基于序贯三支决策的代价敏感文本情感分析方法

范 琴[1] 刘 盾[1] 叶晓庆[1]

摘 要 为了解决文本情感分析的代价不平衡及静态决策中分类代价偏高的问题,文中考虑动态决策过程中产生的误分类代价和学习代价,构建基于序贯三支决策的代价敏感文本情感分析方法. 首先,为了构建多粒度动态决策环境,提出针对文本数据的粒化模型. 然后,引入序贯三支决策模型,构建动态文本分析框架. 最后,利用真实文本评论数据集验证文中方法的有效性. 实验表明文中方法在提高分类质量的同时,明显降低整体的决策代价.

关键词 情感分析, 文本挖掘, 三支决策, 代价敏感, 粒计算

引用格式 范 琴,刘 盾,叶晓庆. 基于序贯三支决策的代价敏感文本情感分析方法. 模式识别与人工智能, 2020, 33(8): 732-742.

DOI 10.16451/j.cnki.issn1003-6059.202008007 **中图法分类号** TP 18

Cost-Sensitive Text Sentiment Analysis Based on Sequential Three-Way Decision

FAN Qin[1], LIU Dun[1], YE Xiaoqing[1]

ABSTRACT To solve the problems of cost imbalance in text sentiment analysis and high classification cost in static decision-making, a cost-sensitive text sentiment analysis method is constructed based on sequential three-way decision, and the misclassification cost and learning cost in dynamic decision-making process are taken into account. Firstly, a granulation model for text data is proposed to construct a multi-level granular structure. Next, sequential three-way decision is introduced to set a dynamic text analysis framework. Finally, real text review datasets are utilized to validate the effectiveness of the proposed method. Experimental results show that the proposed method significantly reduces the overall decision-making cost with the improved classification quality.

Key Words Sentiment Analysis, Text Mining, Three-Way Decision, Cost-Sensitive, Granular Computing

Citation FAN Q, LIU D, YE X Q. Cost-Sensitive Text Sentiment Analysis Based on Sequential Three-Way Decision. Pattern Recognition and Artificial Intelligence, 2020, 33(8): 732-742.

收稿日期: 2020-06-15; **录用日期:** 2020-07-29
Manuscript received June 15, 2020;
accepted July 29, 2020
国家自然科学基金项目(No. 61876157,71571148)、**西南交通大学杨华学者 A 类计划**(No. 201806) **资助**
Supported by National Natural Science Foundation of China(No. 61876157,71571148), Yanghua Scholar Plan(Part A) of SWJTU(No. 201806)
本文责任编委 苗夺谦
Recommended by Associate Editor MIAO Duoqian

1. **西南交通大学 经济管理学院 成都** 610031
1. School of Economics and Management, Southwest Jiaotong University, Chengdu 610031

随着 Web3.0 的成型与发展,电子商务平台趋于成熟,用户评论成为线上销售商家和消费者的关注重点. 作为用户生成内容(User Generated Content, UGC)的表现形式之一,每天都有大量用户评论产生,观点分享变得更加方便和容易. 据研究调查显示,仅 6% 的消费者有发表评论的习惯,其他消费者会留下评论大多是因为对产品或服务感到特别满意或失望,通过评论表达自己的情感倾向. 一些包含大

量用户生成内容的平台,如淘宝、携程等,通过对评论情感极性的分析以吸引潜在顾客、降低客户流失率、维护顾客忠诚度. 总之,对评论进行情感分析不仅可以快速发现用户偏好,促进消费,还能减少用户反馈响应时间,提高产品或服务质量.

文本情感分析[1-2]也称意见挖掘,目的是从文本中挖掘用户观点及情感倾向,现已成功应用于推荐系统[3]、电子商务[4]、智能客服[5]、舆情分析[6]等领域. 目前,常见的情感分析方法主要包括基于语义统计的方法和基于机器学习的方法[7-9].

基于语义统计的方法[10-11]常利用情感词典和语料库挖掘情感词,一般存在语义规则. 当前通用的情感词典有 WordNet 和 HowNet. Hu 等[12]通过 WordNet与自建形容词库确定评论中观点句的情感极性,分类效果较优. 朱嫣岚等[13]基于 HowNet 提出计算中文词汇语义倾向的方法,提高词汇的情感分类精度.

基于机器学习的方法使用相关机器学习算法实现文本情感分析,其中朴素贝叶斯[9]、支持向量机[9]、逻辑回归[14]常被用于解决文本情感分析问题. Pang 等[9]将上述三种算法运用于文本情感分析中,对比分析不同算法的分类精度. Catal 等[14]结合多种机器学习算法对评论进行情感分析,并基于投票机制构建虚假负面评论检测模型.

此外,相比传统机器学习方法,集成学习和深度学习在情感分析问题中通常具有更优的表现. 例如随机森林[15]、梯度提升决策树[16]、极端梯度提升[17]等集成学习方法和卷积神经网络[18]、循环神经网络[19]等深度学习方法被广泛运用. Sadhasivam 等[20]提出基于集成学习的情感分析方法,分类精度较高,并运用于产品推荐中. Zhou 等[21]提出深度多任务学习模型,用于检测文本的情感极性,相比机器学习和基础深度学习算法,分类效果更优. Avinash 等[22]对比集成学习方法与深度学习方法,表明基于分类精度和时间成本两个维度集成学习的表现更优.

通过上述分析可发现,传统情感分析均为静态决策过程,以单步决策划分情感类别为主,忽略决策过程的动态特征. 此外,现有研究主要侧重于追求较高的分类精度[9,12-14,20-22],较少考虑决策行为产生的决策代价及代价的不平衡性问题. 然而,在实际应用中,决策代价也是分类决策中需要考虑的重要因素[23-24]. 考虑到分类代价不平衡性,Li 等[25]提出代价敏感决策方法,用于解决人脸识别问题. Liu 等[26]基于推荐系统中的代价不平衡性,提出动态三支粒推荐算法. 此外,传统文本情感分析方法大多通过一次性静态决策做出判断,使决策产生较高的误分类代价.

以医疗诊断为例,对于一些常见疾病,医生可通过望闻问切快速诊断患者是否患病,而对于疑难杂症,通常需要付出更多的成本以进一步诊断,如时间成本和检查费用成本等. 因此,医生有根据患者病症信息综合衡量进一步检查的必要性,不仅降低误诊率,避免一次性静态诊断造成的决策风险,还能减少相应成本. 因此,在分类过程中有必要考虑决策过程的动态性和分类决策的代价不平衡问题.

为了解决文本情感分析中静态决策导致误分类代价偏高的问题,本文将粒计算(Granular Computing, GrC)思想引入文本情感分析中. 粒计算是解决复杂问题的有效方法,通过信息粒化和基于粒化信息的计算两个步骤完成问题求解. 在粒计算过程中,使用粒子对客观事物进行抽象表示,将粒子聚合为粒层,在粒层间的多层次粒结构中完成问题的动态求解. 一般情况下,存在两种粒结构: 自上而下粒度逐渐细化的拆分型与自下而上粒度逐渐粗化的聚合型[27]. 因此,选择合理的粒结构与粒度划分是构建多粒度动态决策模型的关键.

此外,考虑到决策行为中产生的代价及代价不平衡的问题,本文在文本情感分析中引入序贯三支决策(Sequential Three-Way Decision, S3WD)思想. 序贯三支决策是由三支决策发展而来用于解决不确定性问题的动态方法[28],目前已被广泛应用于图像识别、人脸识别、推荐系统及文本情感分析等领域. Zhang 等[29]根据文本情感分析中产生的误分类代价和时间代价,提出结合集成学习的代价敏感情感分类方法. 张刚强等[30]提出基于 N 元词的多粒度中文情感分析方法,通过实验证明其鲁棒性.

基于上述分析,本文首先结合文本情感分析与粒计算,通过对文本数据粒化构建多粒度动态决策环境. 为了避免粒化维度过高的问题,采用隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)和非负矩阵分解(Nonnegative Matrix Factorization, NMF)构造具有语义解释性的粒子,运用自上而下的粒结构,构建基于封装式、嵌入式、集成式三种准则的粒化模型. 然后,基于多粒度动态决策环境,将序贯三支决策引入文本情感分析,考虑到决策过程中产生的误分类代价和学习代价,提出基于序贯三支决策的代价敏感文本情感分析方法. 最后,利用真实文本评论数据验证方法的有效性.

1 相关概念

1.1 粒计算

粒计算是一种规范化的问题求解范式,研究主要集中于两类:1)关注不确定性处理与计算,如粗糙集模型;2)关注多粒度、多层次、多视角的计算,如商空间模型.

考虑到粒计算特征层次性、信息递增性的特点,Yao[31]将粒计算和决策粗糙集引入三支决策中,提出具有代价敏感性的序贯三支决策模型.因此本文融合粒计算与序贯三支决策,构建具有代价敏感的动态文本情感分析模型.

1.2 序贯三支决策

三支决策(Three-Way Decision,3WD)是由决策粗糙集理论延伸的一种不确定性决策方法,通过"三分而治"的决策思想将论域划分为三部分,采用"分而治之"的策略进行决策,整体框架如图1所示.

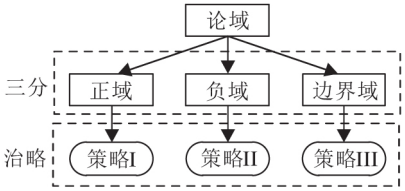

图1 三支决策框架图

Fig. 1 Framework of three-way decision

序贯三支决策是基于三支决策构建的动态决策方法[32].它从粒计算的角度,采用"自上而下,化繁为简"的分治策略,根据多步骤的动态决策框架,形成粒度由粗到细的多层次序贯过程.在序贯三支决策中,除最后一个粒层采用二支决策以外,其它决策层均为三支决策.由于序贯三支过程依据代价最小化进行决策,具有代价敏感性.

1.3 粒子表示

1.3.1 隐含狄利克雷分布

LDA[33]是针对文本数据的概率主题模型,可解决文本中一词多义和多词一义问题.LDA的三层贝叶斯模型是基于概率隐性语义分析模型的延伸,如图2所示,阴影表示模型中可观测值.

假设语料库中有V个非重复词汇,M篇文档,K表示主题数,N_m表示第m篇文档的词汇数量,$z_{m,n}$表示第m篇文档的第n个词的主题,$w_{m,n}$表示可观测的第m篇文档的第n个词.φ_k表示不同主题对应的词分布,θ_m表示不同文档的主题分布,φ_k、θ_m均服从多项分布.狄利克雷分布与多项分布互为共轭分布,因此将狄利克雷分布的参数α、β作为超参数引入LDA中.

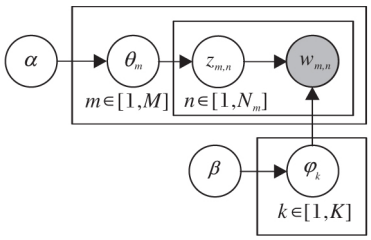

图2 LDA模型图

Fig. 2 LDA model

当超参数α、β给定时,LDA的联合概率分布

$$p(\theta,z,w\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N_m}p(z_n\mid\theta)\,p(w_n\mid z_n,\beta).$$

本文采用Gibbs采样对φ_k和θ_m进行估计求解,参数估计结果如下:

$$\widehat{\varphi}_{k,t}=\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t},\quad \widehat{\theta}_{m,k}=\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m,\neg i}^{(k)}+\alpha_k},$$

$t\in V$.

进而得到LDA的Gibbs采样公式:

$$p(z_i=k\mid z_{\neg i},w)\propto\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m,\neg i}^{(k)}+\alpha_k}\cdot\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t}.$$

1.3.2 非负矩阵分解

NMF[34]是解决维数灾难问题的有效方法,并且NMF中存在非负数约束条件,使结果具有可解释性,因此被广泛用于图像分解、文本聚类等领域.

NMF的核心思想为矩阵分解,对于任意非负矩阵$\boldsymbol{V}_{q\times p}$,存在2个非负矩阵$\boldsymbol{W}_{q\times r}$和$\boldsymbol{H}_{r\times p}$,使

$$\boldsymbol{V}_{q\times p}\approx\boldsymbol{W}_{q\times r}\boldsymbol{H}_{r\times p},$$

其中,$\boldsymbol{W}_{q\times r}$为权重矩阵,$\boldsymbol{H}_{r\times p}$为特征矩阵.NMF原理图如图3所示.

假设语料库中有q个非重复词汇,p篇文档,利用矩阵分解将原始矩阵$\boldsymbol{V}_{q\times p}$分解成文档-主题分布矩阵$\boldsymbol{H}_{r\times p}$和主题-词汇分布矩阵$\boldsymbol{W}_{q\times r}$.

NMF通过不断迭代$\boldsymbol{W}$和$\boldsymbol{H}$使其逼近原始矩阵$\boldsymbol{V}$以求解,等同于如下优化问题:

$$\min D(\boldsymbol{V},\boldsymbol{WH}),$$

$$\text{s.t. }\boldsymbol{W}\geqslant\boldsymbol{0},\ \boldsymbol{H}\geqslant\boldsymbol{0},$$

其中$D(\boldsymbol{V},\boldsymbol{WH})$是衡量$\boldsymbol{V}$与$\boldsymbol{WH}$距离差异性的目标函数,一般可用欧几里得距离和KL散度表示:

$$\| \boldsymbol{V} - \boldsymbol{WH} \|^2 = \sum_q \sum_p (\boldsymbol{V}_{qp} - [\boldsymbol{WH}]_{qp})^2,$$

$$KL(\boldsymbol{V} \| \boldsymbol{WH}) = \sum_q \sum_p \left(\boldsymbol{V}_{qp} \ln \left(\frac{\boldsymbol{V}_{qp}}{[\boldsymbol{WH}]_{qp}} \right) - \boldsymbol{V}_{qp} + [\boldsymbol{WH}]_{qp} \right),$$

$[\boldsymbol{WH}]_{qp}$ 表示矩阵 $\boldsymbol{WH}$ 乘积的第 q 行、第 p 列的元素.

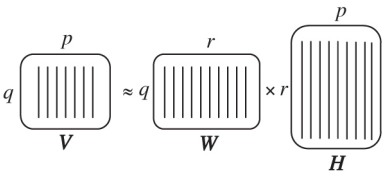

图 3　NMF 原理图

Fig. 3　Illustration of NMF

2　基于序贯三支决策的代价敏感文本情感分析方法

为了解决文本情感分析中代价不平衡及静态决策导致误分类代价偏高的问题，本文提出基于序贯三支决策的代价敏感文本情感分析方法. 一方面使用封装式、嵌入式、集成式方法对文本数据进行粒化; 另一方面，基于文本数据粒化模型搭建多粒度动态决策环境，考虑误分类代价和学习代价，构建序贯三支文本情感分析模型.

2.1　文本数据粒化模型

粒化是对问题空间的划分，涉及粒子、粒化准则和粒结构. 其中，粒子是基本要素，影响粒层与粒结构的形成. 一般文本数据的粒子生成包括两种方式: 直接生成特征向量; 在特征向量的基础上，进一步转化为文本向量表示.

考虑到文本数据直接生成向量表示会导致数据维度过高、稀疏性较大的问题，本文使用 LDA 和 NMF 构造粒子，将数据映射到主题层次的低维特征空间，生成具有语义解释性的粒子. 同时，采用自上而下的粒结构，根据粒化准则完成文本数据粒化，构建具有多层次粒结构的文本数据粒化模型.

针对 LDA，在已确定粒子数 k 的情况下，将 M 篇文档转化为可观测值矩阵 $\boldsymbol{W}_{M \times N_m}$. 通过 Gibbs 采样生成文档-主题向量 $\boldsymbol{\theta}_m$，$\boldsymbol{\theta}_m$ 表示第 m 篇文档的主题分布，$m = 1, 2, \cdots, M$，进而生成粒子集合矩阵

$$\boldsymbol{\theta}_{M \times k} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_M)^{\mathrm{T}},$$

并得到 k 个粒子 $att_1, att_2, \cdots, att_k$. 同理，对于 NMF，在已确定粒子数 r 的情况下，将 p 篇文档转化为原始矩阵 $\boldsymbol{V}_{q \times p}$，通过矩阵分解生成文档-主题分布的粒子集合矩阵 $\boldsymbol{H}_{r \times p}$，得到 r 个粒子 $att_1, att_2, \cdots, att_r$. 上述步骤实现将复杂文本转化为若干主题表示粒子的过程.

完成粒子构造后，需要进一步考虑粒子的重要性程度. 本文采用前向叠加的粒化准则构造粒层，形成自上而下的粒结构. 假设共有 S 个粒子，表示为 att_j，$j = 1, 2, \cdots, S$，根据粒化准则和粒子重要性程度，按降序排列 $att^1, att^2, \cdots, att^S$，则粒子叠加生成 S 个粒层:

$$Gr_1 = [att^1], \cdots, Gr_j = [att^1, att^2, \cdots, att^j], \cdots,$$

$$Gr_S = [att^1, att^2, \cdots, att^S],$$

简记为 Gr_j，$j = 1, 2, \cdots, S$.

基于特征选择原理，本文提出 3 种粒化准则: 封装式、嵌入式和集成式[35]，分别使用支持向量机-递归特征消除算法(Support Vector Machine-Recursive Feature Elimination，SVM-RFE)、信息增益(Information Gain，IG)和随机森林(Random Forest，RF)实现.

封装式方法结合分类算法，通过对比特征子集在分类算法上的精度衡量所选特征的优劣. SVM-RFE 将 SVM 与后向搜索策略结合，泛化能力较强.

SVM-RFE 运用递归消除算法进行特征选择，根据 $att^S, att^{S-1}, \cdots, att^1$ 的顺序消除一定数量的特征，实现特征选择. 假设消除 2 个特征，算法会按照顺序消除 att^S、att^{S-1} 这两个重要性程度最低的特征，这等价于完成 att^S、att^{S-1} 的排序. 为了得到所有特征的重要性程度排序，设定特征选择数量为 1，通过消除 $S-1$ 个特征，完成对所有粒子排序. SVM-RFE 的主要原理是利用 SVM 获得所有粒子的权重因子，剔除权重因子最小的粒子，重复此步骤，直至完成排序. 粒化准则如下:

$$SVM_{rfe}(att_j) = \omega_j^2,$$

其中 ω_j 表示粒子 att_j 的权重因子. $SVM_{rfe}(att_j)$ 值越大，表示粒子 att_j 重要性程度越高.

嵌入式方法结合学习器与特征选择，学习器运行结束的同时完成 $att_1, att_2, \cdots, att_S$ 的排序. 本文将 IG 作为衡量标准，形成如下粒化准则:

$$Gain(D, att_j) = H(D) - H(D \mid att_j),$$

其中，$H(D)$ 表示所有粒子集合的信息熵，$H(D \mid att_j)$ 表示粒子 att_j 的条件熵. $Gain(D, att_j)$ 值越大，表示粒子 att_j 重要性程度越高.

集成式方法借鉴集成学习的思想完成特征选择，文本使用 RF 对 $att_1, att_2, \cdots, att_s$ 进行排序.

RF从样本数据中随机抽样构造N棵决策树,未选中的样本称为袋外数据(Out of Bag, OOB).先使用OOB计算第$k(k=1,2,\cdots,N)$棵树的决策树分类误差,记为err^k_{oob1}.再对OOB中的粒子att_j加入随机噪声,并再次计算分类误差,记为err^k_{oob2}.最后根据加入随机噪声对分类结果的影响判断粒子att_j的重要性程度,形成如下粒化准则:

$$RF(att_j)=\frac{1}{N}\sum_{k=1}^{N}(err^k_{oob2}-err^k_{oob1}).$$

$RF(att_j)$值越大,表示粒子att_j重要性程度越高.

基于上述分析,本文利用LDA和NMF对文本进行粒子表示,提出基于封装式、嵌入式和集成式的粒化准则,并构建6种具有多层次粒结构的文本数据粒化模型:LDA-SVM-RFE、LDA-IG、LDA-RF、NMF-SVM-RFE、NMF-IG和NMF-RF.

2.2 序贯三支文本情感分析方法

序贯三支文本情感分析方法的核心是基于文本数据粒化模型构建的动态决策环境,运用"三分而治"的决策思想实现序贯文本情感分析的过程,使用划分阈值判断文本的情感极性.

2.2.1 序贯三支决策过程

本节采用2.1节的粒化模型构建多层次粒结构,$Gr_j(j=1,2,\cdots,S)$表示粒度由粗到细的S个粒层.在文本情感分析中,x表示文本对象,$P(X\mid x^j)$表示在粒层Gr_j下x属于正情感极性这一概念X的概率.通过对比$P(X\mid x^j)$与三支情感分析阈值对

$$(\alpha_1,\beta_1),(\alpha_2,\beta_2),\cdots,(\alpha_{S-1},\beta_{S-1})$$

或二支情感分析阈值γ_S完成文本分类.在粒层$Gr_j(j=1,2,\cdots,S-1)$中,采用三支情感分析,将文本数据划分至3个不同的域:正域$P_j(x)$、负域$N_j(x)$和边界域$B_j(x)$.当$P(X\mid x^j)$大于等于α_j时,将x划分至$P_j(x)$,情感极性为正;当$P(X\mid x^j)$小于等于β_j时,将x划分至$N_j(x)$,情感极性为负;否则将x划分至$B_j(x)$.

假设给定粒层Gr_j,在粒层Gr_j不再考虑粒层Gr_j前完成文本的分类:

$$\left(\bigcup_{j=1}^{j-1}P_j(x)\right)\cup\left(\bigcup_{j=1}^{j-1}N_j(x)\right).$$

同时,随着粒层的增加,边界域的待划分文本数量逐渐减少.如果到达粒层Gr_{j^*},$j^*=1,2,\cdots,S-1$,使$B_{j^*}(x)=\varnothing$,则结束划分;如果经过粒层Gr_{S-1}划分后使$B_{S-1}(x)\neq\varnothing$,需要在粒层$Gr_S$对上一粒层$B_{S-1}(x)$中的文本进行二支情感分析.通过对比$P(X\mid x^S)$与$\gamma_S$的关系对进行划分.当

$$P(X\mid x^S)\geqslant\gamma_S$$

时,将x划分到$P_j(x)$,情感极性为正;当

$$P(X\mid x^S)<\gamma_S$$

时,将x划分到$N_j(x)$,情感极性为负.

2.2.2 阈值计算过程

在文本情感分析中正负情感极性使用状态集

$$\Omega=\{X,\neg X\}$$

表示,X表示情感极性为正(P),$\neg X$表示情感极性为负(N).针对粒层$Gr_j(j=1,2,\cdots,S-1)$采用的三支情感分析对应三种决策行为:

$$D=\{\text{正极性},\text{负极性},\text{延迟判断}\},$$

记为

$$D=\{a_P,a_N,a_B\},$$

表示将文本数据分别划分到$P_j(x)$、$N_j(x)$和$B_j(x)$.根据决策过程中的状态集和决策行为,粒层$Gr_j(j=1,2,\cdots,S-1)$的三支情感分析代价矩阵如表1所示.表中:λ^j_{PP}、λ^j_{BP}、λ^j_{NP}分别表示若真实情感极性为正时采取a_P、a_N、a_B决策行为的代价;λ^j_{PN}、λ^j_{BN}、λ^j_{NN}分别表示若真实情感极性为负时采取a_P、a_N、a_B决策行为的代价.进一步地,λ^j_{PN}、λ^j_{NP}为接受错误和拒绝正确的误分类代价,λ^j_{BP}、λ^j_{BN}为延迟判断的学习代价.

表1 三支情感分析代价矩阵

Table 1 Cost matrix of three-way sentiment analysis

决策行为	真实情感极性	
	正(P)	负(N)
a_P	λ^j_{PP}	λ^j_{PN}
a_B	λ^j_{BP}	λ^j_{BN}
a_N	λ^j_{NP}	λ^j_{NN}

根据本文中文本情感分析包含X和$\neg X$两种状态集,可得

$$P(X\mid x^j)=1-P(\neg X\mid x^j),$$

则采取a_P、a_N、a_B三种行为的期望损失:

$$Loss_P=\lambda^j_{PP}P(X\mid x^j)+\lambda^j_{PN}P(\neg X\mid x^j),$$

$$Loss_N=\lambda^j_{NP}P(X\mid x^j)+\lambda^j_{NN}P(\neg X\mid x^j),$$

$$Loss_B=\lambda^j_{BP}P(X\mid x^j)+\lambda^j_{BN}P(\neg X\mid x^j).$$

根据贝叶斯决策理论的期望损失最小准则,得到如下决策规则:

1) a_P. $Loss_P\leqslant Loss_B$且$Loss_P\leqslant Loss_N$,判断为正极性,划分到$P_j(x)$;

2) a_N. $Loss_N\leqslant Loss_B$且$Loss_N\leqslant Loss_P$,判断为负极性,划分到$N_j(x)$;

3) a_B. $Loss_B \leqslant Loss_P$ 且 $Loss_B \leqslant Loss_N$, 延迟判断, 划分到 $B_j(x)$.

在文本情感分析中, 误分类代价通常高于延迟判断产生的学习代价, 学习代价通常高于正确分类产生的代价. 此外, 根据负性偏向理论[36], 负极性信息在决策行为中比正极性信息更具影响效用, 这说明将负极性评论判断为正极性评论的代价通常高于将正极性评论判断为负极性的代价. 因此, 有

$$0 \leqslant \lambda_{PP}^{j} \leqslant \lambda_{BP}^{j} < \lambda_{NP}^{j},$$

$$0 \leqslant \lambda_{NN}^{j} \leqslant \lambda_{BN}^{j} < \lambda_{PN}^{j},$$

$$\lambda_{PN}^{j} > \lambda_{NP}^{j}.$$

基于上述分析, 计算阈值对 (α_j, β_j), $j = 1, 2, \cdots, S-1$ 的取值:

$$\alpha_j = \frac{\lambda_{PN}^{j} - \lambda_{BN}^{j}}{(\lambda_{PN}^{j} - \lambda_{BN}^{j}) + (\lambda_{BP}^{j} - \lambda_{PP}^{j})},$$

$$\beta_j = \frac{\lambda_{BN}^{j} - \lambda_{NN}^{j}}{(\lambda_{BN}^{j} - \lambda_{NN}^{j}) + (\lambda_{NP}^{j} - \lambda_{BP}^{j})}.$$

此外, 当 $B_{S-1}(x) \neq \varnothing$ 时, 需要采用二支情感分析对剩下的文本数据进行划分. 这里考虑 a_P 和 a_N 两种决策行为, 将文本划分到 $P_S(x)$ 或 $N_S(x)$ 中. 特别地, 二支情感分析中只存在分类代价, 无学习代价, 表 2 给出粒层 Gr_S 下的二支情感分析代价矩阵.

表 2　二支情感分析代价矩阵

Table 2　Cost matrix of two-way sentiment analysis

决策行为	真实情感极性	
	正(P)	负(N)
a_P	λ_{PP}^{S}	λ_{PN}^{S}
a_N	λ_{NP}^{S}	λ_{NN}^{S}

同理, 采取 a_P 和 a_N 决策行为的期望损失:

$$Loss_P = \lambda_{PP}^{S} P(X \mid x^S) + \lambda_{PN}^{S} P(\neg X \mid x^S),$$

$$Loss_N = \lambda_{NP}^{S} P(X \mid x^S) + \lambda_{NN}^{S} P(\neg X \mid x^S).$$

利用贝叶斯决策理论的期望损失最小准则计算如下:

$$\gamma_S = \frac{\lambda_{PN}^{S} - \lambda_{NN}^{S}}{(\lambda_{PN}^{S} - \lambda_{NN}^{S}) + (\lambda_{NP}^{S} - \lambda_{PP}^{S})}.$$

综上所述, 基于序贯三支决策的代价敏感文本情感分析方法的框架如图 4 所示, 图中, Pr_j 表示 $P(X \mid x^j)$, $j = 1, 2, \cdots, S$. 该方法首先使用粒化准则搭建文本数据粒化模型. 然后基于粒化模型生成多层次粒结构, 有助于构建具有动态决策框架的文本情感分析模型. 最后完成文本情感分析方法搭建, 并进行文本数据划分: 当

$$B_{j^*}(x) = \varnothing, j^* = 1, 2, \cdots, S-1$$

时, 结束划分; 当

$$B_{S-1}(x) \neq \varnothing$$

时, 使用二支情感分析完成整体文本情感极性分类.

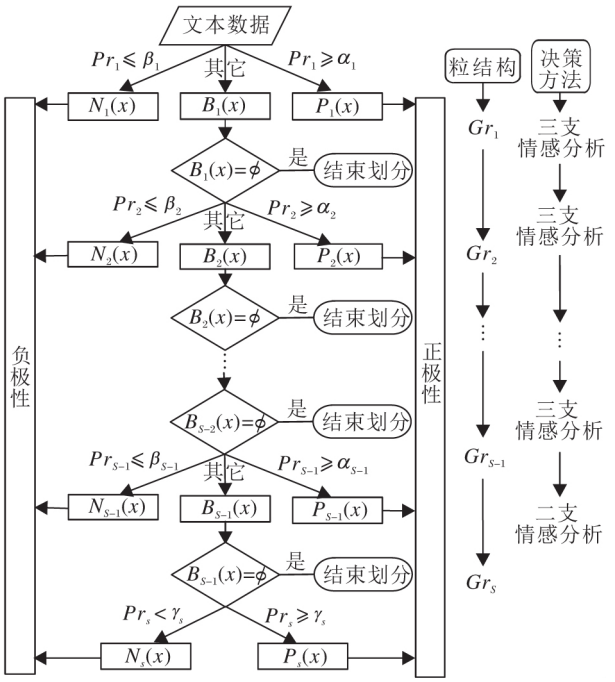

图 4　本文方法框架图

Fig. 4　Framework of the proposed method

3　实验及结果分析

为了验证本文方法的有效性, 选取文本情感分析中常用算法作为基准算法以进行对比分析, 对比方法如下: 朴素贝叶斯(Naive Bayesian, NB)、支持向量机(SVM)、逻辑回归(Logistic Regression, LR)、随机森林(RF)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、极端梯度提升(Extreme Gradient Boosting, XGBoost). 为了验证采用"三分而治"决策方法的有效性, 将代价不敏感二支决策(Cost-Blind Two-Way Decision, CB2WD)、代价敏感二支决策(Cost-Sensitive Two-Way Decision, CS2WD) 和代价敏感三支决策方法(Cost-Sensitive Three-Way Decision, CS3WD) 进行对比分析. 利用真实数据集进行实验.

3.1　实验数据集

本文采用两个携程网数据集: 携程-I(https://www.aitechclub.com/data-detail?data_id=29) 和

携程-H(https://www.aitechclub.com/data-detail?data_id=23).情感标签为1和0,分别表示正极性和负极性.携程-H共10 000条文本评论,经过数据去重和筛选后剩余7 756条评论,包含5 318条正向评论和2 438条负向评论,评论平均长度为128字.本文利用Jieba分词工具对数据进行分词和去除停用词,分词后平均词数为39个.携程-H共10 000条文本评论,经过数据去重和筛选后剩余9 978条评论,包含4 990条正向评论和4 988条负向评论,评论平均长度为109字.Jieba分词后平均词数为34个.

3.2 评价指标

本文主要选取如下两个维度的测度指标评估方法性能:分类质量和分类代价.

在分类质量中,考虑到实验数据中存在正负样本不均衡性,可构建2×2的混淆矩阵,采用精度(Precision)和F1值两个指标衡量分类质量:

$$Precision=\frac{TP}{TP+FP}=\frac{n_{PP}}{n_{PP}+n_{PN}},$$

$$Recall=\frac{TP}{TP+FN}=\frac{n_{PP}}{n_{PP}+n_{NN}},$$

$$F_1=2\cdot\frac{Precision\cdot Recall}{Precision+Recall},$$

其中,真实分类和预测分类分别对应真正例(TP)、真负例(TN)、假正例(FP)、假负例(FN),n_{PP}表示TP的划分数量,n_{NP}表示TN的划分数量,n_{PN}表示FP的划分数量,n_{NN}表示FN的划分数量.

在分类代价中,主要考虑误分类代价和学习代价,可定义第j个粒层Gr_j的代价:

$$C_j=\begin{cases}C_P(\alpha_j,\beta_j)+C_B(\alpha_j,\beta_j)+C_N(\alpha_j,\beta_j), & j=1,2,\cdots,S-1\\ C_P(\gamma_S)+C_N(\gamma_S), & j=S\end{cases}\tag{1}$$

其中,(α_j,β_j)为粒层$Gr_j(j=1,2,\cdots,S-1)$的三支情感分析阈值对,γ_S为粒层Gr_S的二支情感分析阈值,$C_P(\alpha_j,\beta_j)$、$C_P(\gamma_S)$表示将文本划分为正极性产生的分类代价,$C_N(\alpha_j,\beta_j)$、$C_N(\gamma_S)$表示将文本划分为负极性产生的分类代价,$C_B(\alpha_j,\beta_j)$表示对文本采取延迟判断产生的学习代价.

在C_j已知的前提下,假设在粒层$Gr_{j^*}(j^*=1,2,\cdots,S)$完成整体情感分类,总代价

$$TC=C_1+C_2+\cdots+C_{j^*}.$$

令n^j_{PP}、n^j_{NP}、n^j_{BP}表示在粒层Gr_j将真实情感极性为正的文本判断为正极性、负极性及延迟决策的数量;n^j_{PN}、n^j_{NN}、n^j_{BN}表示在粒层Gr_j将真实情感极性为负的文本判断为正极性、负极性及延迟决策的数量;$n^S_{BP}=n^S_{BN}=0$.

一般情况下,认为正确判断文本情感极性不产生代价,即$\lambda^j_{PP}=\lambda^j_{NN}=0$.根据式(1),可得第$j$个粒层$Gr_j$的代价和总代价:

$$C_j=\begin{cases}\lambda^j_{PN}n^j_{PN}+\lambda^j_{BP}n^j_{BP}+\lambda^j_{BN}n^j_{BN}+\lambda^j_{NP}n^j_{NP}, & j=1,2,\cdots,S-1\\ \lambda^S_{PN}n^S_{PN}+\lambda^S_{NP}n^S_{NP}, & j=S\end{cases}$$

$$TC=\begin{cases}\sum\limits_{j=1}^{j^*}(\lambda^j_{PN}n^j_{PN}+\lambda^j_{BP}n^j_{BP}+\lambda^j_{BN}n^j_{BN}+\lambda^j_{NP}n^j_{NP}), & j^*=1,2,\cdots,S-1\\ \sum\limits_{j=1}^{S-1}(\lambda^j_{PN}n^j_{PN}+\lambda^j_{BP}n^j_{BP}+\lambda^j_{BN}n^j_{BN}+\lambda^j_{NP}n^j_{NP})+(\lambda^S_{PN}n^S_{PN}+\lambda^S_{NP}n^S_{NP}), & j^*=S\end{cases}$$

最终可计算情感分类的总体平均决策代价:

$$AC=\frac{TC_{j^*}}{\sum\limits_{j^*}(n^j_{PP}+n^j_{NP}+n^j_{BP}+n^j_{PN}+n^j_{NN}+n^j_{BN})}.$$

3.3 相关参数

本文使用LDA、NMF获取文本数据进行粒子表示.为了确定粒子数量,对比不同粒子数量的数据在NB、SVM、LR、XGBoost这4种机器学习算法下的表现.表3和表4为不同粒子表示方法的实验结果对比.

由表3和表4可知,在LDA和NMF中,当粒子数量为20时,分类准确趋于稳定.因此,将粒子数量的阈值设定为20.进一步地,考虑序贯三支决策的阈值规则,同时参考文献[25]和文献[26]中序贯三支决策的代价设置,设

$$\lambda^j_{PP}=\lambda^j_{NN}=0,\ \lambda^j_{PN}=60,\ \lambda^j_{NP}=40,$$

$$\lambda^j_{BP}=j,\ \lambda^j_{BN}=j,\ 1\leqslant j\leqslant S.$$

基于此种情况,满足

$$0\leqslant\beta_1\leqslant\beta_2\cdots\leqslant\beta_{S-1}\leqslant\gamma_S\leqslant\alpha_{S-1}\cdots\leqslant\alpha_2\leqslant\alpha_1\leqslant 1.$$

表 3　粒子数量不同时 4 种方法的分类准确度对比(LDA)

Table 3　Classification accuracy comparison of 4 methods with different number of granules(LDA)

数据集	方法	粒子数量						
		10	15	20	25	30	35	40
携程 -I	NB	0.7879	0.7739	0.7860	**0.7931**	0.7838	0.7682	0.7612
	SVM	0.8043	0.8091	**0.8137**	0.8083	0.8018	0.7942	0.8093
	LR	0.8036	0.8112	**0.8177**	0.8150	0.8084	0.8040	0.8159
	XGBoost	0.8204	0.8301	0.8351	0.8335	0.8262	0.8236	**0.8354**
	平均值	0.8041	0.8061	**0.8131**	0.8125	0.8051	0.7975	0.8055
携程 -II	NB	0.7833	**0.8042**	0.8039	0.7973	0.7833	0.7687	0.7911
	SVM	0.7948	**0.8086**	0.8066	0.8006	0.7970	0.7895	0.8024
	LR	0.7952	0.8090	**0.8098**	0.8026	0.7984	0.7916	0.8078
	XGBoost	0.8218	0.8484	0.8501	**0.8521**	0.8400	0.8327	0.8426
	平均值	0.7988	**0.8176**	**0.8176**	0.8132	0.8047	0.7956	0.8110

表 4　粒子数量不同时 4 种方法的分类准确度对比(NMF)

Table 4　Classification accuracy comparison of 4 methods with different number of granules(NMF)

数据集	方法	粒子数量						
		10	15	20	25	30	35	40
携程 -I	NB	0.6765	0.6858	0.6788	0.6701	0.6810	**0.6894**	0.6834
	SVM	0.7706	0.7925	**0.8080**	0.7938	0.7706	0.7758	0.7848
	LR	0.7771	0.7964	**0.8093**	0.7977	0.7951	0.7977	0.8080
	XGBoost	0.7758	0.7977	**0.8003**	0.7990	0.7951	0.7899	0.7951
	平均值	0.7500	0.7681	**0.7741**	0.7652	0.7605	0.7632	0.7678
携程 -II	NB	0.7515	0.7555	**0.7623**	0.7435	0.7355	0.6925	0.6853
	SVM	0.7806	0.7906	**0.7986**	0.7966	0.7906	0.7916	0.7866
	LR	0.7700	0.7851	**0.7896**	0.7811	0.7808	0.7801	0.7888
	XGBoost	0.7811	0.8026	0.8227	0.8223	**0.8298**	0.8216	0.8096
	平均值	0.7708	0.7835	**0.7933**	0.7859	0.7842	0.7715	0.7656

3.4　实验结果

为了验证方法的有效性,从分类质量和分类代价上进行对比分析. CSS3WD-SVM-RFE、CSS3WD-RF 和 CSS3WD-IG 表示 CS3WD 和三种粒化模型结合的文本情感分析方法. 在 LDA 和 NMF 中,各方法在 2 个数据集上的分类代价差异如表 5 和表 6 所示.

表 5　各方法在 2 个数据集上的分类结果对比(LDA)

Table 5　Classification result comparison of different methods on 2 datasets(LDA)

算法	携程 -I				携程 -II			
	代价	平均代价	精度	F1	代价	平均代价	精度	F1
NB	115300	14.8640	0.6857	0.8135	172300	17.2680	0.6289	0.7615
SVM	97500	12.5693	0.8053	0.8502	141200	14.1511	0.7901	0.7986
LR	98200	12.6595	0.7822	0.8500	154100	15.4440	0.7859	0.7902
RF	95700	12.3372	0.8243	0.8472	112800	11.3049	0.7927	0.8041
GBDT	94560	12.1903	0.8183	0.8526	110800	11.1044	0.7830	0.8052
XGBoost	90600	11.6798	0.8109	0.8676	111000	11.1245	0.7863	0.8085
平均值	98643	12.7167	0.7878	0.8469	133700	13.3995	0.7612	0.7947
CSS3WD-SVM-RFE	82280	10.6086	0.8264	0.8595	100940	10.1163	0.7866	0.8092
CSS3WD-RF	85160	10.9799	0.8186	0.8565	109700	10.9942	0.7691	0.7933
CSS3WD-IG	85120	10.9747	0.8383	0.8412	100060	10.0281	0.7913	0.8077
平均值	84187	10.8544	**0.8278**	**0.8524**	103567	10.3795	**0.7823**	**0.8034**

表6 各方法在2个数据集上的分类结果对比(NMF)

Table 6 Classification result comparison of different methods on 2 datasets(NMF)

算法	携程-I				携程-II			
	代价	平均代价	精度	F1	代价	平均代价	精度	F1
NB	144500	18.6283	0.6951	0.8154	165600	16.5965	0.6610	0.7555
SVM	116000	14.9542	0.8168	0.8452	149800	15.0130	0.8013	0.7730
LR	133200	17.1716	0.8038	0.8176	140400	14.0710	0.7611	0.7800
RF	93400	12.0407	0.8289	0.8566	108000	10.8238	0.7983	0.8126
GBDT	106000	13.6651	0.8315	0.8571	106900	10.7136	0.8146	0.8269
XGBoost	93700	12.0794	0.8233	0.8608	101200	10.1423	0.8326	0.8179
平均值	114908	14.8135	0.7999	0.8475	128650	12.8934	0.7782	0.7943
CSS3WD-SVM-RFE	90220	11.6323	0.7966	0.8595	83500	8.3684	0.8396	0.8272
CSS3WD-RF	72580	9.3579	0.8552	0.8689	82660	8.2842	0.8514	0.8214
CSS3WD-IG	67900	8.7545	0.9260	0.8429	86440	8.6631	0.8193	0.8329
平均值	76900	9.9149	**0.8593**	**0.8571**	84200	8.4386	**0.8367**	**0.8272**

由于传统分类算法无多层次粒结构和代价敏感性,仅考虑一次静态决策产生的误分类代价,代价设置为 $\lambda_{PN}=60$,$\lambda_{NP}=40$. 由表5和表6可知,传统方法产生的分类代价均高于本文方法,两者存在明显差距. 由此可见,相比传统方法,本文方法分类代价明显降低.

通过表5和表6可看到,本文方法除在分类代价上优势显著以外,在分类质量上也具有一定优势. 在精度方面,本文方法的平均表现最优. 在F1方面,本文方法的表现具有稳定性,平均表现最优. 总之,本文方法分类质量的平均表现情况明显优于传统方法.

此外,为了验证本文决策方法的有效性,考虑单一粒层的决策普遍存在二支决策和三支决策,本文对CS3WD与CB2WD、CS2WD进行对比分析. 表7和表8为本文提出的粒化模型下三种决策方法产生的分类代价和分类效果.

表7 3种决策方法在2个数据集上的分类结果对比(LDA)

Table 7 Classification result comparison of 3 decision-making methods on 2 datasets(LDA)

方法	决策	携程-I				携程-II			
		代价	平均代价	精度	F1	代价	平均代价	精度	F1
SVM-RFE	CB2WD	118251	15.2444	0.7926	0.8462	137149	13.7451	0.7207	0.7683
	CS2WD	93334	12.0338	0.8068	0.8417	131213	13.1502	0.7537	0.7612
	CS3WD	82280	**10.6086**	**0.8264**	**0.8595**	100940	**10.1163**	**0.7866**	**0.8092**
RF	CB2WD	122694	15.8172	0.7949	0.8468	128858	12.9142	0.7358	0.7734
	CS2WD	91747	11.8292	0.8106	0.8432	126543	12.6822	0.7674	0.7632
	CS3WD	85160	**10.9799**	**0.8186**	**0.8565**	109700	**10.9942**	**0.7691**	**0.7933**
IG	CB2WD	118602	15.2897	0.7840	**0.8448**	130750	13.1038	0.7310	0.7725
	CS2WD	96749	12.4741	0.7971	0.8407	128068	12.8350	0.7622	0.7634
	CS3WD	85120	**10.9747**	**0.8383**	0.8412	100060	**10.0281**	**0.7913**	**0.8077**

表8 3种决策方法在2个数据集上的分类结果对比(NMF)

Table 8 Classification result comparison of 3 decision-making methods on 2 datasets(NMF)

方法	决策	携程-I				携程-II			
		代价	平均代价	精度	F1	代价	平均代价	精度	F1
SVM-RFE	CB2WD	107946	13.9159	0.7966	0.8479	100097	10.0318	0.8231	0.7758
	CS2WD	90769	11.7031	**0.8142**	0.8454	98123	9.8339	**0.8404**	0.7691
	CS3WD	90220	**11.6323**	0.7966	**0.8595**	83500	**8.3684**	0.8396	**0.8272**
RF	CB2WD	106447	13.7227	0.7923	0.8471	101533	10.1757	0.8297	0.7635
	CS2WD	91861	11.8439	0.8112	0.8447	99306	9.9525	0.8494	0.7562
	CS3WD	72580	**9.3579**	**0.8552**	**0.8689**	82660	**8.2842**	**0.8514**	**0.8214**
IG	CB2WD	101404	13.0726	0.7960	**0.8483**	102487	10.2712	0.8094	0.7793
	CS2WD	90984	11.7308	0.8129	0.8456	99813	10.0033	**0.8289**	0.7729
	CS3WD	67900	**8.7545**	**0.9260**	0.8429	86440	**8.6631**	0.8193	**0.8329**

从表 7 和表 8 可见,代价敏感三支决策方法优于代价敏感二支决策方法,代价敏感二支决策方法优于精度敏感二支决策方法. 相比其它两种方法,代价敏感三支决策方法在分类代价上优势明显.

由表 7 和表 8 可知,除分类代价以外,本文方法在分类质量上也具有一定优势. 从整体角度上来看,在大多数情况下,本文方法在 F1 指标上表现较好,即使有少数情形不是最优方案,但其与最优结果差距甚微. 与此同时,本文方法在精度指标上表现良好,表明其在分类质量上存在一定优势,可充分说明方法的有效性.

综上所述,通过对分类质量和分类代价两类指标的分析,相比传统机器学习算法及相关二支决策方法,本文方法不仅可提高决策分类质量,而且还可降低决策分类代价.

4　结 束 语

考虑到文本情感分析的代价不平衡性及静态决策导致误分类代价偏高问题,本文提出基于序贯三支决策的代价敏感文本情感分析方法,构建针对文本数据的粒化模型,使用粒化模型生成多层次粒结构,完成动态决策环境搭建. 同时,考虑文本情感分析过程中产生的误分类代价和学习代价,引入序贯三支决策思想,结合粒化模型,实现序贯三支文本情感分析. 实验表明,相比传统机器学习算法和相关二支决策方法,本文方法在分类质量和分类代价上都得到改善.

今后将重点考虑文本情感分析与序贯三支决策融合的粒化机制和粒化模型,并将相关方法应用到其它文本情感分析问题中,在总体分类代价减少的同时,提升分类质量.

参 考 文 献

[1] ZHAO J, LIU K, XU L H. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Computational Linguistics, 2016, 42(3): 595-598.

[2] MEDHAT W, HASSAN A, KORASHY H. Sentiment Analysis Algorithms and Applications: A Survey. Ain Shams Engineering Journal, 2014, 5(4): 1093-1113.

[3] YI S S, LIU X F. Machine Learning Based Customer Sentiment Analysis for Recommending Shoppers, Shops Based on Customers´ Review. Complex and Intelligent Systems, 2020. DOI: 10.1007/s40747-020-00155-2.

[4] ARCHAK N, GHOSE A, IPEIROTIS P G. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Management Science, 2011, 57(8): 1485-1509.

[5] 宋双永,王 超,陈成龙,等. 面向智能客服系统的情感分析技术. 中文信息学报, 2020, 34(2): 80-95.
(SONG S Y, WANG C, CHEN C L, *et al*. Sentiment Analysis for Intelligent Customer Service Chatbots. Journal of Chinese Information Processing, 2020, 34(2): 80-95.)

[6] ZHAN Q Y, ZHUO W, HU W, *et al*. Opinion Mining in Online Social Media for Public Health Campaigns. Journal of Medical Imaging and Health Informatics, 2019, 9(7): 1448-1452.

[7] RANA T A, CHEAH Y N. Aspect Extraction in Sentiment Analysis: Comparative Analysis and Survey. Artificial Intelligence Review, 2016, 46(4): 459-483.

[8] SOMPRASERTSRI G, LALITROJWONG P. Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. Journal of Universal Computer Science, 2010, 16(6): 938-955.

[9] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment Classification Using Machine Learning Techniques [C/OL]. [2020-04-28]. https://arxiv.org/pdf/cs/0205070.pdf.

[10] MOREO A, ROMERO M, CASTRO J L, *et al*. Lexicon-Based Comments-Oriented News Sentiment Analyzer System. Expert Systems with Applications, 2012, 39(10): 9166-9180.

[11] KIM S M, HOVY E. Identifying and Analyzing Judgment Opinions // Proc of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, USA: ACL, 2006: 200-207.

[12] HU M Q, LIU B. Mining and Summarizing Customer Reviews // Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2004: 168-177.

[13] 朱嫣岚,闵 锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006, 20(1): 14-20.
(ZHU Y L, MIN J, ZHOU Y Q, *et al*. Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing. 2006, 20(1): 14-20.)

[14] CATAL C, GULDAN S. Product Review Management Software Based on Multiple Classifiers. IET Software, 2017, 11(3): 89-92.

[15] SRIVASTAVA A, SINGH V, DRALL G S. Sentiment Analysis of Twitter Data: A Hybrid Approach. International Journal of Healthcare Information Systems and Informatics, 2019, 14(2): 1-16.

[16] CAI Y, YANG K, HUANG D P, *et al*. A Hybrid Model for Opinion Mining Based on Domain Sentiment Dictionary. International Journal of Machine Learning and Cybernetics, 2019, 10: 2131-2142.

[17] PARIMALA M, PRIYA R M S, REDDY M P K, *et al*. Spatiotemporal-Based Sentiment Analysis on Tweets for Risk Assessment of Event Using Deep Learning Approach [C/OL]. [2020-04-28]. https://onlinelibrary.wiley.com/doi/full/10.1002/spe.2851.

[18] KIM Y. Convolutional Neural Networks for Sentence Classification [C/OL]. [2020-04-28]. https://arxiv.org/pdf/1408.5882.pdf.

[19] TANG D Y, QIN B, LIU T. Document Modeling with Gated Re-

current Neural Network for Sentiment Classification // Proc of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2015: 1422-1432.

[20] SADHASIVAM J, BABU R. Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm. International Journal of Mathematical Engineering and Management Sciences, 2019, 4(2): 508-520.

[21] ZHOU J, HUANG J X, HU Q V, *et al*. Is Position Important? Deep Multi-task Learning for Aspect-Based Sentiment Analysis [C/OI]. [2020-04-28]. https://link.springer.com/article/10.1007/s10489-020-01760-x.

[22] AVINASH M, SIVASANKAR E. Efficient Feature Selection Techniques for Sentiment Analysis. Multimedia Tools and Applications, 2020, 79(9): 6313-6335.

[23] PECENKA C, DEBELLUT F, BAR-ZEEV N, *et al*. Cost-Effectiveness Analysis for Rotavirus Vaccine Decision-Making: How Can We Best Inform Evolving and Complex Choices in Vaccine Product Selection? Vaccine, 2020, 38(6): 1277-1279.

[24] HANSEN K. Decision-Making Based on Energy Costs: Comparing Levelized Cost of Energy and Energy System Costs. Energy Strategy Reviews, 2019, 24: 68-82.

[25] LI H X, ZHANG L B, HUANG B, *et al*. Sequential Three-Way Decision and Granulation for Cost-Sensitive Face Recognition. Knowledge-Based Systems, 2016, 91: 241-251.

[26] LIU D, YE X Q. A Matrix Factorization Based Dynamic Granularity Recommendation with Three-Way Decisions. Knowledge-Based Systems, 2020, 191. DOI: 10.1016/j.knosys.2019.105243.

[27] **张钹,张铃. 粒计算未来发展方向探讨. 重庆邮电大学学报(自然科学版)**, 2010, 22(5): 538-540.
(ZHANG B, ZHANG L. Discussion on Future Development of Granular Computing. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2010, 22(5): 538-540.)

[28] YAO Y Y, DENG X F. Sequential Three-Way Decisions with Probabilistic Rough Sets // Proc of the 10th IEEE International Conference on Cognitive Informatics and Cognitive Computing. Washington, USA: IEEE, 2011: 120-125.

[29] ZHANG Y B, MIAO D Q, WANG J Q, *et al*. A Cost-Sensitive Three-Way Combination Technique for Ensemble Learning in Sentiment Classification. International Journal of Approximate Reasoning, 2019, 105: 85-97.

[30] **张刚强,刘 群,纪良浩. 基于序贯三支决策的多粒度情感分类方法. 计算机科学**, 2018, 45(12): 153-159.
(ZHANG G Q, LIU Q, JI L H. Multi-granularity Sentiment Classification Method Based on Sequential Three-Way Decisions. Computer Science, 2018, 45(12): 153-159.)

[31] YAO Y Y. Three-Way Decisions with Probabilistic Rough Sets. Information Sciences, 2010, 180(3): 341-353.

[32] YAO Y Y. Granular Computing and Sequential Three-Way Decisions // Proc of the International Conference on Rough Sets and Knowledge Technology. Berlin, Germany: Springer, 2013: 16-27.

[33] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993-1022.

[34] LEE D D, SEUNG H S. Learning the Parts of Objects by Non-negative Matrix Factorization. Nature, 1999, 401(6755): 788-791.

[35] **李邵琴,杜建强,聂 斌,等. 特征选择方法综述. 计算机工程与应用**, 2019, 55(24): 10-19.
(LI Z Q, DU J Q, NIE B, *et al*. Summary of Feature Selection Method. Computer Engineering and Applications, 2019, 55(24): 10-19.)

[36] ESLAMI S P, GHASEMAGHAEI M, HASSANEIN K. Which Online Reviews´ Do Consumers Find Most Helpful? A Multi-method Investigation. Decision Support Systems, 2018, 113: 32-42.

作者简介

范 琴,硕士研究生,主要研究方向为数据挖掘、知识发现、三支决策、粒计算. E-mail: isfanqin@163.com.
(**FAN Qin**, master student. Her research interests include data mining, knowledge discovery, three-way decision and granular computing.)

刘 盾(通讯作者),博士,教授,主要研究方向为数据挖掘、知识发现、粗糙集、粒计算、决策支持系统. E-mail: newton83@163.com.
(**LIU Dun**(Corresponding author), Ph. D., professor. His research interests include data mining, knowledge discovery, rough set, granular computing and decision support systems.)

叶晓庆,博士研究生,主要研究方向为三支决策、机器学习. E-mail: xqxiaoqingye@163.com.
(**YE Xiaoqing**, Ph. D. candidate. Her research interests include three-way decision and machine learning.)