文章编号：2096-1618(2019)05-0501-05

结合关联规则的情感分析模型研究

杨 頔， 文成玉
(成都信息工程大学通信工程学院，四川 成都 610225)

摘要：针对网络中各类格式不规范的评论，提出了一种结合关联规则的情感分析模型，提高网络评论的情感分析精度。即在分析评论的情感倾向基础上，深入挖掘评论对象的所有关联特征，充分考虑文本中每个词对总体情感倾向的影响。该模型首先选择一类有明确评论对象的数据集，采用传统的特征提取方式获得候选特征，再用关联规则挖掘出评论对象的关联特征，最后采用SVM和NB算法进行情感分类。通过实验测试，并对结果对比分析可知：该模型可以有效地提高网络评论情绪分类的准确率、召回率、F值。

关 键 词：情感分析；数据挖掘；关联规则；SVM；NB
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2019.05.011

0 引言

随着互联网的不断发展，越来越多的人热衷于在网络上发表言论，针对一个或者多个评论对象展开讨论。这样产生的网络评论直接反映了用户对人物、事件、产品等的客观看法，同时包含了极大的信息量和潜在价值。因此网络评论是具有一定参考价值的反馈信息，通过捕捉用户在评论中体现的潜在需求，就能够实现信息的预测、辅助决策；挖掘评论中用户的意见，就能实时关注评论的风向并实现舆情的监控[1]。最常见的就是商家根据用户的评论对产品进行改进，并向用户推荐相关商品，这些都能带来巨大的商业价值。

那么通过对网络评论进行情感分析，就能够发掘评论背后的价值。情感分析不仅能判断出文本的情感极性，还能获得更多主观信息。不过，网络评论具有数据量大、内容复杂，而文本句式简短、格式不太规范等特点，传统文本情感分析算法并不适用。周咏梅等[2]提出一种基于中文情感词典的方法，利用词典将句子中的词语进行自动分解，再计算各自的情感色彩强度。但对于网络这样复杂的语言环境，很多词语往往存在变化的情感色彩，同一个词可作为多种词性使用，不能一概而论。

随着分析精度等要求的不断增高，情感分析算法从早期的情感词典过渡到了时下热门的机器学习。赵刚等[3]在传统情感词典方法的基础上，进行了词典的扩充，最后运用机器学习方法设计了一个网络评论情感分析模型。这样的前期准备工作较为复杂，包括收集若干个情感词典进行整合去重等，最终才能保证较高的准确率。周杰等[4]提出了一个评论数据情感分析模型，虽然应用范围较小，但选取了不同的特征数据集等进行实验，得到了情感词和相关词语共同作为特征有助于情绪分类的结论。姜杰等[5]提出结合机器学习与规则的情绪分析算法，即对两种算法取长补短来提升情绪分类的效果，但仍有一定的改进范围。本文同样选择规则和机器学习算法结合的方法，采用的关联规则算法能够挖掘出情感词与其他论据词语的所有搭配，并筛选出关联性强的搭配作为评论特征。这样选出来的特征能为机器学习提供准确的分类信息，达到更佳的实验效果。

1 基于关联特征的情感分析

人脑之所以能准确地判断句子的类型，是因为关注到了句子中每一个词语，从而能整体地认识句子的情感。因此，文本情感分析要尽可能地做到对人脑思维的模拟，清晰地掌握特定语境下词语对象之间的联系。文中主要通过建立规则去联系这些词语，以实现复杂的思维网络，并采用机器学习的方法有监督地去学习，进而达到高分类精度的要求。以此构造的情绪分类模型见图1。

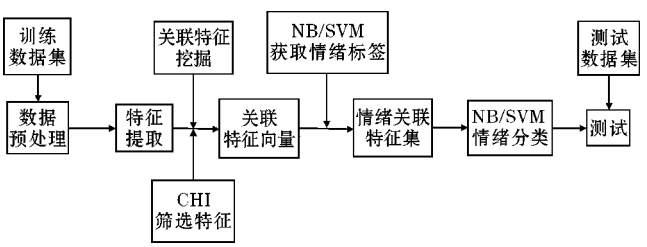

图1 基于关联特征的情绪分类模型

收稿日期：2019-04-09

通过对数据集提取到的特征再挖掘关联特征,能够捕捉到某一语境下词语之间的联系,再用机器学习算法为评论特征添加情绪标签,扩展了基本的特征模板,形成一个情绪关联特征集,最后再依据该特征集实现情绪分类。

1.1 文本预处理和特征选择

首先要对句子中的词语进行分割操作,就要选择适用于汉语分词的工具。例如,英文文本的句子间有空格间隔,处理起来就很简单,但中文词与词之间的间隔没有那么明显。如果是人工处理的话,很容易就可以看出词与词之间的边界。为了让计算机正确拆分句子,本文使用 Python 中文分词工具“jieba”。“jieba”分词自带一个包含词语出现次数的词典,通过查询词典可以列出句子所有的切分方式,再基于词频找出最有可能的切分方式[6]。在分词过程中,相邻句子间的关联性词都会保留下来,并且“jieba”分词支持繁体分词,十分适合语言形式复杂的网络环境。

分词处理后会产生大量的单个词语,如果直接将这些词语作为分类的特征,就会出现计算量超负荷的现象,因此要进行特征的降维,即对这些词语进行筛选。然而,很多特征提取方法采用了独立性假设,不太符合语言文字实际的分布情况。这些方法往往忽略了单词间的语法和顺序,无法了解单词间的关联程度。例如,文献[6]采用基于词频的 TF-IDF 方法,选择在某一个句子中出现次数多,而在其它句子出现次数相对少的词作为特征,并赋予高权重。但单纯以词频作为单词重要性的度量,对于词语所处的位置没有敏感性,就会造成特征的错误选择。本文选择基于卡方统计量(CHI)的方法,该方法根据各个词语与目标特征的卡方值大小来筛选特征,比较适合评论文本的特征选择。因为卡方值的大小体现了词语与特征的相关性大小,卡方值越大则代表该特征比较重要。

1.2 关联特征挖掘

采用关联规则来挖掘评论数据之间的关联性,用规则的形式表示对象的某些特征一起出现的规律或模式。由于考虑到了评论文本中各类词语的搭配,就能抓住情绪词、情绪词和评论对象之间的关联性。首先,关联规则将数据集中每一条记录定义为事务,记录中的每一个项定义为项目,那么项目的集合就称为项集。

那么规则就可表示为项集→项集,描述规则的参数有支持度(support),定义为

$$\sup(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{|T|} \tag{1}$$

参数置信度(confidence)定义为

$$\mathrm{conf}(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{\sigma(X)} \tag{2}$$

其中,$\sigma(X\cup Y)$为数据库中项集 X、Y 同时出现的事务频数,T 为全部事务的总和,$\sigma(X)$为项集 X 出现的事务频数。

为实现关联规则挖掘,需要选出频繁出现的项集。符合规定的最小支持度条件的项集就是频繁项集。再依据这些项集产生关联规则,并依照给定的最小置信度筛选出强关联规则[8]。实现该过程的算法主要有 Apriori 和 FP-growth,前者要求不断扫描数据库以生成大量候选集,耗时较长,不适合网络评论这种量级的数据。后者仅要求扫描两次,其最大优点是不用生成候选集合,平均效率远高于 Apriori。FP-growth 通过形成一个 FP-tree 来存放扫描记录。为了构建 FP-tree,首先要对数据库扫描两次,第一次扫描数据库每个元素项出现的次数并进行计数,第二次扫描则只关注那些频繁出现的元素。

第一步,扫描评论数据库的记录,生成一级频繁项集,并列出库中包含该项的记录数,如表 1 所示。

表 1 评论数据库记录

特征 ID	元素项	包含该项的记录数
C1	配置/技术/性能/产品/设备/东西…	2610
C2	价格/定价/售价…	1730
C3	贵/昂贵/鸡肋/渣/无价值/麻烦/无意义…	1506
C4	支持/期待/喜欢…	2341
C5	创新/升级/智能/科技…	1275

第二步,再次扫描,对于出现在表 1 中的每一项元素,依照出现次序组合并排序,作为候选特征组合,如表 2 所示。

表 2 评论文本候选特征组合

元素项集	候选特征组合
【价格性能麻烦】	[C1,C2,C3]
【升级喜欢设备价格】	[C1,C2,C4,C5]
【售价昂贵】	[C1,C2,C4,C5]
【产品定价创新】	[C1,C2,C5]
【东西鸡肋】	[C1,C3]
【价格贵】	[C2,C3]
【售价期待产品】	[C1,C2,C4]
【配置渣】	[C1,C3]
【定价支持】	[C2,C4]

扫描生成 FP-tree 过程如图 2 所示,首先选择“NULL”作为根节点。表 2 中第一条记录[C1,C2,C3]对应 FP-tree 中的第一条分支{[C1:1],[C2:1],[C3:1]},第二条记录[C1,C2,C4,C5] 与前一条记录有相同的前缀[C1,C2],因此[C1]、[C2]的支持度加 1,并在[C1,C2]节点下添加节点[C4:1]、[C5:1],第三条记录[C2,C3]作为 FP-tree 的一个分支,更新相关节点的支持度。

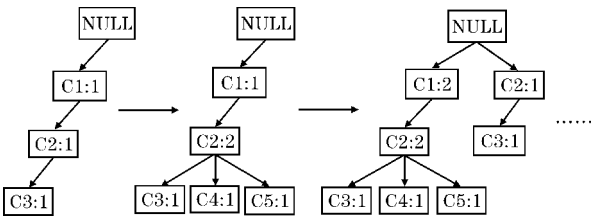

图 2　扫描过程

扫描完记录之后,就会形成最终的 FP-tree。扫描到第九条记录形成的 FP-tree 如图 3 所示。

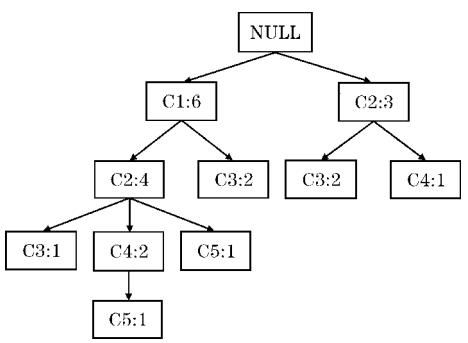

图 3　扫描至第九条记录所生成的 FP-tree

从树的底部开始挖掘频繁项集,在树中以 C5 为尾节点的链共有两条,分别是{[C1:6],[C2:4],[C4:2],[C5:1]}和{[C1:6],[C2:4],[C5:1]}。这两条链表示[C1,C2,C4,C5]、[C1,C2,C5]在数据库中分别出现了 1 次,所以将{[C1:1],[C2:1],[C4:1],[C5:1]}和{[C1:1],[C2:1],[C5:1]}称为 C5 的条件模式基。将 C5 的条件模式基作为新的事务数据库,计算每一行记录中各种物品的支持度,设定最小支持度来选出频繁项集。而 C3 与 C5 同时出现的次数较少,C3 就会因为支持度较低而被去掉。最终构建一棵新的树如图 4 所示,这棵树被称为 C5 的条件 FP-tree。

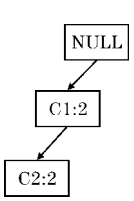

图 4　C5 的条件 FP-tree

从图 4 可以看出,C5 的条件 FP-tree 中满足支持度阈值的节点剩下 2 个,以 C5 为尾节点的频繁项集有[C5:2]、[C2,C5:2]、[C1,C5:2]、[C2,C1,C5:2]。同样地,以 C4 为尾节点的频繁项集有[C4:3]、[C1,C4:2]、[C2,C4:3]、[C1,C2,C4:2],C3 为尾节点的频繁项集有[C3:3]、[C1,C3:3]、[C2,C3:3]、[C1,C2,C3:1],C2 为尾节点的频繁项集有[C2:7]、[C1,C2:4],C1 为尾节点的频繁项集有[C1:6]。最后,规定合适的最小置信度,选择符合要求的项集作为挖掘到的强关联规则。

1.3　情绪分类算法

对于上一节挖掘出的强关联规则中的数据采用机器学习算法分类,以训练评论数据库中每条文本的情绪标签为标准,给每条关联规则添加情绪标签。最后,根据带有情感标签的关联规则去对测试评论数据实现情感分类。本文研究的问题本质是一个情感的正负性分类问题,训练数据中包括评论文本和文本的情绪标签 pos(积极)或者 neg(消极),采用 NB 和 SVM 两种算法训练数据并获得测试结果。

NB 算法主要利用概率论方法,两个事件同时发生的概率为一个事件发生的条件概率与另一个事件发生概率的乘积。即公式 $P(x,y)=P(x\mid y)P(y)$,其中 x 代表数据库中的特征集,y 代表数据库中的分类标签集。采用 Bayes 定理就能得到 $P(y\mid x)$,根据数学公式 $P(y\mid x)=\frac{P(x\mid y)\times P(y)}{P(x)}$ 进行计算[9]。当满足后验概率为最大的时候,对应的分类标签就是特征所属文本的分类标签。

SVM 算法首先要将每条包含几类特征的待分类文本映射为三维空间中的一个点,然后在这个三维空间产生一个超平面去分隔这些点。参考人工标注的分类结果,SVM 会自动调整超平面的位置实现样本点的分隔,此时超平面的位置会达到一个最优的状态[10]。所谓的最优状态即存在与超平面距离最近的两个样本点,这两个点属于不同的两个类别,且它们之间的距离 d 为最大值。其中,d 根据平面表达式 $w\cdot x+b=0$ 表示为 $\frac{2}{\|\omega\|}$。总之,SVM 算法的核心就是形成一个满足 d 为最大值的超平面,此时超平面的分类效果最佳。

2　实验与结果分析

2.1　实验评论数据

实验评论数据源自用户对某国外品牌电子产品的

新品发布的评论集。通过对用户评论的情感倾向分析,可以得出用户群体对这款新产品的支持程度,以此可推测出用户购买产品的可能性。并且,这些评论发布在各种社交平台上,就会对看到这些评论的人造成一定的影响。通过爬虫代码在各大社交网站上爬取针对该新产品的评论数据,总数据量达 8532 条,选择 7000 条作为训练集,余下 1532 条为测试集。数据集先用“jieba”实现分词,再对比相关词典去除一些无含义的词,剩下的词全部作为候选特征。

2.2 情感关联特征提取与分类

情绪分类的效果很大程度取决于提取特征的有效性,为了体现关联特征对情绪分类效果的影响,采用 3 种特征挖掘方法:

(1)基于卡方校验的 CHI 方法,假设候选特征与分类标签“pos”和“neg”相互独立,计算候选特征与两类标签的卡方值。在一定误差允许的范围内,卡方值越大的候选特征与分类标签的关联度越高,设定一定的最小卡方值阈值筛选出满足条件的特征。

(2)基于关联规则中的 FP-growth 方法,用候选特征构成基本事务集,构建关于基本事务集的 FP-tree 来发现频繁项集,找出强关联的特征。

(3)CHI 与 FP-growth 结合的方法,将两种方法挖掘出的特征合并为一个特征项集,关联特征为主要特征,CHI 特征为补充特征。

分类方法采用朴素贝叶斯和支持向量机,分别采用这两种方法为以上 3 种特征添加情绪标签,形成不同的情绪关联特征集再进行情绪分类。

2.3 实验结果分析

情绪分类实验效果以准确率、召回率、*F* 值为参考指标,分类结果见表 3。

表 3 最小支持度变化对实验结果的影响

Min_sup	FP-growth + NB			FP-growth + SVM		
	准确率	召回率	*F*	准确率	召回率	*F*
0.01	0.5812	0.5733	0.5791	0.6237	0.6132	0.6176
0.05	0.5793	0.5680	0.5767	0.6201	0.6141	0.6017
0.10	0.5721	0.5714	0.5623	0.6219	0.6079	0.6129
0.15	0.5638	0.5571	0.5489	0.6176	0.6035	0.6008

由表 3 可知,采用 FP-growth 算法并将最小支持度设为0.01时,获得到的关联特征用于 NB、SVM 情绪分类效果最佳,因为 3 种参考指标的值都最高。那么,选取最小支持度为0.01进行对比实验,实验结果见表 4。

表 4 情绪分类结果对比

方法	NB			SVM		
	准确率	召回率	*F*	准确率	召回率	*F*
CHI	0.5618	0.5521	0.5597	0.5714	0.5644	0.5706
FP-growth	0.5812	0.5733	0.5791	0.6237	0.6132	0.6176
FP-growth + CHI	0.6137	0.6014	0.6031	0.6307	0.6274	0.6223

对比实验结果发现,无论采用 NB 还是 SVM 的情况下,FP-growth 都比 CHI 具有更好的实验效果。尽管 CHI 方法也考虑到了特征之间的关联性,但是实验结果说明 FP-growth 方法更适合挖掘特征的关联性,适合于网络评论类文本的特征挖掘。而将 CHI 和 FP-growth 结合起来的方法提升了分类的效果,因为 CHI 补充了一些 FP-growth 未发现到的关联特征。同时,在 3 种不同的特征挖掘方法下,SVM 分类算法在 3 个指标上都高于 NB 算法,也说明了 SVM 算法比 NB 算法更适用于处理关联性特征。

3 结束语

文中将关联规则用于情感分析,并同机器学习结合实现了网络评论的情绪分类。通过采用 FP-growth 算法挖掘关联特征、CHI 方法补充关联特征的方法,促进了有效特征的提取。采用 SVM 算法为特征添加情绪标签的方法,评论情绪分类的准确率等明显增加。以此建立的网络评论情感分析模型也具有较好的实验效果,说明了关联规则挖掘特征对评论文本的情感分析有一定的帮助。

此外,根据实验数据显示,提出的模型在准确率等方面仍有待提高。今后,还可以结合其他的特征挖掘算法来进行优化,并选用多种不同的机器学习分类方法来实验,以进一步提升情感分析的效果。

参考文献:

[1] 林钦和,刘钢,陈荣华. 基于情感计算的商品评论分析系统[J]. 计算机应用与软件,2014,31(12):39-44.

[2] 周咏梅,杨佳能,阳爱民. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报(工学版),2013,43(6):27-33.

[3] 赵刚,徐赞. 基于机器学习的商品评论情感分析模型研究[J]. 信息安全研究,2017,3(2):166-170.

[4] 周杰,林琛,李弼程. 基于机器学习的网络新闻评论情感分类研究[J]. 计算机应用,2010,30(4):1011-1014.

[5] 姜杰,夏睿. 机器学习与语义规则融合的微博情感分类方法[J]. 北京大学学报(自然科学版),2017,53(2):247-254.

[6] 于重重,操镭,尹蔚彬,等. 吕苏语口语标注语料的自动分词方法研究[J]. 计算机应用研究,2017,34(5):1325-1328.

[7] 李言武,郑勇. 基于语义扩展的汉语全覆盖关键词提取算法[J]. 控制工程,2018,25(7):1326-1334.

[8] John D. Holt, Soon MChung. Multipass Algorithms for Mining Association Rules inTextDatabases[J]. Knowledge and Information Systems, 2001, 3(2): 13-17.

[9] 刘爽,赵景秀,杨红亚,等. 文本情感分析综述[J]. 软件导刊,2018,17(6):1-4.

[10] 杨经,林世平. 基于 SVM 的文本词句情感分析[J]. 计算机应用与软件,2011,28(9):225-228.

[11] Hang C, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews [C]. Proceedings of the 21 National Conference on Artificial Intelligence. New York: Mountation View, 2006: 1265-1270.

[12] 刘思,朱福喜,阳小兰,等. 基于分类关联规则的微博情绪分析[J]. 计算机工程与设计,2016,37(12):3361-3365.

[13] 明均仁. 融合语义关联挖掘的文本情感分析算法研究[J]. 图书情报工作,2012,56(15):99-103.

[14] LipikaDey, SkMirajulHaque. Opinion mining from noisy text data[J]. International Journal on Document Analysis and Recognition (IJDAR), 2009, 12(3): 32-34.

[15] Ramanathan Narayanan, Bing Liu, Alok-Choudhary. Sentiment Analysis of Conditional Sentences[C]. In: Proceedings of the 2009 Conference on EMNLP. Morristown, USA: ACL, 2009: 180-189.

Research on Emotional Analysis Model based on Association Rules

YANG Di,　WEN Chengyu

(College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: For varieties of non-standard format comments on the Internet, An emotional analysis model combined with association rule is proposed to improve the emotional analysis accuracy of Internet commentsBased on the analysis of the emotional tendency of comments, it deeply explores all the related features of the comment objects, and fully considers the effect of each word in the text on the overall emotional tendency. Firstly, a kind of dataset with clear comment object is selected, and the candidate feature is obtained by traditional feature extraction method. Then, the association feature of the comment object is mined by association rules, and finally uses SVM and NB algorithm to classify emotion. The experimental examination and comparative analysis of result suggest that this model can effectively improve the accuracy, recall rate and F value of emotional classification of Internet comments.

Keywords: emotional analysis; data mining; association rules; SVM; NB