基于三支决策的多粒度文本情感分类模型

张越兵[1,2] 苗夺谦[1,2] 张志飞[1,3]
(同济大学计算机科学与技术系 上海 201804)[1]
(同济大学嵌入式与服务计算教育部重点实验室 上海 201804)[2]
(同济大学大数据与网络安全研究中心 上海 200092)[3]

摘 要 文本情感分类是一项重要的自然语言处理任务,具有广泛的应用场景。以往的情感分类方法过于注重分类准确率,忽略了训练和分类过程的时间代价,而且使用的特征大多为词袋特征,存在维度高、可解释性差的缺点。针对这些问题,将粒计算的思想运用于文本数据的三层粒度结构(词-句-篇章),提出一种具有强可解释性的文本情感分类特征——SSS(Sentence-level Sentiment Strength)特征,SSS 特征每一维度代表文章中每个句子的情感强度值;同时,在分类过程中,利用三支决策方法将待分类对象划分为 3 个区域,位于正域和负域的对象直接划分至正类和负类中,使用 SVM(Support Vector Machine)+SSS 特征对位于边界域的对象做进一步分类。实验结果显示,SSS 特征由于自身的低维特性,能够大大降低特征提取和模型训练过程所耗费的时间成本,结合了三支决策方法的 SVM 能够进一步提高分类准确率,而且三支决策方法可以减少分类过程所耗费的时间。

关键词 情感分类,三支决策,多粒度,支持向量机
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.035

Multi-granularity Text Sentiment Classification Model Based on Three-way Decisions

ZHANG Yue-bing[1,2] MIAO Duo-qian[1,2] ZHANG Zhi-fei[1,3]
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)[1]
(Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)[2]
(Research Center of Big Data and Network Security, Tongji University, Shanghai 200092, China)[3]

Abstract Text sentiment classification is a very important branch of natural language processing. Researchers focus on the accuracy of sentiment classification but ignore the time cost of training and classification. Bag-of-words feature used in most methods for text sentiment classification has high dimension and bad interpretability. To solve the above problems, we presented a multi-granularity text sentiment classification model based on three-way decisions for document-level sentiment classification. With the aid of granular computing, we made a structure of text that contains three levels of granularity — word, sentence and document, and presented a new kind of feature — SSS (sentence-level sentiment strength) feature which represents a document, in which the value of each dimension is the sentence-level sentiment strength. In classification process, we firstly utilized three-way decisions method to divide the objects into three regions. The objects in positive region and negative region are classified into positive class and negative class, respectively. We employed the state-of-the-art classifier — SVM to classify the objects in boundary region. Experimental results show that combining three-way decisions method and SVM can improve the accuracy of classification. The SSS-feature reduces the time-cost of feature extraction and training greatly because of its low dimension. Three-way decisions method can reduce the time-cost of classification, and they can ensure good performance in classification accuracy at the same time.

Keywords Sentiment classification, Three-way decisions, Multi-granularity, SVM

到稿日期:2017-01-11 **返修日期**:2017-03-08 **本文受国家自然科学基金**(61273304, 61673301),**高等学校博士学科点专项科研基金**(20130072130004)**资助**。

张越兵(1991—),**男**,**博士生**,CCF **学生会员**,**主要研究方向为机器学习、自然语言处理、文本情感分析**,E-mail:yuebing_zhang@hotmail.com;**苗夺谦**(1964—),**男**,**博士**,**教授**,CCF **会员**,**主要研究方向为粗糙集、粒度计算、数据挖掘、机器学习**,E-mail:dqmiao@tongji.edu.cn;**张志飞**(1986—),**男**,**博士**,**讲师**,CCF **会员**,**主要研究方向为机器学习、自然语言处理、文本情感分析**,E-mail:zhifeizhang@tongji.edu.cn。

1　引言

文本情感分析（亦称观点挖掘[1]）是自然语言处理领域的一个研究热点[2-3]，其主要任务是从文本数据中挖掘出人类的观点和情感。根据文本语言的粒度，可以将文本情感分类问题分为词、句、篇章 3 个层次[4]。关于文本情感分类的现有研究工作主要有两个研究方向[5]：1）基于词典的方法，主要利用情感词典来标注文本情感极性或情感强度[2,6]；2）基于语料的方法，将文本情感分类问题作为文本分类问题的一种特例[7]，并使用机器学习方法从文本中提取出合理的特征来完成分类问题[3]。文本情感分类问题除了包含二元分类任务（积极或消极）外，还包含多元分类任务。

前人的研究重点在于如何提高分类的准确率，然而针对互联网大规模文本数据的文本情感分类任务不可避免地会面临算法执行的时间成本问题。若要在可接受时间内完成大数据的分类任务，则要求算法模型具有较高的运行效率。以往提出的情感分类方法所使用的特征大部分为词袋特征，词袋特征维度高，限制了模型训练和分类的效率；此外，词袋特征仅考虑文本中出现的词语的统计特性，可解释性较差。针对上述问题，本文提出一种处理篇章级文本的多粒度情感分类模型（下文简称为 TWD-SSS）。TWD-SSS 结合了基于词典的方法和基于语料库的方法，从词-句-篇章 3 个粒度来处理文本数据。TWD-SSS 的主体分为两部分：在词-句层次，利用通用情感词典、领域情感词典以及由词向量训练的多层感知机分类器发现情感词语，再结合语法知识提取句子情感强度，为后续构建机器学习方法的 SSS 特征（Sentence-level Sentiment Strength Feature）做准备；在句-篇章层次，引入作者写作时的情感走向模型，根据文章中每一个句子的情感强度值计算获得文章的情感强度值，并将其作为后续三支决策的判定依据。三支决策方法将待分类的对象划分为 3 个区域，将位于正域和负域的对象直接划分至正类和负类中，利用 SVM 分类器和 SSS 特征对位于边界域的对象进一步分类。经实验验证，TWD-SSS 模型在中英文文本上的表现都优于其他主流分类模型。

2　相关工作与基础知识

2.1　基于词典的方法

基于词典的方法通常采用已存在的情感词典来完成情感分类任务。情感词典是一种特殊的词典，其中标注了词语的情感极性或情感强度。情感词典又分为通用情感词典和领域情感词典两种：通用情感词典适用于绝大部分领域的情感分类任务，其中包括 WordNet 和 HowNet；领域情感词典仅适用于特定领域的情感分类任务。Turney[8] 于 2002 年提出一种具有代表性的基于词典的无监督学习方法，将评论数据进行二元分类，将一条评论中所有词语的情感倾向的平均值作为当前评论情感分类的指标。Ding[9] 利用否定词（例如“not”“never”等）和转折词（例如“but”）来提高基于词典方法的准确率。Thelwall 等[10] 提出使用情感词典和语法规则来计算推特（Twitter）的情感强度。同一词语在不同语境中可能具有不同的情感极性，这是基于词典的方法中错误分类的主要来源，对此 Heeryon 等[11] 提出了一套数据驱动的方法使情感词典适应不同的领域，这套方法的核心思想是通过比较正负类评论中情感词出现的比例和正负类评论自身的比例来判定该删除哪些情感词或者修改其情感极性。

2.2　基于语料库的方法

Pang 等[7] 首次使用机器学习方法（包括朴素贝叶斯、最大熵模型和支持向量机）实现文本情感分类，实验结果表明支持向量机（SVM）＋词袋特征（Bag-of-words）的效果最优。在该工作的基础上，许多研究者将重点放在有效特征的发现和设计上。Katz 等[12] 提出了一种基于语境的方法 ConSent，其在无噪声和有噪声的文本上均有效。Franco-Salvador 等[13] 使用元学习（meta-learning）将几种主流方法结合，其中包括词袋模型、n-grams 模型和基于词典资源的分类方法，旨在解决跨领域的文本情感分类问题。Zhang 等利用粗糙集，从否定句[14] 和强语义模糊性词语[15] 的角度研究文本情感分类问题。Das 等[16] 提出情感分类模型 DS，将否定词“NOT”附在句子否定范围之后，例如“The book is not interesting”会被转换成“The book is interesting-NOT”。Li 等[17] 提出将每段文本分为极性转换（polarity-shifted）和极性非转换（polarity-unshifted）两部分，针对两部分分别训练分类器并组合用于情感分类任务。

2.3　三支决策

三支决策理论是由粗糙集理论发展而来的一种新的决策思想[18]，近年来已经受到了各领域学者的广泛关注。相比于二支决策而言，三支决策的处理方法更合理：当对象提供的信息不足但以支撑决策时采用延迟决策，以等待更多信息来完成最终的决策。因此，三支决策可以规避分类信息不足但却盲目决策造成的风险。在决策粗糙集的公式化表达中，$X\subseteq U$ 是全集的子集，状态集合可以表示为 $\Omega=\{X,X^c\}$，X 和 X^c 分别表示一个对象属于 X 和不属于 X。为了方便，子集和其对应的状态都用 X 表示。状态 X 对应的动作集合表示为 $\wedge=\{P,B,N\}$，其中 P,B,N 分别表示判定对象的 3 种动作，即 $x\in POS(X)$，$x\in BND(X)$，$x\in NEG(X)$。三支决策的损失函数由各个状态下的行为风险决定，如表 1 所列，其中 λ_{PP}，λ_{BP} 和 λ_{NP} 分别表示当对象属于子集 X 时采取动作 P,B,N 产生的损失，λ_{PN}，λ_{BN} 和 λ_{NN} 分别表示当对象不属于子集 X 时采取动作 P,B,N 产生的损失。

表 1　三支决策的损失函数

	$X(P)$	$X^c(N)$
P	λ_{PP}	λ_{PN}
B	λ_{BP}	λ_{BN}
N	λ_{NP}	λ_{NN}

根据最小风险决策规则[19] 得：

（P）当 $P_r(X\mid[x])\geqslant\alpha$ 时，$x\in POS(X)$；

（B）当 $\beta<P_r(X\mid[x])<\alpha$ 时，$x\in BND(X)$；

（N）当 $P_r(X\mid[x])\leqslant\beta$ 时，$x\in NEG(X)$。

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$\beta=\frac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

其中，$0\leqslant\beta<\alpha\leqslant1$。

3 TWD-SSS 模型

本文提出的 TWD-SSS 模型由两步组成：第一步使用基于词典的方法获取句子的情感强度，进一步构造 SSS 特征用于文档的情感分类；第二步将文档的 SSS 特征放入由三支决策和支持向量机组成的分类器中，完成文档的情感分类任务。整体框架如图 1 所示。

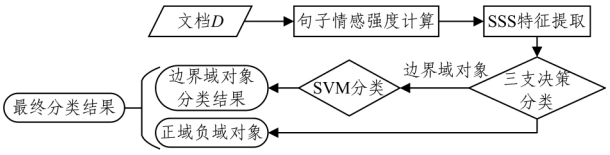

图 1 TWD-SSS 模型的整体框架图

3.1 句子的情感强度标注

3.1.1 情感词语的发现

情感词语的发现包括 3 个部分：

1)使用通用情感词典 HowNet[1] 发现情感词语；

2)使用领域情感词典(使用 Prop-dep[20] 在语料库中训练得到)发现情感词语；

3)根据现有的情感词典训练一个多层感知机(Multi-layer perceptron)分类器，对于单纯依赖情感词典没有提取到情感词的句子，进一步发现其情感词语。

步骤 2)使用 Prop-dep 训练得到的领域情感词典在一定程度上缓解了因情感词典覆盖不足而造成的情感信息缺失问题，然而在实验过程中发现，仍然有一定比例的句子没有被检测出任何情感词。为了进一步解决因情感词典覆盖不足造成的情感缺失问题，将步骤 3)作为补充。步骤 3)中的多层感知机结构为 3 层，其中输入层为 100 个神经元，隐藏层为 20 个神经元，输出层(softmax 层)为 2 个神经元。利用 word2vec[2](采用 skip-gram 模型，词向量维度为 100，训练的窗口大小为 5)训练维基百科语料[3] 得到词向量，查询现有情感词典中情感词语对应的词向量，用作训练数据的输入值。对于单纯依赖情感词典没有提取到情感词语的句子，将其中所有词语对应的词向量作为测试数据的输入值。假设输出层的 2 个输出值为 O_p 和 O_n，最终的决策规则为：

(P)当 $O_p-O_n\geqslant\gamma$ 时，输入词语被判定为正极性情感词；

(N)当 $O_n-O_p\geqslant\gamma$ 时，输入词语被判定为负极性情感词；

(B)当 $\gamma>O_p-O_n>-\gamma$ 时，输入词语被判定为中性词。

其中，γ 设定为 0.2。

3.1.2 句子情感强度的提取

将文档分割成若干句子，每个句子作为一个独立的对象来提取情感强度。句子情感强度的计算公式如下：

$$S(i)=\mathrm{Tr}\left(\frac{\sum neg(w(Pos,i))-\sum neg(w(Neg,i))}{NW(i)}\right) \quad (1)$$

其中，$S(i)$表示句子 i 的情感强度值。句子情感强度提取的步骤如下：1)扫描句子中出现在通用情感词典和领域情感词典中的情感词，如果通过情感词典没有检测出情感词，则使用上述多层感知机分类器对句子中的所有词语进行分类，将分类结果最明确的若干个词作为该句的情感词。式(1)中的 $w(Pos,i)$和 $w(Neg,i)$分别表示句子 i 通过此步骤检测出的正极性词语和负极性词语。

2)使用 Stanford Parser 对句子进行句法分析，判断情感词是否存在否定修饰。式(1)中的 $neg(w(Pos/Neg,i))$表示情感词语经过否定修饰后的情感值，$neg(w(Pos,i))$的计算方法如下：

$$neg(w(Pos,i))=\begin{cases}-1, & w(Pos,i)\text{存在否定修饰}\\ 1, & w(Pos,i)\text{不存在否定修饰}\end{cases} \quad (2)$$

$neg(w(Neg,i))$的计算方法同理。

3)对经过否定修饰判定的情感词进行计数，并除以当前句子的总词数，以融入句子的长度信息。式(1)中 $NW(i)$表示句子 i 中的总词数。

4)考虑句际关系，对于存在转折关系的前后两个句子，弱化前面句子的情感强度。式(1)中 $Tr(S_i)$表示对存在转折关系的句子计算情感强度，其计算方法如下：

$$Tr(S_i)=\begin{cases}\delta S_i, & S_i\text{和}S_{i+1}\text{存在转折关系}\\ S_i, & S_i\text{和}S_{i+1}\text{不存在转折关系}\end{cases} \quad (3)$$

其中，δ 为 0～1 的常量。

经过以上 4 个步骤得到句子情感强度值 $S(i)$，用作后续 SSS 特征的生成和三支决策的划分依据。

3.1.3 SSS 特征的构造

根据粒计算的思想，构造如图 2 所示的 SSS 特征结构。与词袋特征相比，SSS 特征由每个句子的情感强度值串联得出，保留了句子的位置信息；SSS 特征维度与文档中句子的数目相同，远远低于词袋特征维度，在数据量大的情况下可以大大减少训练和分类的时间成本。

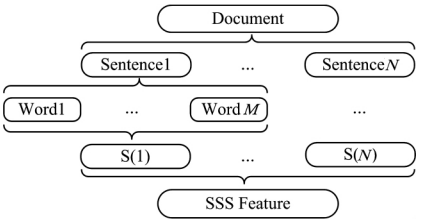

图 2 SSS 特征的结构示意图

3.2 篇章情感分类

根据上述计算得出各个句子的情感强度，进而计算篇章的情感强度值，由三支决策理论将对象划分为正域、负域、边界域 3 个部分。将正域和负域的对象直接对应划分至正类和负类，对处于边界域的对象，由 SVM+SSS 特征进一步分类，最终将所有对象分类完毕。

3.2.1 篇章情感强度的计算

任何一个作者在撰写主观文本(长文本)时都会有一个情感基调和情感走向过程，篇章级文本情感分类可以看作是由情感走向过程求解情感基调的任务，因此对于篇章级文本，以

[1] http://www.keenage.com

[2] http://word2vec.googlecode.com/svn/trunk

[3] https://dumps.wikimedia.org

句子为最小粒度建立文本情感走向模型来完成情感分类任务，如图 3 所示。

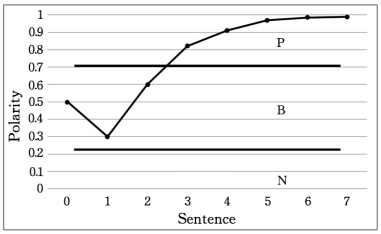

图 3　文本情感走向模型示意图

在图 3 中，横坐标表示按文档原有顺序排列的句子；纵坐标表示当前累计的文档情感强度值；P，B，N 分别表示三支决策的正域、边界域和负域。文本情感走向模型基于如下两个假设：

1）以句子为最小粒度来说，作者撰写主观文本时的情感值是连续的，整个文档的情感基调是每个句子情感强度的累积。

2）截止目前，累积的某一情感极性的强度越强，整个文档的情感极性越难达到其对立面，且情感强度达到某一阈值后可忽略后续句子，直接做出不可扭转的决策结果。

第一条假设容易理解。对于第二条假设，可以做出如下比较形象的类比解释：若某一个人做出的令你讨厌的事情越多，则你会越难喜欢这个人，当令你讨厌的事情多到一定程度时，你会拒绝与这个人产生任何联系，即不可扭转的决策结果。

基于以上假设，截止到句子 i 的文档情感强度值 $DS(i)$ 的计算公式如下：

$$DS(i)=\frac{1}{1+\mathrm{e}^{-\mu\sum_{k=1}^{i}S(i)}} \tag{4}$$

其中，公式的主体为 Sigmoid 函数，μ 为控制 Sigmoid 函数横向拉伸的参数。

3.2.2　篇章情感的三支划分

假设句子 N 为文档的最后一个句子，则 $DS(N)$ 被用作文档的三支决策划分，与阈值 α 和 β 进行比较。为获得三支决策划分的阈值，根据经验设置对象划分的损失函数，如表 2 所列，根据损失函数计算得出阈值 α 和 β，再根据 2.3 节的决策规则 P，B，N，可以将对象划分至正域、边界域和负域中。

表 2　三支决策损失函数的经验值设定

	$X(P)$	$X^c(N)$
P	0	10
B	4	3
N	9	0

经三支决策方法分类后，落入正域和负域的对象分别被划分至正类和负类，落入边界域的对象需要进一步分类。在取得 SSS 特征后，将其放入 SVM 进行再一次分类，综合三支决策方法和 SVM 得到最终的分类结果。

4　实验结果与分析

在中文和英文语料库上，比较 SSS 特征和基于 unigram、bigram、情感词典的词袋特征（基于情感词典的词袋特征的每一维代表情感词典中的一个词，SLBOW）的提取时间，同时比较上述特征用于 TWD-SSS 模型和 SVM，NB，DS[16]，LSS[17] 模型的分类准确率和分类时间。中文语料的分词使用中文分词工具 NLPIR[1] 来完成。由于特征提取和模型训练的时间成本与实验环境紧密相关，因此给出实验环境配置：操作系统 Windows10，编程语言 Python，内存 4GB DDR3，CPU 双核 2.4GHz。

4.1　实验数据

为了验证 TWD-SSS 模型在中文文本和英文文本上的有效性，在中文语料库和英文语料库上分别进行了实验。英文语料库选用康奈尔大学的电影评论 polarity dataset v2.0[2]，其中包括标注好正、负极性的电影评论各 1000 篇。中文语料库选用谭松波在携程（http://www.ctrip.com）上采集的酒店评论 ChnSentiCorp-Htl-ba-600[3]，其中包括正、负极性的酒店评论各 3000 篇。详细信息如表 3 所列，其中文档中的句子数和词数均为平均值。

表 3　语料库信息

语料库	文档数	句子数	词数
电影评论（英文）	1000＋1000	32	648
酒店评论（中文）	3000＋3000	13	82

4.2　不同特征和不同分类方法下的准确率比较

在两个语料库中进行三折交叉实验对比，各种特征和方法组合下的准确率如表 4 和表 5 所列，其中部分结果来自文献[16，21]。

表 4　不同方法在英文电影评论中的准确率

分类器＼特征	SSS	SLBOW	unigrams	bigrams
NB	0.786	0.824	0.787	0.773
SVM	0.754	0.805	0.728	0.771
DS	N/A	0.825	0.819	0.827
LSS	N/A	0.844	0.825	0.842
TWD-SSS	**0.839**	**0.879**	**0.823**	**0.815**

表 5　不同方法在中文酒店评论中的准确率

分类器＼特征	SSS	SLBOW	unigrams	bigrams
NB	0.788	0.807	0.844	0.869
SVM	0.810	0.791	0.827	0.862
DS	N/A	0.806	0.833	0.866
LSS	N/A	0.813	0.847	0.872
TWD-SSS	**0.871**	**0.816**	**0.836**	**0.869**

TWD-SSS 方法可以看作是三支决策与支持向量机的结合，比较表 4 和表 5 中 TWD-SSS 和 SVM 在各种特征表示下的性能可以看出，TWD-SSS 在各类特征表示中的准确率均优

1）http://ictclas.nlpir.org

2）http://www.cs.cornell.edu/people/pabo/movie-review-data/

3）http://www.searchforum.org.cn/tansongbo/corpus

于 SVM，证明了三支决策方法应用在情感分类中的有效性。

从表 4 和表 5 中可以看出，SLBOW 特征在英文语料库中的表现要优于在中文语料库中的表现，这是因为相比于英文评论，中文评论的词语更加复杂，情感词典中词语的覆盖程度较低，因此 SLBOW 特征在中文语料库中的表现稍差，当然这与情感词典的质量以及文本领域也有关系。此外，SSS 特征应用在除 TWD-SSS 以外的方法中的效果并不好，因为 SSS 特征的特点是维度低、可解释性强，而一般的分类模型并没有利用好 SSS 特征的可解释性，除去可解释性，维度低就成为了 SSS 特征的劣势，直接将其放入一般的分类模型中也不能取得像 BOW 这样高维度特征的准确率，而 TWD-SSS 则通过三支决策方法和情感走向模型充分利用了 SSS 特征的可解释性。

4.3 不同特征下的时间开销比较

在 TWD-SSS 模型中，SSS 特征的可解释性得到了充分利用，特征维度低就成为了 SSS 特征的优势，因为低维度的特征可以大大减少特征提取、模型训练和分类过程的时间成本。如表 6 所列，在英文语料中基于 unigram 的 BOW 特征维度是 16165，而 SSS 特征维度只有 32，特征提取耗费的时间也大大缩减。SSS 特征的低维特性除了可以减少特征提取的时间，还可以减少分类过程的时间，如图 4 和图 5 所示。低维度的 SSS 特征可以减少特征提取和模型训练的时间，三支决策方法可以减少分类过程所需的时间，这是由于在边界域对象被处理之前，已经有一部分数据被三支决策方法过滤，由于三支决策方法的损失函数是根据经验值设定的，因此三支决策过程所耗费的时间与后续的分类过程相比几乎可以忽略。

表 6 英文电影评论语料库中各种特征的维数和提取时间

特征	平均维数	提取时间/s
SSS	32	188.42
SLBOW	10871	284.43
unigrams	16165	305.66
bigrams	16165	313.17

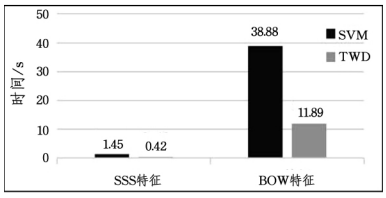

图 4 TWD-SSS 与 SVM 在英文电影评论语料上的分类时间比较

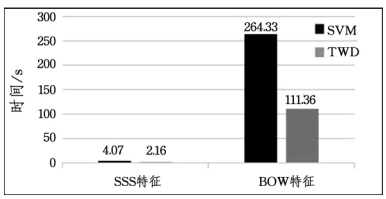

图 5 TWD-SSS 与 SVM 在中文酒店评论语料上的分类时间比较

在 TWD-SSS 方法中，无论对边界域对象的分类使用哪种特征，在进行正域、负域和边界域划分时使用的总是 SSS 特征，因此 SSS 特征最适用于 TWD-SSS 方法。综合不同分类器和特征来看，TWD+SSS 的组合已经拥有不错的分类准确率，但并不是最优的(英文电影评论语料库 TOP4，中文酒店评论语料库 TOP2)，然而 TWD-SSS 方法的运行效率却远远超过其他方法，而且考虑到 SSS 特征的低维度特性，它还拥有很大的改进空间，并且从粒计算的角度来说，SSS 特征具有很强的可解释性：篇章的分类依赖于句子粒度的数值，句子粒度情感强度的获取依赖于句子中的词语。词袋特征在处理篇章级文本情感分类时，跨越了句子这一粒度，这实则抛弃了许多信息；而 SSS 特征则很好地整理了这些信息，而且考虑到了句子之间的位置信息。

4.4 参数 μ 对情感分类性能的影响

三支决策中文档情感强度值的计算对整个模型有很大影响，其中参数 μ 的设定对三支决策部分的准确率、覆盖率以及整体准确率的影响如图 6 所示，其中覆盖率表示由三支决策处理的数据占整个测试数据的比例。从图 6 可以看出，随着 μ 的增大，三支决策的覆盖率逐渐增大，而且 μ 趋近于 0 时，数据极难达到三支决策的阈值判定标准，三支决策覆盖率趋近于 0；μ 趋近于无限大时，数据非常容易落在三支决策的正域和负域内，三支决策的覆盖率逼近 100%。μ 越小，三支决策的判定越严苛；μ 越大，三支决策的判定越宽松。三支决策的准确率趋势与覆盖率相反，当 μ 越小时，三支决策的判定很严苛，经由三支决策判定的数据量很小，但是判定的准确率很高；当 μ 越来越大时，宽松的标准增大了三支决策的覆盖率，却降低了三支决策的准确率。从整体准确率来看，μ 取 1 时整体准确率最高。

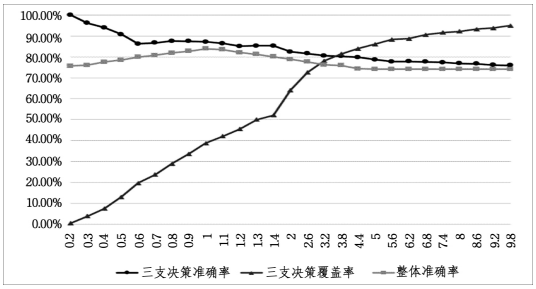

图 6 式(4)中 μ 对三支决策准确率和覆盖率以及整体准确率的影响

4.5 三支决策阈值 α 和 β 对情感分类性能的影响

三支决策中阈值的设定对三支决策准确率和覆盖率的影响如图 7 和图 8 所示，与参数 μ 的设定类似，随着阈值 α 和 β 的变化，三支决策准确率和覆盖率的变化呈对称状态，原因与 μ 类似，只是阈值的设定更直接地影响了三支决策的判定标准。

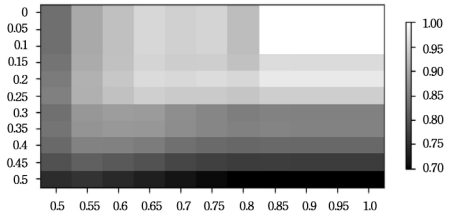

图 7 三支决策的阈值 α(横坐标)和 β(纵坐标)对三支决策准确率的影响

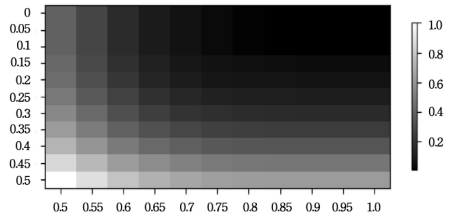

图 8　三支决策的阈值 α(横坐标)和 β(纵坐标)对三支决策覆盖率的影响

4.6　跨领域情感分类效果

为了验证模型的有效性，对 TWD-SSS 方法进行了跨领域数据集的训练和测试，在电影评论数据集上训练模型，在酒店评论数据集上进行测试(其中在计算句子情感强度的步骤中使用的领域情感词典依旧是从各自领域训练得来的)，实验结果如图 9 所示。

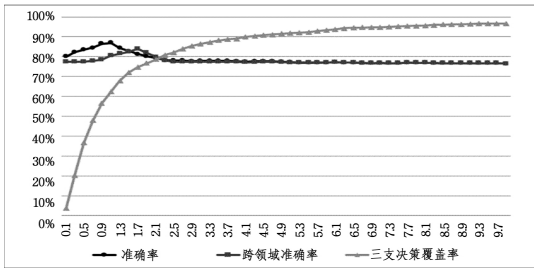

图 9　跨领域情感分类效果在不同 μ 值(横坐标)下的对比

从图 9 中可以看到，使用其他领域数据集训练模型得出的准确率稍逊于使用本领域数据集训练模型得出的准确率，但并不存在效果急剧下降的情况，这说明模型训练没有出现过拟合的情况。从理论角度解读图 9 中的实验结果：由于三支决策方法不同于支持向量机等机器学习方法，三支决策不存在过拟合的问题，因此过拟合的风险仅存在于处理边界域对象所采用的支持向量机方法中。由于训练支持向量机时所使用的 SSS 特征维度很低，而且三支决策又处理了一部分测试数据，因此 TWD-SSS 模型不易出现过拟合现象，这也与图 9 所示的实验结果相吻合。另外从图 9 中还可以看出，使用跨领域数据集训练得到的模型取得准确率峰值点的 μ 值比使用本领域数据集训练得到的模型取得准确率峰值点的 μ 值要大，结合 μ 值越大三支决策覆盖率越大的特性，跨领域训练模型时，其准确率更加依赖于三支决策，因为三支决策在处理不同领域数据时的表现很稳定(虽然不高)，这也是三支决策不同于机器学习方法的一个特点。

结束语　本文提出一种基于三支决策的多粒度文本情感分类模型，旨在更好地解决篇章级文本情感分类问题。主要的创新点为：传统的词袋特征应用在篇章级文本情感分类任务中，其维度高、可解释性差，而且忽略了文本的位置信息，对此提出一种基于文本粒度结构的 SSS 特征；将三支决策方法和情感走向模型融合起来，结合 SSS 特征构成 TWD-SSS 模型来处理情感分类任务。实验结果表明，将三支决策方法和 SVM 分类器相结合后在各种特征上的准确率都要优于仅使用 SVM 分类器的准确率，而且三支决策方法可以降低分类过程的时间成本。SSS 特征的低维特性可以大大减少特征提取和模型训练的时间，而且 SSS 特征由于维度很低，因此具有很大的优化空间。本文主要考虑文本情感分类的二元分类任务，后续工作会考虑构建适用于多分类任务的 TWD-SSS 模型，并将继续完善 SSS 特征的构造过程。

参 考 文 献

[1] LIU B. Sentiment analysis: Mining opinions, sentiments, and emotions[M]. Cambridge University Press, 2015.

[2] FENG S, FU Y C, YANG F, et al. Blog sentiment orientation based on dependency parsing [J]. Journal of Computer Research and Development, 2012, 49(11): 2395-2406. (in Chinese)
冯时，付永陈，阳峰，等. 基于依存句法的博文情感倾向分析研究[J]. 计算机研究与发展，2012，49(11)：2395-2406.

[3] WANG S G, LI D Y, WEI Y J. A method of text sentiment classification based on weighted rough membership [J]. Journal of Computer Research and Development, 2011, 48(5): 855-861. (in Chinese)
王素格，李德玉，魏英杰. 基于赋权粗糙隶属度的文本情感分类方法[J]. 计算机研究与发展，2011，48(5)：855-861.

[4] FELDMAN R. Techniques and applications for sentiment analysis[J]. Communications of the ACM, 2013, 56(4): 82-89.

[5] ZHAO Y Y, QIN B, LIU T. Sentiment analysis [J]. Journal of Software, 2010, 21(8): 1834-1848. (in Chinese)
赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848.

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

[7] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques [C]// Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. ACL, 2002: 79-86.

[8] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL, 2002: 417-424.

[9] DING X W, LIU B, YU P S. A holistic lexicon-based approach to opinion mining [C]// Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 231-240.

[10] THELWALL M, BUCKLEY K, PALTOGLOU G. Sentiment strength detection for the social web [J]. Journal of the American Society for Information Science and Technology, 2012, 63(1): 163-173.

[11] CHO H, KIM S, LEE J, et al. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews [J]. Knowledge-Based Systems, 2014, 71: 61-71.

[12] KATZ G, OFEK N, SHAPIRA B. ConSent: Context-based sentiment analysis [J]. Knowledge-Based Systems, 2015, 84: 162-178.

(下转第 215 页)

相关领域知识的依赖，即使在支持度设置得较小时，也不会产生大量的规则，并且挖掘到的规则的置信度更高。

参 考 文 献

[1] HAN J，KAMBER M. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2006：1-27.

[2] RAKESH A，SRIKANT R. Fast Algorithms for Mining Association Rules[C]// Proceedings of International Conference on Very Large DataBases. Santiago，Chile：ACM Press，1994：21-30.

[3] 李锦泽，叶晓俊. 关联规则挖掘算法研究现状[C]// 计算机技术与应用进展——全国计算机技术与应用. 安徽：中国科学技术大学出版社，2007：9-14.

[4] CUI L，GUO J，WU L D. Algorithm for Mining Association Rules Based on Dynamic Hashing and Transaction Reduction [J]. Computer Science，2015，42(9)：41-44. (in Chinese)
崔亮，郭静，吴玲达. 一种基于动态散列和事务压缩的关联规则挖掘算法[J]. 计算机科学，2015，42(9)：41-44.

[5] XIE Z P，LIU Z T. Concept Lattice and Association Rule Discovery [J]. Journal of Computer Research & Development，2000，37(12)：1415-1421. (in Chinese)
谢志鹏，刘宗田. 概念格与关联规则发现[J]. 计算机研究与发展，2000，37(12)：1415-1421.

[6] LI Y，LI T，CAI J J，et al. Extracting Succinct Association Rules Based on Concept Lattice[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science)，2007，27(3)：44-47. (in Chinese)
李云，李拓，蔡俊杰，等. 基于概念格提取简洁关联规则[J]. 南京邮电大学学报(自然科学版)，2007，27(3)：44-47.

[7] OUYANG J H，WANG Z J，LIU D Y. An Improved Association Rule Algorithm with Dynamically Weighted Characteristic[J]. Journal of Jilin University (Science Edition)，2005，43(3)：314-319. (in Chinese)
欧阳继红，王仲佳，刘大有. 具有动态加权特性的关联规则算法[J]. 吉林大学学报(理学版)，2005，43(3)：314-319.

[8] DUAN J，DAI J F. Algorithm of Mining Weighted Association Rules Based on Multiple Supports[J]. Journal of Tianjin University，2006，39(1)：114-118. (in Chinese)
段军，戴居丰. 基于多支持度的挖掘加权关联规则算法[J]. 天津大学学报，2006，39(1)：114-118.

[9] LI J，CERCONE N. A Rough Set Based Model to Rank the Importance of Association Rules [C]// Proceedings of Rough Sets，Fuzzy Sets，Data Mining，and Granular Computing. Berlin Heidelberg：Springer Press，2005：109-118.

[10] HU K，LU Y，ZHOU L，et al. Integrating Classification and Association Rule Mining：A Concept Lattice Framework[C]// Proceedings of New Directions in Rough Sets，Data Mining，and Granular-Soft Computing. Berlin Heidelberg：Springer Press，2003：443-447.

[11] SIM A，INDRAWAN M，ZUTSHI S，et al. Logic-Based Pattern Discovery [J]. IEEE Transactions on Knowledge and Data Engineering，2010，22(6)：798-811.

[12] ZAKI M. Scalable Algorithms for Association Mining[J]. IEEE Transactions on Knowledge and Data Engineering，2000，12(3)：372-390.

[13] ZAKI M，GOUDA K. Fast Vertical Mining using Diffsets [C]// Proceedings of International Conference on Knowledge Discovery and Data Mining. Washington DC：ACM Press，2003：326-335.

[14] CHEN C H，LAN G C，HONG T P，et al. Mining High Coherent Association Rules with Consideration of Support Measure [J]. Expert Systems with Applications，2013，40(16)：6531-6537.

[15] AN J R，WANG H P，ZHANG L B，et al. A Compression Matrix Algorithm for Mining Association Rules Based on Mapreduce[J]. Journal of Chongqing University of Technology(Natural Science)，2016，30(2)：95-100. (in Chinese)
安建瑞，王海鹏，张龙波，等. 一种基于 MapReduce 的压缩矩阵关联规则挖掘算法[J]. 重庆理工大学学报(自然科学版)，2016，30(2)：95-100.

(上接第 193 页)

[13] FRANCO-SALVADOR M，CRUZ F L，TROYANO J A，et al. Cross-domain polarity classification using a knowledge-enhanced meta-classifier [J]. Knowledge-Based Systems，2015，86：46-56.

[14] ZHANG Z F，MIAO D Q，NIE J Y，et al. Sentiment uncertainty measure and classification of negative sentences [J]. Journal of Computer Research and Development，2015，52(8)：1806-1816. (in Chinese)
张志飞，苗夺谦，聂建云，等. 否定句的情感不确定性度量及分类[J]. 计算机研究与发展，2015，52(8)：1806-1816.

[15] ZHANG Z F，MIAO D Q，YUE X D，et al. Sentiment analysis with words of strong semantic fuzziness [J]. Journal of Chinese Information Processing，2015，29(2)：68-78. (in Chinese)
张志飞，苗夺谦，岳晓冬，等. 强语义模糊性词语的情感分析[J]. 中文信息学报，2015，29(2)：68-78.

[16] DAS S，CHEN M. Yahoo！for Amazon：Extracting market sentiment from stock message boards [C]// Proceedings of the Asia Pacific Finance Association Annual Conference. 2001：43.

[17] 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. ACL，2010：635-643.

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

[19] YAO Y Y，ZHAO Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences，2008，178(17)：3356-3373.

[20] QIU G，LIU B，BU J，et al. Expanding domain sentiment lexicon through double propagation [C]// Proceedings of the 21st International Joint Conference on Artificial Intelligence. Morgan Kaufmann，2009：1199-1204.

[21] XIA R，XU F，ZONG C Q，et al. Dual sentiment analysis：Considering two sides of one review [J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(8)：2120-2133.