

基于序贯三支决策的多粒度情感分类方法

张刚强 刘 群 纪良浩

(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

摘 要 如何对评论数据进行正确的情感分类是情感分析中的重要研究内容。从粒计算和认知学角度,提出了一种基于序贯三支决策的多粒度中文评论情感分类方法。首先,基于评论数据集的特点,根据评论中情感信息量的多少,提出一种由粗到细的多粒度情感信息表示方法;然后,结合序贯三支决策的思想在不同粒度依据情感信息进行逐步计算,对边界域评论序贯地进行三支决策;最后,根据不同粒度的决策阈值和成本对评论做出最终的情感分类。对比实验结果表明,该方法在 3 个经典评论数据集上获得了更好的结果,具有更高的分类正确率和更强的鲁棒性。

关键词 情感分类,序贯三支决策,多粒度,认知

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.12.024

Multi-granularity Sentiment Classification Method Based on Sequential Three-way Decisions

ZHANG Gang-qiang LIU Qun JI Liang-hao

(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract How to classify the review data correctly is important research content in sentiment analysis. From the perspective of granular computing and cognitive science, this paper proposed a multi-granularity sentiment classification method for Chinese reviews based on sequential three-way decisions. Firstly, based on the characteristics of review data, a coarse-to-fine multi-granularity sentiment information representation method is put forward according to the amounts of sentiment information existing in the review. Then, by combining the principle of sequential three-way decisions, the calculation is gradually executed in different sentiment information granularity and the sequenced three-way decision is carried out for the boundary reviews. Lastly, according to the decision thresholds and costs in different granularities, the final sentiment classification is provided for the review data. The experimental results show that the proposed method achieves better performance, and performs higher classification accuracy and stronger robustness on three classic datasets.

Keywords Sentiment classification, Sequential three-way decisions, Multi-granularity, Cognitive

1 引言

随着互联网和社交网络的快速发展,各种评论已经成为一种新型的文本,大量地出现在人们的日常生活中。常见的评论主要有产品评论、新闻评论和社交评论等。好的产品评论可以吸引更多的用户,促进产品的销售;消极的评论能够为产品供应商和卖家提供有效的改进建议,促其产品改良以增加市场的竞争力。如何对大量评论数据进行快速、高效和低成本的情感分析,已经成为当下自然语言处理中的热点之一。情感分析^[1],也称作意见挖掘或评论挖掘,是分析人们对产品、服务及其属性等的情感、评价和态度的研究。情感分析经过多年的研究与发展,其内容主要包括主客观识别、情感分类和情感极性强度判别。情感分类作为情感分析中的重要组

成部分,具有重要的研究价值。文献[2-4]对情感分析的相关研究工作进行了综述和分析。常见的情感分类方法主要有基于情感词典与规则的方法^[5]和基于特征的机器学习方法^[6-8]。基于情感词典和规则的方法主要依赖于情感词典的构建和规则的选取,就不同领域的评论数据而言,种子情感词典的分类效果不佳,领域情感词典的迁移能力弱;而传统的机器学习方法主要通过提取数据的特征来训练模型以获得输出,不同的特征提取方式对分类结果有很大的影响。三支决策^[9](Three-way Decisions)是由 Yao 提出的一种符合人类认知的决策模式,其主要思想是在决策过程中,如果人们因为现有信息不精确或者不完整而无法做出接受或拒绝的决定,则采用第三种方式延迟决策。在现有研究成果中,将三支决策思想应用于情感分类任务中获得了较好的实验效

到稿日期:2018-12-13 返修日期:2018-03-02 本文受国家自然科学基金(61572091),重庆市产业类重点主题专项(cstc2017zdcy-zdyfx0091),重庆市人工智能技术创新重大主题专项重点研发项目(cstc2017rgzn-zdyfx0022)资助。

张刚强(1993—),男,硕士生,主要研究方向为智能信息处理、情感分析,E-mail:tszqg2015@163.com;刘 群(1969—),女,博士,教授,主要研究方向为智能信息处理、复杂网络、数据挖掘,E-mail:liuqun@cqupt.edu.cn(通信作者);纪良浩(1977—),男,博士,教授,主要研究方向为复杂网络、智能信息处理。

果。文献[10-11]将三支决策的决策规则应用于情感词典分类中;文献[12]提出了一种将主题特征与三支决策理论相融合的多标记情感分类方法;文献[13]在多个文本粒度中研究了上下文有关、主题依赖和情绪分类的相关问题。以上相关研究是在单一粒度利用有限的信息或者尽量完整的情感信息进行情感极性类的三支决策,虽然取得了较好的实验结果,但是也存在一些问题:利用有限信息或者在单一粒度并不能够获得准确的分类结果;应用完整情感信息进行情感分类的成本较高。

字词是评论情感表达的最基本形式之一,也是机器学习方法选取特征的基本形式之一,但是仅考虑情感词难免会丢失掉评论中可能包含情感的重要信息,从而影响最终的情感分类结果。基于此,本文根据评论中表达情感信息量的多少,提出一种由粗到细的多粒度情感信息表示方法。对于评论数据激增的现状,如果能够在较粗粒度下进行正确的情感分类,将极大降低分类成本,提高分类效率;同时,随着粒度细化,所表达的情感的信息量逐渐增加,将为情感分类提供更多的依据,分类准确性也将进一步增加。因此,多粒度情感信息表示方法对于高效的情感分类将是一种有益的选择方案。受到序贯三支决策^[14](Sequential Three-way Decisions)思想的启发,本文从粒计算和认知学角度出发,将序贯三支决策的思想与多粒度文本情感信息表示方法相结合,提出了一种基于序贯三支决策的多粒度中文评论情感分类方法,该方法将多粒度的情感信息作为序贯三支决策的序贯粒,在每一粒度分别进行三支决策,最后根据序贯三支决策不同粒度下阈值的设置原则和决策成本实现评论最终的情感分类。实验结果表明,基于序贯三支决策的多粒度中文评论情感分类方法相比情感词典、贝叶斯和 SVM 情感分类方法具有更高的分类正确率和更强的鲁棒性。

本文第 2 节简要介绍了相关理论,包括粗糙集理论、序贯三支决策和粒计算;第 3 节详细阐述了文中提出的基于序贯三支决策的多粒度中文评论情感分类方法;第 4 节介绍了实验数据集和评价指标,并在 3 个数据集中对实验结果进行了分析与对比;最后总结全文并就下一步工作进行了展望。

2 序贯三支决策与粒计算

三支决策的思想被广泛应用于日常生活中,人们对于有充足证据证明接受或拒绝的事物能即刻做出接受或拒绝的判断,而对于不能立即做出决策的事物往往采取延迟决策。三支决策也是一种符合人类认知的决策模式,在医疗诊断、论文审查和管理科学等多个学科及研究领域中有十分广泛的应用。三支决策在认知计算时代将会扮演一个重要的角色,并产生深远的影响^[15]。

2.1 粗糙集理论

Pawlak^[16]提出的经典粗糙集模型是一种处理不确定性问题的方法,自 20 世纪 80 年代以来,在理论和应用中取得了很多研究成果。粗糙集和概率粗糙集^[17]的相关定义的具体描述如下。

定义 1 假设信息表 $M=(U, A_i, V_a, I_a)$ 是一个四元组,其中,有穷非空集合 U 是对象的论域, A_i 是所有属性的有穷

非空集合, V_a 是属性 $a \in A_i$ 取值的非空集合, $I_a:U \rightarrow V_a$ 是从 U 到 V_a 的映射函数。对于 $\forall X \subseteq U$, X 的下近似 $\underline{apr}(X)$ 和上近似 $\overline{apr}(X)$ 将论域划分为正域 $POS(X)$ 、边界域 $BND(X)$ 和负域 $NEG(X)$ 3 个互斥的部分,其定义分别如式(1)–式(3)所示:

$$POS(X)=\underline{apr}(X) \tag{1}$$

$$BND(x)=\overline{apr}(X)-\underline{apr}(X) \tag{2}$$

$$NEG(X)=U-\overline{apr}(X) \tag{3}$$

定义 2^[18] 在信息表中, (α, β) -上近似和下近似将论域划分为 (α, β) -正域 $POS_{(\alpha, \beta)}$ 、 (α, β) -边界域 $BND_{(\alpha, \beta)}$ 和 (α, β) -负域 $NEG_{(\alpha, \beta)}$ 3 个部分,其定义分别如式(4)–式(6)所示:

$$POS_{(\alpha, \beta)}(X)=\{x \in U | Pr(X|[x]) \geq \alpha\} \tag{4}$$

$$BND_{(\alpha, \beta)}(X)=\{x \in U | \beta < Pr(X|[x]) < \alpha\} \tag{5}$$

$$NEG_{(\alpha, \beta)}(X)=\{x \in U | Pr(X|[x]) \leq \beta\} \tag{6}$$

其中, $Pr(X|[x])$ 是分类的条件概率, $[x]$ 是包含 x 的等价类。

2.2 序贯三支决策

三支决策规则^[19]是为了给概率粗糙集的 3 个域提供一种合理的语义解释而提出的一种在不确定或不完整信息条件下的决策方式。对应于粗糙集的 3 个域构造的三支决策规则分别为接受、拒绝和不承诺规则。在实际问题中,选择不同的阈值会产生不同的决策结果,因此如何选择合理的阈值成为解决问题的关键。为此, Yao 等将贝叶斯决策过程引入概率粗糙集模型中,根据不同的决策成本选择总体风险最小的决策代价作为损失函数,但是三支决策没有考虑缺乏决策过程中获取和使用信息的代价。在现实问题中,获取解决问题的有效信息往往是一个递进过程,一般所获得的初始信息并不能支持其直接进行三支决策。基于此, Yao 等提出了序贯三支决策方法^[14],将三支决策作为序贯三支决策的一个中间过程,旨在获取信息成本最小化的同时取得更好的决策效果。

序贯三支决策的思想^[20]是,按从粗粒度到细粒度的顺序,在每一粒度中,当现有信息不足以支持做出接受或者拒绝的决策时,选择延迟决策,等下一粒度补充信息之后再再进行三支决策判断。而对于常见的二支决策问题,这一过程随着信息量的增加和决策成本的考虑,在合适的粒度获得二支决策的结果。三支决策理论是传统二支决策理论的拓展^[9]。二支决策和三支决策在解决实际问题时各有优缺点,二支决策适用于信息充足或者获取的信息代价较小的情况,但现实中可获取的初始信息往往不能支持决策者即刻做出合适的决策,对于情感分类问题尤其如此。

对评论情感分类问题而言,其最终结果是二支的,即需要指出一条评论属于“积极类”还是“消极类”。由于情感的表达具有模糊性、不确定性,因此如果直接强行进行二支决策并不能够获得好的分类效果。同样地,在单一粒度下所能获取的信息是有限的,无法直接得到准确的分类结果。三支决策为我们提供了一种折中的解决思路,可将不能完全确定情感类别的评论划分到边界域中。序贯三支决策针对边界域数据的决策设计了一种逐步计算的方式,在每一步增加与决策相关

的重要信息,即将单一粒度转换为更细粒度继续决策,直至做出最终决策,如果信息获取成本太大,则直接以二支决策得出结果,而这样的实现效果通常优于一般方法,且整体分类代价更低。

2.3 粒计算

粒计算(Granular Computing)是当前计算智能研究领域模拟人类思维和解决复杂问题的新方法^[21]。粒化和粒的计算是粒计算中的两个基本问题,其中粒化是基于有用信息将问题空间中的实例聚集成不同的粒,粒的计算是对具有某些概念粒的表示、特征化和描述解释。序贯三支决策采用了粒计算中逐步计算的实现方式^[14],因此构建合适的序贯信息粒对决策的正确性有很大的影响。粒计算作为信息处理的典型方法,能够很好地解决这个问题。

3 基于序贯三支决策的多粒度中文评论情感分类方法

3.1 多粒度情感信息表示方法

文本情感分类中的粒度划分一般包括词语级、句子级和篇章级,相关研究主要是在某一粒度进行特征的提取与选择,取得了较多的研究成果。但仅在某一较粗粒度,即所含情感信息较少的粒度下进行特征选择,并不能完全提取情感信息,使得分类正确率不高;而仅在最细粒度对所有特征进行选择,共同判定情感,则将导致效率低下。考虑到粒计算中“逐步计算”的思想,以评论中情感信息量的多少,由粗粒度到细粒度对评论数据集进行多个粒度的划分。

定义 3 假设对于任意评论对象 $x \in U$, 不同粒度的情感信息由不同的属性表达,情感属性子集 $A_i = \{A_1, A_2, \dots, A_n\}$ 是不同粒度下情感信息的具体描述,则评论对象 x 在不同情感信息粒度(Sentiment Information Granularity)的属性简记为 $SIGr_{A_i}(x)$ 。由粗粒度到细粒度的情感属性的序列表达如式(7)所示:

$$SIGr_{A_1}(x) \subset SIGr_{A_2}(x) \subset \dots \subset SIGr_{A_n}(x) \subset A_i \quad (7)$$

对于只有“积极类”和“消极类”的情感分类问题,评论对象在不同的情感信息粒度下,正域、边界域和负域之间存在如式(8)一式(10)的关系:

$$POS_{SIGr_{A_1}}(X) \subseteq POS_{SIGr_{A_2}}(X) \subseteq \dots \subseteq POS_{SIGr_{A_n}}(X) \quad (8)$$

$$BND_{SIGr_{A_n}}(X) \subseteq BND_{SIGr_{A_{n-1}}}(X) \subseteq \dots \subseteq BND_{SIGr_{A_1}}(X) \quad (9)$$

$$NEG_{SIGr_{A_1}}(X) \subseteq NEG_{SIGr_{A_2}}(X) \subseteq \dots \subseteq NEG_{SIGr_{A_n}}(X) \quad (10)$$

3.2 基于序贯三支决策的多粒度中文评论情感分类方法

基于序贯三支决策的多粒度中文评论情感分类方法的流程如下:

- 1) 对评论数据集进行预处理;
- 2) 首先结合评论数据的特点,根据评论中情感信息量的多少,对数据集进行由粗到细的粒度划分,获得由粗到细多个粒度的序贯情感表示;
- 3) 在较粗粒度下进行三支决策情感极性分类;
- 4) 对于较粗粒度进行三支决策之后划入边界域的评价对象,选择更细粒度的情感信息,继续进行三支决策,最后根据决策阈值获得情感分类结果。

具体流程如图 1 所示。

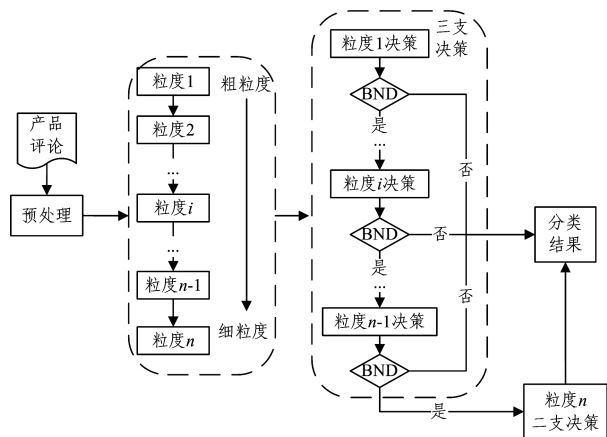


图 1 基于序贯三支决策的多粒度中文评论情感分类流程图

Fig. 1 Flowchart of multi-granularity sentiment classification method for Chinese reviews based on sequential three-way decisions

3.2.1 预处理

预处理部分主要是对评论数据集进行分词和停用词去除操作。本文主要应用 jieba 分词工具,其支持用户自定义词典,进行分词时能够获得较好的效果。

3.2.2 基于序贯三支决策的多粒度中文评论情感分类方法

在文本情感分类任务中,最常见的方法是基于某一粗粒度或者细粒度进行情感特征的提取,然后采用情感词典、规则或机器学习方法进行情感分类。这种方法太过简单,在粗粒度下进行情感分类时不能获得较好的分类效果,而在细粒度下进行较全的情感特征的提取之后,将会增加分类的成本。为了尽量节约分类成本,获得更好的情感分类效果,本文将序贯三支决策思想和多粒度情感信息表示方法相结合,提出了基于序贯三支决策的多粒度中文评论情感分类方法,即以多粒度的情感信息作为序贯三支决策的序贯粒,在考虑每一粒度决策阈值和成本的同时逐步计算获得最终的分类结果。

N-gram 语言模型是一种统计自然语言模型,常见的有 unigram, bigram, trigram 等模型。N-gram 模型根据前 $N-1$ 个词来预测第 N 个词,当前词 w_i 的概率只取决于前面 $i-1$ 个词,如式(11)所示:

$$P(w_i | w_1, w_2, \dots, w_{i-1}) = P(w_i | w_{i-N+1}, w_{i-N+2}, \dots, w_{i-1}) \quad (11)$$

情感信息的表达可以是多个粒度的,最简单的有词语粒度、短语粒度、句子粒度和评论粒度的划分。文本评论的情感表达由字、词和词组短语具体体现,每一粒度所表达的情感信息不尽相同。

N-gram 语言模型构成了一种多粒度情感信息的序贯表达。首先,以 unigram 作为最粗粒度的特征项进行三支决策,将情感信息不足以支持分类的评论归入到边界域。然后,以 bigram 作为更细粒度的特征项对这些边界域的评论进行三支决策,进一步将仍然不能分类的评论继续归入到边界域,以此方式构建序贯情感信息粒,逐步完成三支决策。在整个分类过程中,同时考虑决策成本和特征项选择的代价,以此为依据选择合适的粒度进行二支决策,最终获得评论情感分类的结果。

通过以上总结,本文将序贯三支决策的思想和多粒度情

感信息表达方式相结合,提出了基于序贯三支决策的多粒度中文评论情感分类算法,详细描述如算法 1 所示。

算法 1 基于序贯三支决策的多粒度中文评论情感分类方法 (MultiGr-S3DSC)

输入:中文评论数据集 $X=\{x_1,x_2,\cdots,x_n\}$,多粒度情感信息的属性 $SIGr_{A_i}(x)$,不同粒度的阈值对 $\{(\alpha_i,\beta_i,\gamma_i)\}$

输出:决策结果,即 POS 和 NEG

1. 初始化 $U_1=U,U_n=\emptyset,i=1,POS=\emptyset,NEG=\emptyset,BND=\emptyset$

2. for $i=1,2,\cdots,n-1$ do //三支决策过程

3. //对于 $\forall x\in X$,计算其在第 i 粒度的条件概率:

$$P(X|[x_i])=P(X)\frac{P([x_i]|X)}{P([x_i])}$$

4. $POS_{SIGr_{A_i}}=\{x\in U_i|P(X|[x_i])\geq\alpha_i\}$

5. $BND_{SIGr_{A_i}}=\{x\in U_i|\beta_i<P(X|[x_i])<\alpha_i\}$

6. $NEG_{SIGr_{A_i}}=\{x\in U_i|P(X|[x_i])\leq\beta_i\}$

7. $POS=POS\cup POS_{SIGr_{A_i}}$ //划入正域的评论

8. $NEG=NEG\cup NEG_{SIGr_{A_i}}$ //划入负域的评论

9. $U_{i+1}=BND_{SIGr_{A_i}}$ //划入边界域的评论

10. $i=i+1$

11. end for

12. if $U_n\neq\emptyset$ then //最细粒度二支决策过程

13. $POS_{SIGr_{A_n}}=\{x\in U_n|P(X|[x_i])\geq\gamma_n\}$

14. $NEG_{SIGr_{A_n}}=\{x\in U_n|P(X|[x_i])<\gamma_n\}$

15. $POS=POS\cup POS_{SIGr_{A_n}}$

16. $NEG=NEG\cup NEG_{SIGr_{A_n}}$

17. end if

3.2.3 序贯三支决策的阈值计算和设置原则

三支决策粗糙集模型利用 2 个状态集和 3 个行动集来描述决策过程。对于二支情感分类问题,状态集 $\Omega=\{X,X^C\}$ 分别表示某一评论对象属于“积极类”和“消极类”,行动集 $A=\{a_p,a_b,a_n\}$ 分别表示对某一评论对象采取接受、延迟和拒绝行动,即分别将某一对象划入正域 $POS(X)$ 、边界域 $BND(X)$ 和负域 $NEG(X)$ 中。由于评论对象 x 在不同状态下采取不同的行动会产生不同的损失, λ_{pp} , λ_{bp} 和 λ_{np} 分别表示对象 x 属于 X 时采取 a_p,a_b 和 a_n 3 种行动时的损失函数; λ_{pn} , λ_{bn} 和 λ_{nn} 分别表示对象 x 属于 X^C 时采取 a_p,a_b 和 a_n 3 种行动时的损失函数。

根据贝叶斯决策准则,需要选择期望损失最小的行动集作为最佳行动方案。根据采取 a_p,a_b,a_n 3 种行动时的期望损失可得到定义 2 中 3 个域的划分,其中:

$$\alpha=\frac{(\lambda_{pn}-\lambda_{bn})}{(\lambda_{pn}-\lambda_{bn})+(\lambda_{bp}-\lambda_{pp})} \tag{12}$$

$$\beta=\frac{(\lambda_{bn}-\lambda_{nn})}{(\lambda_{bn}-\lambda_{nn})+(\lambda_{np}-\lambda_{bp})} \tag{13}$$

$$\gamma=\frac{(\lambda_{pn}-\lambda_{nn})}{(\lambda_{pn}-\lambda_{nn})+(\lambda_{np}-\lambda_{pp})} \tag{14}$$

三支决策是一种符合人类认知的决策模式,在具体的决策过程中对不同的域采取不同的策略。文献[14]指出了序贯三支决策在计算过程中的一条基本法则,即只有在必须或者有益于决策的情况下才进行更细粒度的决策。这为不同粒度下三支决策阈值的设置提供了依据,即粗粒度选择更大的接受阈值和更小的拒绝阈值。 $i=1,2,\cdots,n$ 表示由粗粒度到细粒度的序列,则不同粒度的阈值的具体描述如式(15)所示:

$$0\leq\beta_i<\alpha_i\leq 1,1\leq i\leq n,\beta_1\leq\beta_2\leq\cdots\leq\beta_i<\alpha_i\leq\cdots\leq\alpha_2\leq\alpha_1 \tag{15}$$

在三支决策粗糙集模型中引入了损失函数的概念,依据贝叶斯风险最小决策理论,可以计算出每一粒度的决策阈值,损失函数通常是具体情况依据经验给定的。本文算法 (MultiGr-S3DSC)依据序贯三支决策阈值的设置原则,初始最粗粒度采取不同决策时的损失,同时考虑到随着粒度的细化需要获取的信息的成本也在增加,更细粒度即使做出正确的决策,也给予一定的损失。由于三支决策解决的是二分类问题,因此在逐步计算的过程中,给两个类别正确划分到正域、边界域和负域的对象以同样的惩罚方式,对从粗粒度到细粒度的正域、边界域和负域的损失函数设置了一种与粒度相关的线性递增函数。即随着粒度的细化,正域、边界域和负域的损失分别第 1 粒度的基础上增加 $(n-1)$, $2(n-1)$ 和 $3(n-1)$,其中 $n>1$ 表示当前粒度,以此作为序贯三支决策中每一粒度的阈值。本文实验是在 3 个情感信息粒度下进行的,在不同粒度采取不同决策的一组损失函数的详细信息如表 1 所列,其中 μ 是单位损失。

表 1 不同粒度采取不同决策时的损失

Table 1 Losses of different granularities employing different decisions

粒度	损失					
	λ_{pp}	λ_{bp}	λ_{np}	λ_{pn}	λ_{bn}	λ_{nn}
第 1 粒度	0	2μ	11μ	9μ	3μ	0
第 2 粒度	1μ	4μ	14μ	12μ	5μ	1μ
第 3 粒度	2μ	6μ	17μ	15μ	7μ	2μ

由式(12)一式(14)以及表 1 中的损失可以计算出每一粒度的决策阈值对 $\{\alpha_i,\beta_i,\gamma_i\}$,第 1 粒度的阈值对 $\alpha_1=0.75$, $\beta_1=0.25$,这符合序贯三支决策阈值的设置原则。

利用贝叶斯理论,评论对象 x 属于“积极类” X 和属于“消极类” X^C 的条件概率分别如式(16)和式(17)所示:

$$P(X|[x])=\frac{P(X)P([x]|X)}{P([x])} \tag{16}$$

$$P(X^C|[x])=1-P(X|[x]) \tag{17}$$

其中, $P([x])$ 可由全概率公式得到,如式(18)所示:

$$P([x])=P(X)P([x]|X)+P(X^C)P([x]|X^C) \tag{18}$$

4 实验与分析

4.1 实验数据集和评估方法

为验证本文所提方法的性能,将电脑评论(Notebook)、书评(Book)和酒店评论(Hotel)作为 3 个实验数据集,每个数据集中正、负类各 2000 条。

对于只有积极和消极两个类别的情感分类问题,根据其真实类别与预测类别的组合结果,形成如表 2 所列的分类结果混淆矩阵。其中 TP,FP,TN,FN 分别表示真正例、假正例、真反例和假反例对应的样本数。查准率 P 、查全率 R 、 $F1$ 值和正确率 Acc 的定义分别如式(19)一式(22)所示:

$$P=\frac{TP}{TP+FP} \tag{19}$$

$$R=\frac{TP}{TP+FN} \tag{20}$$

$$F1=\frac{2PR}{P+R} \tag{21}$$

$$Acc=\frac{TP+TN}{TP+FN+FP+TN}$$

(22)

表 2 分类结果的混淆矩阵

Table 2 Confusion matrix of classification results

真实情况	预测结果	
	Positive	Negative
Positive	TP(True Positive)	FN(False Negative)
Negative	FP(False Positive)	TN(True Negative)

4.2 多粒度特征提取与降维

N-gram 语言模型在英文处理时以单词作为特征,若直接以 unigram,bigram 等方式提取中文评论特征并不能获得较好的实验效果。例如:“这台笔记本的性价比高”根据 unigram 提取的候选特征为“‘这’‘台’‘笔’‘记’‘本’‘的’‘性’‘价’‘比’‘高’”;根据 bigram 提取的候选特征为“‘这台’‘台笔’‘笔记’‘记本’‘本的’‘的性’‘性价’‘价比’‘比高’”。从 unigram 和 bigram 提取的候选特征中可以看出,并没有多少直接能够表达评论情感的候选特征,且这与一般的先对中文进行分词处理的方式有所不同。

由于中文评论中情感的表达大多是针对产品的某一特性利用情感词和短语来进行表达的,因此本文在实验中以分词后的词语作为基本的特征单元,同时,依据 3.1 节提出的多粒度情感信息表示方法,构建了 3 个粒度的中文评论情感候选特征,即一元词、双词词组以及一元词和双词组合作为序贯三支决策的 3 个粒度来提取评论中的情感信息,以此构建中文评论的序贯特征。例如:首先对“这台笔记本的性价比高”进行中文分词,可得“‘这台’‘笔记本’‘的’‘性价比’‘高’”,则该例的 3 个粒度的候选特征的表示形式如表 3 所列。其中,第 1 粒度是例子所有的一元词候选特征,在第 1 粒度能够提取到直接表达情感的情感词;第 2 粒度是例子所有双词词组形式的候选特征,在第 2 粒度能够提取到评论中表达情感的情感短语;第 3 粒度是所有一元词和双词组合的候选特征。

表 3 评论的多粒度候选特征表示

Table 3 Multi-granularity candidate feature representation of reviews

粒度序列	提取的候选特征				
第 1 粒度	‘这台’	‘笔记本’	‘的’	‘性价比’	‘高’
第 2 粒度	‘这台 笔记本’ ‘笔记本 的’ ‘的 性价比’ ‘性价比 高’				
第 3 粒度	‘这台’	‘笔记本’	‘的’	‘性价比’	‘高’
	‘这台 笔记本’ ‘笔记本 的’ ‘的 性价比’ ‘性价比 高’				

由于随粒度的增加评论的特征数量也在增加,且分词后的评论中有一些像‘的’这样的无用特征词,因此为了加快计算速度以及提高分类准确率,文中在去除停用词之后采用 χ^2 统计量 CHI(Chi-square Statistic)衡量特征项和类别之间的关联程度,寻找情感信息量丰富的特征来减少特征的数量。例如候选特征表中“性价比 高”可以为判定为“积极类”的评论提供更多的情感决策依据。本文在实验中使用文本分类中常用的词频 TF 权重算法计算各特征词的权重值。

由图 2 可以看出,对于中文评论数据集,相对于直接以 unigram,bigram 和 unigram+bigram 作为 3 个粒度特征,分词后以词语、双词词组和两者组合分别作为 3 个粒度特征的方式能够得到更好的实验效果,这表明文中提出的多粒度情感信息表示方法是有效的。

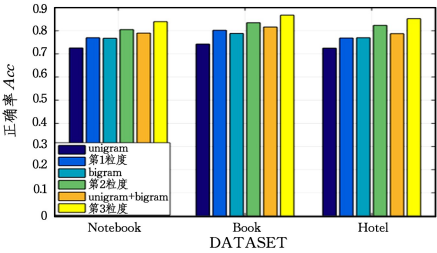


图 2 贝叶斯分类器下 N-gram 模型和多粒度情感信息表示方法的性能比较

Fig. 2 Performance comparison of N-gram model and multi-granularity sentiment information representation method under Bayesian classifier

4.3 实验结果的对比与分析

4.3.1 实验阈值的确定

序贯三支决策的阈值设置的原则是随着粒度的细化,选择更小的接受阈值和更大的拒绝阈值。本文基于 3.2.3 节中关于阈值的设置原则,选择不同的初始损失函数分别计算 3 个粒度的阈值对。图 3 是以第 1 粒度不同阈值对 α, β 为坐标轴时对实验最终分类正确率散点的拟合曲线。

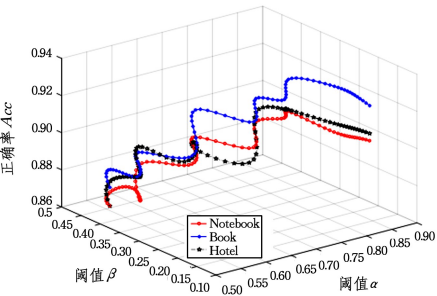


图 3 第 1 粒度在不同阈值对下的正确率对比

Fig. 3 Comparison of precision of first granularity under different thresholds

由图 3 可知,在电脑评论和酒店评论两个数据集中,当第 1 粒度的两个阈值 α, β 分别为 0.75 和 0.25 时能够获得更高的分类正确率;在书评数据集中当第 1 粒度的阈值对 α, β 分别为 0.8 和 0.21 时能取得最好的实验结果。

图 4 和图 5 分别是以第 1 粒度不同阈值对 α, β 为坐标轴时,对“积极类”数据集的查准率和查全率的实验结果的散点拟合曲线。由图 4 和图 5 可知,当实验获得较高的查准率时,查全率往往偏低,这是由于查准率和查全率是一对矛盾的度量,当查准率较高时查全率较低,当查全率较高时查准率较低。

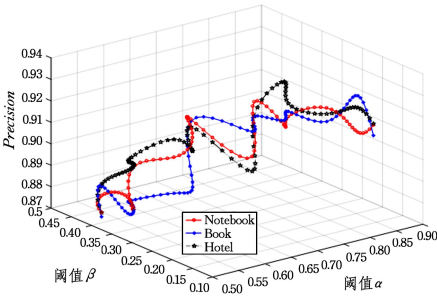


图 4 第 1 粒度在不同阈值对下“积极类”的查准率对比

Fig. 4 Comparison of precision of positive class for first granularity under different thresholds

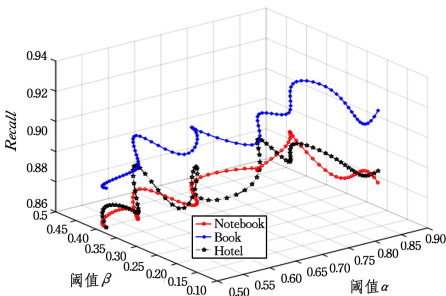


图5 第1粒度在不同阈值对下“积极类”的查全率对比

Fig.5 Comparison of recall of positive class for first granularity under different thresholds

通过上述多组实验可以发现,本文中关于序贯三支决策损失函数的设置是合理的。

4.3.2 不同方法实验结果的对比与分析

本文选取已有情感分类算法,包括基于情感词典与规则(Lexicon)的方法^[5]、SVM^[6]和贝叶斯^[8]情感分类方法进行实验对比,同一数据集的不同方法的正确率比较结果如表4所列。

表4 正确率 Acc 的对比结果

Table 4 Comparison results of Acc

方法	电脑评论	书评	酒店评论
Lexicon	0.738	0.728	0.762
NB	0.788	0.816	0.788
SVM	0.883	0.866	0.847
本文方法	0.896	0.930	0.897

由表4可知,相比于SVM情感分类方法,在电脑评论、书评和酒店评论3个数据集中,本文方法的正确率分别提高了1.3%,6.37%和5%,这说明本文方法具有更高的正确率且更有效,将多粒度情感信息表示方法与序贯三支决策思想相结合能对中文评论实现很好的情感分类。

由表5和表6可以看出,在书评和酒店评论数据集中,与其他几种情感分类方法相比,本文方法在查准率和查全率两个分类标准上都获得了较高值。在电脑评论数据集中,相比SVM情感分类方法,本文方法的“积极类”的查准率和“消极类”的查全率并不都是最高的,但总体来讲,本文方法相比其他几种情感分类方法还是获得了更高的查准率和查全率。且由表7可知,本文方法相对于其他经典的情感分类方法在“正类”和“负类”评论中均具有更高的F1值。

表5 查准率 P 的对比结果

Table 5 Comparison results of P

方法		电脑评论	书评	酒店评论
Lexicon	Positive	0.7066	0.6688	0.7227
	Negative	0.7807	0.7879	0.8175
NB	Positive	0.8303	0.7935	0.7869
	Negative	0.7553	0.8428	0.7892
SVM	Positive	0.9101	0.8805	0.8343
	Negative	0.8593	0.8530	0.8607
本文方法	Positive	0.9041	0.9151	0.9026
	Negative	0.8882	0.9461	0.8915

表6 查全率 R 的对比结果

Table 6 Comparison results of R

方法		电脑评论	书评	酒店评论
Lexicon	Positive	0.814	0.832	0.8495
	Negative	0.662	0.624	0.674
NB	Positive	0.724	0.855	0.79
	Negative	0.852	0.7775	0.786
SVM	Positive	0.85	0.8475	0.868
	Negative	0.916	0.885	0.828
本文方法	Positive	0.886	0.948	0.89
	Negative	0.906	0.912	0.904

由表4—表7可知,本文提出的多粒度情感信息表示方法与序贯三支决策思想相结合能够获得更高的分类正确率和F1值,且能够保证较高的查准率和查全率,这说明本文方法对于中文评论情感分类问题而言是可行且有效的。

表7 F1 值的对比结果

Table 7 Comparison results of F1

方法		电脑评论	书评	酒店评论
Lexicon	Positive	0.7565	0.7415	0.781
	Negative	0.7165	0.6964	0.7388
NB	Positive	0.7735	0.8231	0.7884
	Negative	0.8007	0.8088	0.7876
SVM	Positive	0.879	0.8637	0.8508
	Negative	0.8867	0.8687	0.844
本文方法	Positive	0.895	0.9313	0.8963
	Negative	0.897	0.9287	0.8977

4.3.3 跨领域情感分类正确率对比与分析

为了验证模型的鲁棒性,对本文方法进行了跨领域情感分类的实验验证,当第1粒度的阈值对 α, β 分别为0.8和0.21时取得了最好的实验结果,不同分类方法的分类正确率对比如表8所列,其中 $B \rightarrow H$ 表示源领域数据是书评(Book),目标领域数据是酒店评论(Hotel),其他表示以此类推,其中,对比实验结果来自文献[22]和文献[23]。

表8 不同分类方法跨领域情感分类正确率 Acc 的对比结果

Table 8 Comparison of Acc of different classification methods for interdisciplinary sentiment

领域	SVM	文献[22]	文献[23]	本文方法
$B \rightarrow H$	0.5488	0.790	0.8598	0.8613
$B \rightarrow N$	0.5625	0.776	0.8098	0.8448
$H \rightarrow B$	0.600	0.683	0.8095	0.8178
$H \rightarrow N$	0.7225	0.784	0.8488	0.8548
$N \rightarrow B$	0.5488	0.650	0.7418	0.7465
$N \rightarrow H$	0.7250	0.791	0.8643	0.8520

由表8可知,对于跨领域的情感分类问题,与文献[22]中提出的两阶段的分类框架和文献[23]中提取情感关键句的方法相比,除了电脑评论作为源域、酒店评论作为目标域这一组实验的分类正确率稍低于文献[23]的结果外,在其他5个跨领域实验中本文方法均获得了更好的分类正确率。特别是当源域为书评、目标域为电脑评论时本文方法的分类正确率比文献[23]提高了3.5%,这说明本文中的多粒度情感分类方法优于提取情感关键句方法。该组实验也再次验证了基于序贯三支决策的多粒度情感分类方法对中文评论数据集进行跨

领域情感分类的有效性,也证明了本文方法对于中文评论情感分类具有更强的鲁棒性。

结束语 本文从认知学的角度出发,结合粒计算思想,对评论中的情感信息进行了多粒度的划分,并将序贯三支决策的思想应用于文本情感分类任务中。通过分析评论中情感信息粒之间的序贯性,构建了一种基于序贯三支决策的多粒度情感分类方法。实验结果表明,所提出的情感分类方法具有更强的鲁棒性,在 3 个数据集中获得了较高分类正确率和 $F1$ 值的同时也保证了更高的查准率、查全率,且在跨领域情感分类中也获得了更高的分类正确率。本文只是初步探索了序贯三支决策思想在情感分类问题中的应用,下一步工作将基于本文提出的多粒度情感信息表示方法,构建其他形式更加一般的序贯情感信息粒,对情感极性强度进行判别。

参 考 文 献

[1] LIU B. Sentiment Analysis and Opinion Mining[J]. Synthesis Lectures on Human Language Technologies, 2016, 30(1): 152-153.

[2] CAMBRIA E. Affective Computing and Sentiment Analysis[J]. IEEE Intelligent Systems, 2016, 31(2): 102-107.

[3] MEDHAT W, HASSAN A, KORASHY H. Sentiment analysis algorithms and applications; A survey[J]. Ain Shams Engineering Journal, 2014, 5(4): 1093-1113.

[4] RANA T A, CHEAH Y N. Aspect extraction in sentiment analysis; comparative analysis and survey[J]. Artificial Intelligence Review, 2016, 46(4): 459-483.

[5] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2): 267-307.

[6] ZOU H, TANG X, XIE B, et al. Sentiment Classification Using Machine Learning Techniques with Syntax Features[C]//International Conference on Computational Science and Computational Intelligence. IEEE Computer Society, 2015: 175-179.

[7] AGARWAL B, MITTAL N. Machine Learning Approach for Sentiment Analysis[M]//Prominent Feature Extraction for Sentiment Analysis. Springer International Publishing, 2016: 21-45.

[8] TRIPATHY A, AGRAWAL A, RATH S K. Classification of Sentiment Reviews using N-gram Machine Learning Approach [J]. Expert Systems with Applications, 2016, 57(C): 117-126.

[9] YAO Y Y. An Outline of a Theory of Three-Way Decisions [C]//International Conference on Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012: 1-17.

[10] ZHOU Z, SHANG L. A sentiment analysis method based on dynamic lexicon and three-way decision [J]. Journal of Shandong University (Engineering Science), 2015, 45(1): 19-23. (in Chinese)

周哲,商琳.一种基于动态词典和三支决策的情感分析方法[J]. 山东大学学报(工学版), 2015, 45(1): 19-23.

[11] ZHOU Z, ZHAO W, SHANG L. Sentiment Analysis with Automatically Constructed Lexicon and Three-Way Decision[M]//Rough Sets and Knowledge Technology. Springer International Publishing, 2014: 777-788.

[12] WANG L, HUANG H X, WU B, et al. Emotion analysis of text based on topics and three-way decisions [J]. Computer Science, 2015, 42(6): 93-96. (in Chinese)

王磊,黄河笑,吴兵,等.基于主题与三支决策的文本情感分析[J]. 计算机科学, 2015, 42(6): 93-96.

[13] ZHANG Z, WANG R. Applying Three-way Decisions to Sentiment Classification with Sentiment Uncertainty[C]//International Conference on Rough Sets and Knowledge Technology. Springer International Publishing, 2014: 720-731.

[14] YAO Y Y, DENG X. Sequential three-way decisions with probabilistic rough sets[C]//IEEE International Conference on Cognitive Informatics & Cognitive Computing. IEEE, 2011: 120-125.

[15] YAO Y Y. Three-Way Decisions and Cognitive Computing[J]. Cognitive Computation, 2016, 8(4): 543-554.

[16] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[17] YAO Y Y. Probabilistic approaches to rough sets[J]. Expert Systems, 2003, 20(5): 287-297.

[18] LIU D, YAO Y Y, LI T R. Three-way decisions-theoretic rough sets [J]. Computer Science, 2011, 38(1): 246-250. (in Chinese)

刘盾,姚一豫,李天瑞.三枝决策粗糙集[J]. 计算机科学, 2011, 38(1): 246-250.

[19] YAO Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180(3): 341-353.

[20] YAO Y Y. Granular Computing and Sequential Three-Way Decisions [C]//International Conference on Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2013: 16-27.

[21] WANG G Y, ZHANG Q H, HU J. An overview of granular computing[J]. CAAI Transactions on Intelligent Systems, 2007, 2(6): 8-26. (in Chinese)

王国胤,张清华,胡军.粒计算研究综述[J]. 智能系统学报, 2007, 2(6): 8-26.

[22] WU Q, TAN S. A two-stage framework for cross-domain sentiment classification[J]. Expert Systems with Applications, 2011, 38(11): 14269-14275.

[23] ZHANG S, LIU H, YANG L, et al. A Cross-Domain Sentiment Classification Method Based on Extraction of Key Sentiment Sentence[M]//Natural Language Processing and Chinese Computing. Springer International Publishing, 2015: 90-101.