率。因此,本文在上述研究内容的基础上,提出了 IGSC 算法。并将结合情感赋权重的信息增益特征选择方法 IGSC 与原始信息增益算法 IG 进行比较,实验中特征词的数量范围为 1~101,步长为 5,共有 21 个特征词数取值。图 6 描述了改进信息增益算法 IGSC 与原始算法 IG 在分类准确率、召回率和 F 值上的实验对比。

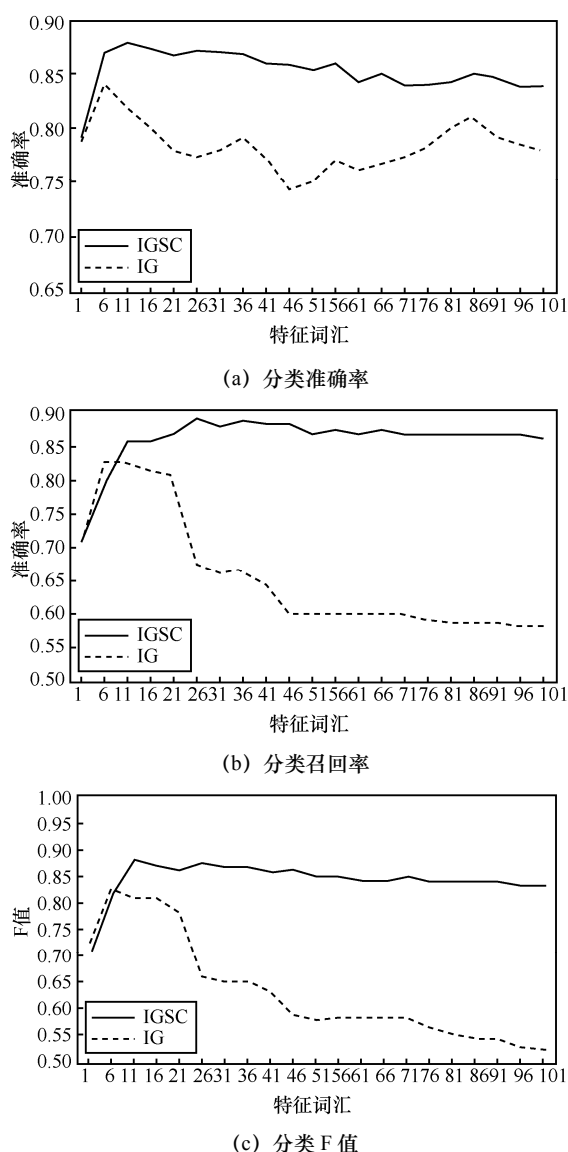


图 6 改进信息增益算法 IGSC 与原始算法 IG 实验对比

从图 6 可知, IGSC 算法在特征词数选取范围内的分类准确率都高于 IG 算法, 且 IGSC 算法分类精度较 IG 算法有很大的提高且分类准确率趋势更稳定; 在分类召回率方面, 特征维度为 21 之后, IG 算法性能骤然下降, 而 IGSC 算法却有明显提升, 两者召回率最多相差 30%。综上可知, 将情感词赋权重的特征选择算法 IGSC 与原始信息增益特征选择算法 IG 相比, 其分类准确率、召回率和 F 值 3 个方面都得到了较大的提升, 分类效果更优。

为了验证 IGSC 算法与同类算法在分类精度方面的性能, 本文将 IGSC 算法与参考文献[11]中所提出的 KELM 算法进行实验对比。由于 KELM 算法采用了 BOOK 和 NOTEBOOK 两个领域的语料集, 两个领域的语料集属于较为相近的应用领域, 为此, 本文仅对 BOOK 语料集中的数据评论信息进行测试。该测试集来源于谭松波博士搜集整理的中文文本语料, 实验中的数据选用正负向文本各 1 000 篇。对 IGSC 算法、KELM 算法以及参考文献[11]在进行算法比较时使用的 SVM 和 NB 算法进行实验对比, 实验中仍选取特征词的数量范围为 1~101, 步长为 5, 共有 21 个特征词数取值, 实验结果分别如图 7 所示。

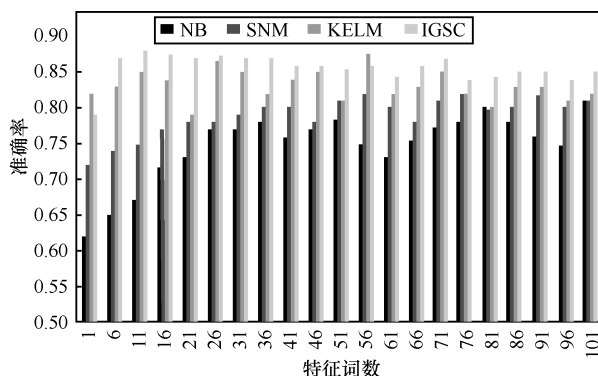


图 7 分类精度比较结果

从图 7 中可以看出, 本文所提出的 IGSC 算法在分类精度方面总体优势优于其他 3 类算法, 且分类精度的结果较为稳定。其原因是本文所提



出的算法考虑到了类内的分布因子和均衡概率等信息, 结合了情感词对文本分类的影响, 并采取了为情感词赋权重的方式, 使得分类结果更加准确, 且达到了较好的效果。

5 结束语

本文分析了原始的信息增益算法 IG 只注重特征词出现的文档频率, 而忽视了类内词频位置分布和不平衡数据集等方面对特征选择的影响问题, 通过增加类内词频分布因子和均衡概率的方法对其进行改进; 而后, 针对目前信息增益特征选择在文本情感分类中忽略了情感因素, 将情感词典和特征选择进行结合, 实现了对文本中的情感词进行特征选择和对情感词赋权重的两种改进特征选择方法选择旅游评论文本数据集进行测试, 其结果表明, 改进的文本情感分类特征选择方法在分类准确率、召回率和 F 值方面均得到了提升, 并且达到了很好的分类稳定性, 有效地验证了改进的文本情感分类特征选择方法的可行性。对于不同的语料及应用, 权重值的设定存在不确定性, 因此, 研究如何根据分类准确率自适应地调整情感词权重, 并最终提高分类准确率是本文下一步的工作。

参考文献:

- [1] 马晓玲, 金碧漪, 范并思. 中文文本情感倾向分析研究[J]. 情报资料工作, 2013(1): 52-56.
MA X L, JIN B Y, FAN B S. An analysis of Chinese text emotional tendency[J]. Information and Documentation Services, 2013(1): 52-56.
- [2] DAS S, CHEN M. Yahoo! for amazon: extracting market sentiment from stock message boards[R]. 2001.
- [3] TURNEY P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]// Meeting on Association for Computational Linguistics, July 7-12, Philadelphia, Pennsylvania. New York: ACM Press, 2002: 417-424.
- [4] MORAES R, VALIATI J F, NETO W P G, et al. Document-level sentiment classification: an empirical comparison between SVM and ANN[J]. Expert Systems with Applications, 2013, 40(2): 621-633.
- [5] HADDI E, LIU X H, SHI Y. The role of text pre-processing in sentiment analysis[J]. Procedia Computer Science, 2013(17): 26-32.
- [6] KANG H, YOO S J, HAN D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews[J]. Expert Systems with Applications, 2012, 39(5): 6000-6010.
- [7] YU L C, WU J L, CHANG P C. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news[J]. Knowledge-Based Systems, 2013(41): 89-97.
- [8] 高伟, 王中卿, 李寿山. 基于集成学习的半监督情感分类方法研究[J]. 中文信息学报, 2013, 27(3): 120-126.
GAO W, WANG Z Q, LI S S. Semi-supervised sentiment classification with a ensemble strategy[J]. Journal of Chinese Information Processing, 2013, 27(3): 120-126.
- [9] 赵传君, 王素格, 李德玉. 基于分组提升集成的跨领域文本情感分类[J]. 计算机研究与发展, 2015, 52(3): 629-638.
ZHAO C J, WANG S G, LI D Y. Cross-domain text sentiment classification based on grouping-Adaboost ensemble[J]. Journal of Computer Research and Development, 2015, 52(3): 629-638.
- [10] 张越兵, 苗夺谦, 张志飞. 基于三支决策的多粒度文本情感分类模型[J]. 计算机科学, 2017, 44(12): 188-193, 215.
ZHANG Y B, MIAO D Q, ZHANG Z F. Multi-granularity text sentiment classification model based on three decision making[J]. Computer Science, 2017, 44(12): 188-193, 215.
- [11] 于海燕, 陈丽如, 郑文斌. 基于核超限学习机的中文文本情感分类[J]. 中国计量学院学报, 2016, 27(2): 228-233.
YU H Y, CHEN L R, ZHENG W B. Chinese text sentiment classification based on kernel extreme learning machines[J]. Journal of China Jiliang University, 2016, 27(2): 228-233.
- [12] 朱宪莹, 刘箴, 金炜, 等. 基于特征融合的层次结构微博情感分类[J]. 电信科学, 2016, 32(7): 106-114.
ZHU X Y, LIU Z, JIN W, et al. Hierarchical micro-blog sentiment classification based on feature fusion[J]. Telecommunications Science, 2016, 32(7): 106-114.
- [13] CATAL C, NANGIR M. A sentiment classification model based on multiple classifiers[J]. Applied Soft Computing, 2017(50): 135-141.
- [14] ONAN A, KORUKOĞLU S, BULUT H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification[M]. Oxford: Pergamon Press, Inc. 2016.
- [15] 杜慧, 徐学可, 伍大勇. 基于情感词向量的微博情感分类[J]. 中文信息学报, 2017, 31(3): 170-176.
DU H, XU X K, WU D Y. A sentiment classification method based on sentiment-specific word embedding[J]. Journal of Chinese Information Processing, 2017, 31(3): 170-176.
- [16] LIU Y, BI J W, FAN Z P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm[J]. Information Sciences, 2017(394-395): 38-52.
- [17] 李然, 林政, 林海伦. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55(1): 30-52.
LI R, LIN Z, LIN H L. Summary of text sentiment analysis[J]. Journal of Computer Research and Development, 2018, 55(1): 30-52.
- [18] LIU S M, CHEN J H. A multi-label classification based approach for sentiment classification[J]. Expert Systems with Applications, 2015, 42(3): 1083-1093.
- [19] ZHANG X, LI W, LU S. Emotion detection in online social net-

work based on multi-label learning[C]//22nd International Conference on Database Systems for Advanced Applications, March 27-30, 2017, Suzhou, China. Berlin: Springer, 2017: 659-674.

- [20] LI J, RAO Y, JIN F, et al. Multi-label maximum entropy model for social emotion classification over short text[J]. Neurocomputing, 2016(210): 247-256.

[作者简介]



刘洛辛（1976-），男，博士，燕山大学信息科学与工程学院教授，河北省信息传输与信号处理重点实验室硕士生导师，主要研究方向为物联网、无线传感器网络。



陈晶（1976-），女，博士，燕山大学信息科学与工程学院副教授，河北省计算机虚拟技术实验室硕士生导师，主要研究方向为 Web 服务、社交网络情感分析。



王麒媛（1991-），女，燕山大学信息科学与工程学院硕士生，主要研究方向为无线传感器网络、情感分析。