基于强化表征学习深度森林的文本情感分类

韩 慧[1] 王黎明[1] 柴玉梅[1] 刘 箴[2]
(郑州大学信息工程学院 郑州 450001)[1] (宁波大学信息科学与工程学院 浙江 宁波 315211)[2]

摘 要 为了有效实现评论文本的情感倾向性预测,在深度森林模型的基础上提出一种基于强化表征学习的深度森林算法BFDF(Boosting Feature of Deep Forest)来对文本进行情感分类。首先,提取二元特征与情感语义概率特征;其次,对二元特征中的评价对象做聚类处理以及特征融合;然后,改进深度森林级联层的表征学习能力,避免特征信息逐渐削减;最后,将AdaBoost方法融入到深度森林,使深度森林注意到不同特征的重要性,进而得到改进的模型BFDF。在酒店评论语料集上进行了实验验证,实验结果证明了该方法的有效性。

关键词 情感分类,特征提取,深度森林,AdaBoost
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.027

Text Sentiment Classification Based on Deep Forests with Enhanced Features

HAN Hui[1] WANG Li-ming[1] CHAI Yu-mei[1] LIU Zhen[2]
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)[1]
(School of Information Science and Technology, Ningbo University, Ningbo, Zhejiang 315211, China)[2]

Abstract To effectively realize the sentiment orientation prediction of the review text, based on the deep forest model, a deep forest algorithm BFDF (Boosting Feature of Deep Forest) was proposed to classify the text. Firstly, the binary features and emotional semantic probability features are extracted. Secondly, the evaluation objects in the binary features are clustered and made features fusion. Then, the deep forest cascade characterization learning ability is improved to avoid the gradual reduction of feature information. Finally, the AdaBoost method is integrated into the deep forest, so that the deep forest notices the importance of different features, and the improved model BFDF is obtained. The experimental results on the hotel commentary corpus demonstrate the effectiveness of the proposed method.

Keywords Sentiment classification, Feature extraction, Deep forest, AdaBoost

1 引言

随着计算机和互联网技术的迅猛发展,人们越来越倾向于利用网络来获取信息并发布信息[1]。其中最为显著的是电子商务,它的迅猛发展给人们带来了极大便利,电商的出现使人们足不出户就可以买到自己心仪的商品。然而网络存在双面性,虽然电商为人们节省了时间,但是在面对网络中丰富多样的商品时,消费者如何做出正确的选择也成为一大难题。

针对这一现象,越来越多的研究者倾向于对消费者评论进行情感分析,即将评论分为正负两类[2],这不仅可以为商家提供有效的反馈信息,也可以为广大群众提供有价值的参考信息。

用户评论的情感分析是对非结构化的评论语句进行分析以识别用户的情感状态的过程[3]。由于商品评论句式结构的特殊性,一条评论中包含了对产品多方面属性的评论,其中有好评也有差评,因此识别评价对象成为对产品评论进行情感分析的过程中的重要步骤。Jo等[4]从此角度出发,提出了ASUM模型,该模型可以针对产品的不同属性,分析其评价词的情感倾向。

实现文本情感分析不仅需要提取合适的文本特征,选择分类器也至关重要。本文采用Zhou等[5]提出的深度森林DF(Deep Forest)为原型,并在此基础上做出改进,将改进的分类器作为本文的分类器。DF模型使用级联结构让深度森林做表征学习,当输入高维数据时,可以通过多粒度扫描结构使表征学习能力有所提升,进而可使DF注意到上下文或结构。文献[5]已经通过实验证明了DF模型的有效性。

综合以上研究,本文首先提取文本的二元特征,即评价对象-评价词极性特征对;再根据评价对象的TF-IDF值对评价对象做聚类处理,以确定二元特征的最终维度k;而后融合二元特征与情感语义概率特征作为文本的最终特征。为了避免随着级联层深度的加深,特征所携带的信息流逐渐削弱,本文改进了级联层的表征学习结构;并且为了使模型注意到文本

到稿日期:2018-06-12 返修日期:2018-09-13 本文受国家自然科学基金项目(U1636111)资助。

韩 慧(1992-),女,硕士,主要研究方向为自然语言处理,E-mail:18337149649@163.com;**王黎明**(1963-),男,博士,教授,CCF高级会员,主要研究方向为现代软件工程技术、分布式人工智能、数据挖掘等,E-mail:ielmwang@zzu.edu.cn(通信作者);**柴玉梅**(1964-),女,硕士,教授,主要研究方向为机器学习、数据挖掘和自然语言处理;**刘 箴**(1965-),男,博士,研究员,主要研究方向为虚拟现实、情感计算。

特征的重要性，本文将 AdaBoost 方法融入到改进之后的模型中，得到最终的深度森林模型 BFDF，从而实现评论文本的情感倾向预测。

2　相关工作

情感分析作为自然语言处理的一个子任务，已经深受广大研究人员的青睐。其中，细粒度的情感挖掘是近年来研究的热门领域。Zhao 等[6]通过统计评价对象与情感词的共现，以及二者间的依赖模式，构建了一种情绪图模型，以实现对评价对象和情感词的提取。Liu 等[7]提出用异构图构建语义关系和情感关系，然后用联合排序算法估算每个候选词的置信度，从而实现对评价对象和情感词的提取。以此作为出发点，本文提出提取文本的二元特征，即将评价对象-评价词极性对作为文本特征。

GU 等[8]以 SBV 极性传递法为核心，引入 ATT 定中关系等算法来实现对评价对象的提取。Kamal 等[9]结合文本的语言和语义分析，提出了一种基于规则的系统，以实现对评价对象和情感词对的提取。王素格等[10]利用依存分析，构建了含有特征和观点词语的组块规则，以实现对特征-观点对的提取。Saru 等[11]采用基于图的协同排序算法提取目标词与观点词。但是文献[8]只考虑了主谓关系，因此存在局限性。文献[9-11]对隐式及缺省评价对象的提取不够完善。因此，本文在预处理阶段实现句子结构的统一化，通过规则成对提取评价对象-评价词极性对，此过程也包括对隐式及缺省评价对象的处理。此外，本文对二元特征还进行了聚类处理，以确定其最终维度。

冯时等[12]提出了一种基于依存句法分析技术的算法 SOAD，来对博文搜索结果进行情感倾向性分析。Xiao 等[13]将情感词典与依存分析相结合，首先改进了情绪词典，然后基于语法分析计算情感词在语句中的情感权重，实现了情感句的判别。考虑到二元特征缺乏对文本的语义信息的提取，本文结合句法分析提取文本的情感语义概率特征，并将其与二元特征进行融合作为最终特征。

Lev 等[14]在 Deep Forest 模型的基础上提出了 Siamese Deep Forest(SDF)模型，实现了相似度度量学习。朱晓妤等[15]将深度森林模型运用到火焰检测，改进了多粒度扫描结构并提取出了抽象特征，再使用深度森林模型进行火焰检测。王海洋[16]对 Deep Froest 模型进行了改进，并将改进后的模型用于文本分类，实现了很好的分类效果。Lev 等[17]通过给决策树分配权重，改进了原始的距离度量算法。杨峰等[18]将 RPN 网络与深度森林模型相结合，有效地实现了船舰探测。考虑到深度森林模型的简单、高效性，本文采用深度森林模型为原型，将 AdaBoost 集成方法应用于深度森林的每一层，增强其对特征的识别能力，并且改进了级联层的表征学习能力，以更好地实现评论文本的情感倾向性判别。

3　理论基础

3.1　相关定义

定义 1(二元特征)　$S=\{O,p\}=\{\langle o_1,p_1\rangle,\langle o_2,p_2\rangle,\cdots,\langle o_n,p_n\rangle\}$ 是该文本的二元特征表示。其中，o_i 表示提取的文本的评价对象，p_i 表示修饰评价对象的评价词极性，$p_i=\{-1,1\}$。

本文在提取完二元特征之后，会依据评价对象的 TF-IDF 值对评价对象做 k-means 聚类处理，以确定二元特征的最终维度 k。之后，继续提取文本的情感语义概率特征并将其与二元特征进行融合。

定义 2(情感语义概率特征)　$W=\langle P,N,F,G\rangle$ 表示情感语义概率特征集合。其中，P 为正向情感词概率特征，N 为负向情感词概率特征，F 为非情感词否定概率特征，G 为负面情感概率特征。

3.2　深度森林

由文献[5]可知，深度森林可以理解为基于树的集成方法，主要包括两种结构：多粒度扫描与级联森林，其定义如下。

定义 3(多粒度扫描)　$W=\langle X^N,v,b,l\rangle$，其中，$X$ 为原始输入特征，N 为其维度，v 为扫描窗口维度，b 为扫描步长，l 为扫描窗口数量。则经过扫描之后的特征数 r 变为 $r=(N-v)/s+1$。

这里的多粒度扫描类似于 CNN 的多粒度卷积核，v 可以取不同值，以实现真正意义上的多粒度采样，从而获得更多的特征子样本。

定义 4(级联森林)　$CF=\{z,F,t,c\}$ 表示级联森林。其中，$z=\{1,2,\cdots,Z\}$ 代表级联森林的级数，每一级包含 m 个森林 F，$m=\{1,2,\cdots,M_z\}$。而 F 是分别由 t 棵决策树组成的随机森林和完全随机森林，$t=\{1,2,\cdots,T_{m,z}\}$，$c=\{1,2,\cdots,C\}$ 代表样本的类别标签。

在训练阶段，级联森林的每一级会生成对样本 x 的类分布向量，如式(1)所示：

$$P^{(t,m)}(x)=(p_1^{(t,m)}(x),p_2^{(t,m)}(x),\cdots,p_c^{(t,m)}(x)) \tag{1}$$

其中，$p_c^{(t,m)}$ 是每棵决策树计算的样本 x 属于类别 c 的概率。然后每个森林会根据该概率得到自己对样本 x 的类分布估计，表示为：$V^m(x)=(V_1^m(x),V_2^m(x),\cdots,V_c^m(x))$。根据文献[5]，$V_c^m(x)$ 可以写成：

$$V_c^m(x)=T_m^{-1}\sum_{t=1}^{T_m}p_c^{(t,m)}(x) \tag{2}$$

然后，级联森林将每一级的输出结果与原始特征向量相结合作为下一级森林的输入，表示为：$x\leftarrow(x,V_1(x),V_2(x),\cdots,V_C(x))$。以此类推，直到准确率不再上升，停止训练。

4　BFDF 的文本情感分类

在文本情感分类任务中，为了得到文本情感分类器，需要从训练样本中提取出能够描述文本情感差异的特征向量，并将该特征向量作为 BFDF 的输入特征，才能最终训练得到文本情感分类器。在提取特征之前，需要对文本做适当的预处理。

1)为了实现文本句子结构的统一化，将文本中无法表示完整语句的标点符号以及文本中出现的空格均用句号替换。

2)使用哈尔滨工业大学开发的语言技术平台 LTP(Lan-

guage Technology Platform)对句子进行分词、词性标注、依存句法分析和语义角色标注等处理。

4.1 提取文本二元特征

二元特征的提取主要包括 3 个步骤,即抽取评价对象、确定修饰该评价对象的评价词极性以及特征后处理。

4.1.1 抽取评价词并确定评价词极性

本文是针对产品评论进行的情感分类,在产品评论中,评价词多以形容词为主,少数情况下会出现名词性的评价词,因此本文只抽取形容词及名词性评价词。具体分析如下:

1)对文本进行分词等预处理,然后根据情感词典识别句中的评价词,并标注其情感极性 p,$p=\{-1,1\}$。

2)通过分析文本的语义依存关系,判断评价词是否有否定标记"mNeg"修饰。若有,则改变其极性;否则不变。

4.1.2 抽取评价对象

形容词性评价词在句子中多以谓语角色出现,但是考虑到消费者描写评论的不规范性,评价词在句中也会充当宾语的角色。以上两种情况的分析结果如图 1 所示。

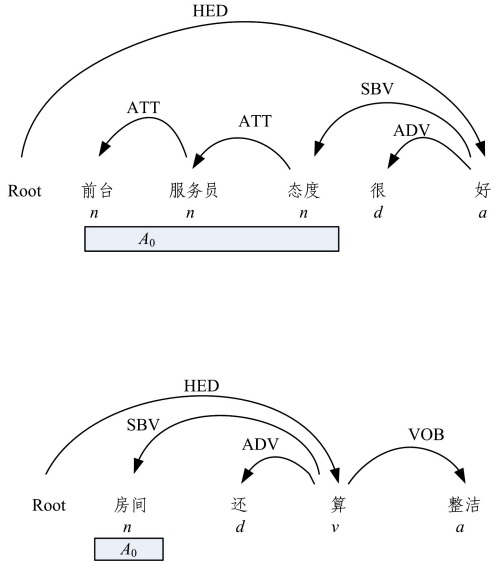

图 1 例句的语义依存分析结果

Fig. 1 Semantic dependency analysis results

由图 1 可以分析出,无论评价词充当谓语(VOB)还是宾语(SBV),其评价对象均可由 A_0 表示。规则 1 如下所示:

规则 1 假设 w 为评价词,则 if w=谓语 or 宾语,then $o=A_0$。

若评价对象为代词,则需要确定该代词指代的具体对象。规则 2 如下所示。

规则 2(隐式评价对象处理) if A_0=代词,then $o=pre\text{-}sentence\ A_0$。

若上述规则并未找到评价对象,则可能在预处理阶段已将评价对象与评价词分割在两个句子中,这时本文从上下文语义关联的角度考虑,给缺失的评价对象赋值。规则 3 如下所示。

规则 3(缺失值赋值) if $o=\varnothing$,then $o=pre\text{-}sentence\ A_0$。

4.1.3 特征后处理

上述过程完成之后,每条文本均会得到形如 $S=\{\langle o_1,p_1\rangle,\langle o_2,p_2\rangle,\cdots,\langle o_n,p_n\rangle\}$ 表示的二元特征集合。但是评论文本带有强烈的主观意识,这会使得每条文本提取的评价对象复杂且不尽相同。针对此问题,本节首先根据式(3)计算评价对象 o 的 $TF\text{-}IDF$ 值:

$$TF\text{-}IDF_o=\frac{n_{o,c}}{\sum_{i=1}^{m}n_{m,c}}\times\log\left(\frac{N}{n_o+1}\right) \tag{3}$$

其中,c 表示文档的类别标签,本文中 $c=\{-1,1\}$;$n_{o,c}$ 表示在 c 类文档中评价对象 o 出现的次数;$n_{m,c}$ 表示在 c 类文档中所有的评价对象的数目;N 表示文档总数;n_o 表示包含评价对象 o 的文档数。

然后根据上述结果对评价对象 o 做 k-means 聚类处理,将文本的评价对象集合分成 k 类,分别代表文本的 k 个特征,而每一类的情感极性 p_c 由该类中包含的特征对 $\langle o_i,p_i\rangle$ 决定,如式(4)所示:

$$p_c=\begin{cases}1, & \sum p_o>0\\ 0, & \sum p_o=0\\ -1, & \text{else}\end{cases} \tag{4}$$

分析可知,聚类操作不仅可以降低二元特征的维度,也可以最终确定每条文本的特征数 k。以此作为 BFDF 模型的 k 个输入特征,即 $X=(C_i,p)$,其中 $C_i=\{0,1,\cdots,k-1\}$,$p=\{-1,1\}$。

算法 1 提取二元特征 (EFGC)

输入:经过预处理之后的文本 t

输出:文本的二元特征

```
1. for w in t
2.   if w=形容词或名词
3.     if match(w,pos)
4.       p=1;
5.       if 有否定词修饰 w
6.         p=-1;
7.       end if
8.     else if match(w,neg)
9.       p=-1;
10.    end if
11.    if w=谓语 or 宾语
12.      o=A0;
13.    else if A0=代词 or A0=Ø
14.      o=pre-sentence A0;
15.    end if
16.  end if
17. end for
18. tfidf=TF-IDF(o);
19. 对 o 作 k-means 处理,分成 k 类;
20. for o in t
21.    if Ci(o)∈Ci
22.       p(Oi)=Σ_{a=1}^{m} p(oa);
23.       if p(Oi)>0
24.          p(Ci)=1;
25.       else p(Ci)=-1
26.    else
27.       p(Ci)=0;
28. end for
```

算法中,pos 为正向情感词典,neg 为负向情感词典,$match()$为匹配函数。第 3－17 行提取二元特征;第 18－19 行根据提取的评价对象的 $TF\text{-}IDF$ 值,实现评价对象聚类;第 20－28 行根据聚类结果,将每个文本中未涉及的类别的情感极性赋值为 0,对于同一类中含有不同情感极性的评价对象,利用第 22－25 行确定该类别最终的情感极性。

4.2　提取文本情感语义概率特征

由于文本二元特征没有考虑文本语义信息,因此,本节结合情感词词典、转折词词典、递进词词典、情感词影响因子词典与依存句法分析,引入文本情感语义概率特征。每个特征的获取原理如下所示。

4.2.1　正向情感词概率特征

本文在现有情感词典的基础上构建了正向情感词词典。对于评论文本 T_i,经过预处理之后,首先根据情感词典识别句中的正向情感词,同时给情感分赋予初始值 $posSScore_0=1$。然后根据语义依存分析结果,判断该词是否存在否定词以及程度副词修饰,并根据式(5)改变其情感分值。

$$posSScore=\prod_{i=1}^{m}\beta\prod_{j=1}^{n}(-1)\times posSScore_0 \tag{5}$$

其中,β 为程度副词的强度值,本文中 $\beta=\{1.1,1.2,1.4,1.6,1.8,2\}$。根据上述分析,可以得到每条文本含有正向情感词特征的情感句权重值,如式(6)所示:

$$PosScore=\frac{1}{l}\sum_{i=1}^{l}posSScore_i \tag{6}$$

其中,$posSScore_i>0$。分析发现,式(6)实现了特征值归一化,特征值的取值范围为[0,2],因此可以根据式(7)使样本在该特征下取 L 个特征值。

$$L=(max-min)/b \tag{7}$$

其中,max 和 min 分别为区间(开区间或闭区间)的最大值和最小值;b 为步长,这里可将步长取 0.5。那么样本的正向情感词特征可以取 4 个特征值,即 $p\_pos=\{0.5,1,1.5,2\}$。

4.2.2　负向情感词概率特征

本文在现有情感词典的基础上构建了负向情感词词典。对于评论文本 T_i,首先进行相同的文本预处理,然后根据情感词典识别句中的否定情感词,并赋予其初始情感分值 $negSScore_0=-1$。根据语义依存分析结果,判断该情感词是否存在程度副词修饰,并根据式(8)更新其情感分值。

$$negSScore=\prod_{i=1}^{m}\beta\times negSScore_0 \tag{8}$$

进而可以得到含有负向情感词特征的情感句权重值,如式(9)所示:

$$NegScore=\frac{1}{m}\sum_{i=1}^{m}negSScore_i \tag{9}$$

分析发现,特征值的取值范围为[－2,－1),因此根据式(7),将步长取 0.2,那么样本的负向情感词特征可以取 5 个特征值,即 $p\_neg=\{-2,-1.8,-1.6,-1.4,-1.2\}$。

4.2.3　非情感词否定概率特征

给定评论文本 T_i,对预处理结果进行分析,若句中不含情感词,则考虑否定词在判断语句情感倾向时的作用。即如果语义分析中存在“mNeg”标识,则判定此句为负面情感句,并赋予情感分值 $mNegScore=-1$。最后计算该特征的情感权重,如式(10)所示:

$$mNeg\_Score=\frac{1}{N}\sum_{i=1}^{n}|mNegScore_i| \tag{10}$$

分析发现,特征值的取值范围为[0,1]。因此根据式(7),将步长取 0.15,那么样本的非情感词否定特征可以取 6 个特征值,即 $p\_mNeg=\{0.15,0.3,0.45,0.6,0.75,0.9\}$。

4.2.4　负面情感概率特征

给定评论文本 T_i,对预处理结果进行分析,如果语句中不包含情感词,也没有否定词,则可以根据转折词特征与递进词特征来判断语句的情感倾向性。

1)转折词特征

例　宽带要另外收费不爽。不过有数字电视算是弥补一下了。

根据情感词特征分析出第一句为负面情感句。第二句虽然不包含情感词,但是不难判断此句为正向情感句。这时可以结合转折词词典,依据预处理结果,规定如果语句第一个词($id=0$)是转折词,则赋予与前一句相反的情感分值。

2)递进词特征

例　太旧了。而且酒店内有 KTV。

根据情感词特征分析出第一句为负面情感句,

而第二句虽不包含情感词和否定词,但是不难判断出此句仍为负面情感句。这时可以结合递进词词典,规定如果语句第一个词($id=0$)是递进词,则赋予与前一句相同的情感分值。

给定评论文本 T_i,完成句子级的情感分析之后,即可提取负面情感概率特征,如式(11)所示:

$$neg\_prob=\begin{cases}1, & negCount/N>1/3\\ -1, & \text{else}\end{cases} \tag{11}$$

其中,1/3 是根据评论文本特性设定的阈值。因此,负面情感概率特征有两个特征值。

4.3　BFDF 的实现

4.3.1　改进级联层的表征学习能力

如 3.2 节所述,深度森林级联层每一级的输入向量是由前一级森林输出的类分布向量与初始特征向量拼接而成的,即 $\boldsymbol{P}^{(t,m,z)}(x)=(P_c^{(t,m,z-1)}(x),X^d)$,本文中 $c=\{-1,1\}$。随着森林深度的不断加深,特征向量所携带的文本信息不断退化,从而导致分类结果曲线起伏不稳定。文献[16]称此问题为稀疏连通性,并针对该问题提出了解决办法,将深度森林级联层每一级的输入向量改进为将原始向量与之前每一级森林的输出进行结合,以防止信息流的削弱,如式(12)所示:

$$\boldsymbol{P}^{(t,m,z)}=(P_c^{(t,m,1)},\cdots,P_c^{(t,m,z-1)},X^d) \tag{12}$$

但是本文考虑到这种改进可能会引发以下问题:随着深度森林级联层的不断加深,每一级输入的特征向量的维度会不断增加,从而会不断加大时间复杂度。因此,本文在此基础上继续做出改进,将深度森林级联层每一级的输入向量改进为将原始向量与之前每一级森林输出的类分布向量的平均值进行结合,如式(13)所示:

$$\boldsymbol{P}^{(t,m,z)}=(\frac{1}{z-1}\sum_{z=1}^{z-1}P_c^{(t,m,z)},X^d) \tag{13}$$

可以分析出，这样改进之后，每一级输入的特征向量不仅考虑了原始特征向量的影响，也考虑了此前每一级森林的分类结果的影响。因此这种改进方式不仅可以保留文本特征信息，也可以解决文本特征的维度不断扩大的问题。改进之后的结构如图 2 所示。

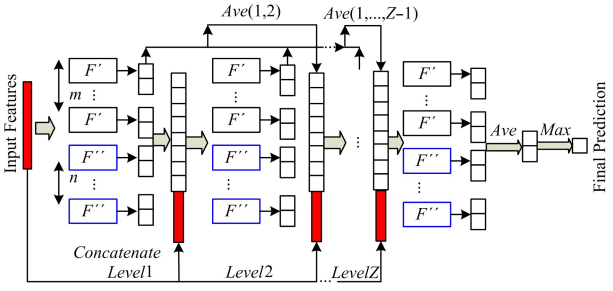

图 2 改进之后的深度森林模型

Fig. 2 Improved deep forest model

如图 2 所示，在改进后的深度森林中，级联层每一层的输入向量不再是将原始向量与每一级森林的输出向量进行简单结合，而是将原始向量与之前每一级森林输出结果的平均值进行结合。例如第 Z 层的输入向量 $\boldsymbol{P}^{(t,m,z)}$ 可以简单地表示为 $\boldsymbol{P}^{(t,m,z)}=(Ave(\boldsymbol{P}^{(t,m,1)},\cdots,\boldsymbol{P}^{(t,m,z-1)}),X^d)$。

4.3.2 融入 AdaBoost

完成特征提取以后，每个文本包含 N 个特征，$N=k+4$，这 N 个特征可以表示为：$X=\{\langle C_i,p\rangle,P,F,G,N\}$，$i=\{0,1,\cdots,k-1\}$。由于原始的深度森林模型没有考虑到不同的特征对分类结果的贡献不一，因此本文提出将 AdaBoost 方法融入到深度森林的每一层，通过给特征赋予初始权值 $w_i=1/N$，让模型在训练过程中根据每一层的分类结果 $\boldsymbol{P}^{(t,m,z)}(X_w)$，减小对分类效果影响大的特征的权值，而增加对分类结果影响小的特征的权值，使得在接下来的训练过程中，模型更加注重这些特征，从而学习到分类性能更好的深度森林。融入 AdaBoost 方法后的深度森林算法如算法 2 所示。

算法 2 特征强化的深度森林算法(BFDF)

输入：原始数据集 X^R，类别标签 $y=\{-1,1\}$

输出：最终分类结果

1. 文本预处理
2. 提取二元特征
3. k-means(TF-IDF(o))
4. $X^d=$ EFGC(t)$=\{(C_0,p),(C_1,p),\cdots,(C_{k-1},p)\}$
5. 提取情感语义概率特征
6. 特征融合：$X^d\leftarrow(X^d,P,N,F,G)$
7. 设置级联层的随机森林数 m
8. 设置随机森林中的决策树数 n
9. 初始化每个特征的权重 $w_i=1/N$，$N=k+4$
10. for z=1 to Z
11. 使用带有初始权值的 X^d 训练级联森林：
12. $h_z=DF(X^d,y,w_i)$
13. 计算每一级森林的加权类分布向量：
14. $\boldsymbol{P}^{(t,m,z)}(X_w)=(p_c^{(t,m,z)}(X_w))$，$c=\{-1,1\}$
15. 计算分类误差率：
16. $e_z=P(DF_z(X_i\neq y_i))=\sum_{i=1}^{N}w_iI(DF_z(X_i\neq y_i))$
17. 计算分类器 DF(X)的系数：
18. $\alpha_z=\frac{1}{2}\log\frac{1-e_z}{e_z}$
19. 更新特征向量：
20. $X_z=(X^d,\frac{1}{Z-1}\sum_{z=1}^{z-1}P_c^{(t,m,z)})$，$c=\{-1,1\}$
21. 更新特征权值：
22. $w_{z+1,i}=\frac{w_{z,i}\exp(-\alpha_z y_i DF_z(X))}{\sum_{i=1}^{N}w_{z,i}\exp(-\alpha_z y_i DF_z(X))}$
23. end for
24. final_pred$=$argmax $p_c^{(t,m,Z)}(X_w)$，$c=\{-1,1\}$

算法第 3 行计算评价对象的 *TF-IDF* 值，然后根据这个结果对评价对象进行聚类处理，获得 $N=k+4$ 个文本特征。为了让模型识别出每个特征的不同的重要性，第 10－24 行在深度森林模型中引入了 AdaBoost 集成方法。在开始训练前，先初始化特征的权重值(第 9 行)，然后根据深度森林每一级的分类误差率(第 15 行)，实现特征的权值更新(第 22 行)。在此过程中，还实现了对深度森林表征学习能力的改进(第 20 行)。

5 实验结果与分析

5.1 数据集

为了论证 BFDF 算法在文本特征提取以及文本情感分类方面的优势，本文选用谭松波收集整理的一个中文酒店评论语料集。该语料是从携程网爬取并经过整理而成的。该语料集包含评论文本 10 000 篇，其标签是二分类(positive/negative)，其中正面评论文本 7000 条，负面评论文本 3000 条，去除不能进行在线分析的文本，评论文本分别还有 6942 条和 2963 条，总计 9905 条。实验过程中，采用分割函数(分割参数选取为 0.2)将数据集的 80%用于模型训练，剩余 20%用于测试。为防止过拟合，实验采用 5 折交叉验证。

5.2 深度森林的参数选择

本实验首先考虑到本文提取的文本特征并不表示空间或者序列关系，因此不采用多粒度扫描窗口。本节采用的数据集为总的语料集，即 9905 条文本。

1)每个级联层随机森林的大小

级联森林的每一层是由若干个随机森林以及完全随机森林组成的，两种随机森林的大小会对预测结果产生直接影响，因此，为随机森林选择合适的值是关键所在，实验结果如图 3 所示。

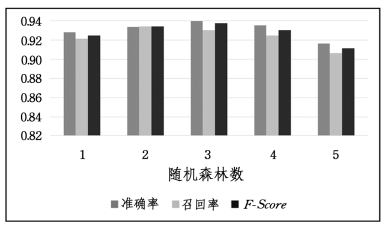

图 3 评价指标与随机森林数关系图

Fig. 3 Relationship between evaluation index and random forest quantity

由图 3 得知，当随机森林数 $m=3$ 时，准确率最高达到

94.53%，此时召回率为 93.02%，*F-Score* 值最高达到 93.77%。在 m 取其他值的情况下，所得效果均有降低。因此，本实验对该参数的选择为 $m=3$。

2）每个随机森林包含的决策树大小

由于级联森林最后的预测结果依赖于每棵决策树的预测结果，即对于给定的未知标签样本，其最终所属情感类别是由投票机制决定的。因此，为决策树选择合适的值是关键所在，实验结果如图 4 所示。

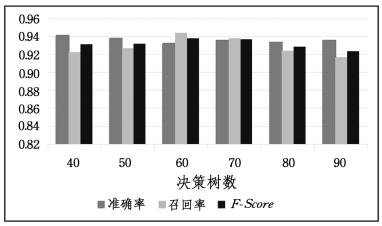

图 4　评价指标与决策树数的关系

Fig. 4　Relationship between evaluation index and number of decision trees

由图 4 得知，当每个随机森林中决策树的数量 n 为 40 时，准确率最高达到 94.09%。当 $n=60$ 时，准确率达到 93.23%，召回率达到 94.42%，*F-Score* 值最高达到 93.82%。因此，本实验中 n 设置为 60。

5.3　聚类算法中 k 值的选择

在对二元特征进行后处理的过程中，本文会对提取的评价对象求 *TF-IDF* 值，而后据此对评价对象做 k-means 聚类处理，将其分为 k 类，并与该类对应的情感极性一起，作为深度森林的 k 个输入特征。由于聚类算法中 k 值的选择决定了文本的二元特征数，进而会影响实验结果，因此需要找到一个合适的 k 值使结果达到最优。本文在负向语料集和正向语料集上均进行了实验，结果如图 5、图 6 所示。其中，负面评论文本有2963 条，正面评论文本有 6942 条。

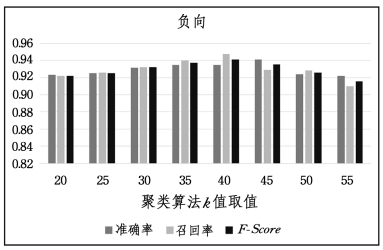

图 5　在负向语料集上准确率随 k 值变化的趋势图

Fig. 5　Trend of accuracy with k on negative corpus set

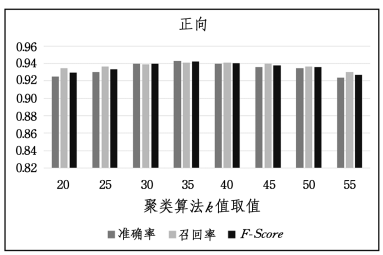

图 6　在正向语料集上准确率随 k 值变化的趋势图

Fig. 6　Trend of accuracy with k on positive corpus sets

由图 5 可以得知，当 $k=40$ 时，准确率达到 93.45%，召回率达到 94.74%，*F-Score* 值最高达到 94.09%。因此，确定负向语料集中 k 的取值为 40。

由图 6 得知，当 $k=35$ 时，准确率达到 94.31%，召回率达到 94.09%，*F-Score* 值最高达到 94.2%。因此，确定正向语料集中 k 的取值为 35。

5.4　模型对比实验

本节将本文提出的文本情感分类算法 BFDF 与原始模型 Deep Forest、SVM[19]、PCNN + Dropout[20]、LSTM-CSP-WE[21]、WLDA[22] 以及王娜娜提出的方法[23]，在此数据集上进行实验并对比结果，以式(14)－式(16)所示的准确率(Precision)、召回率(Recall)和 F-Score 作为评价标准。

$$Precision=\frac{N_r}{N_c} \tag{14}$$

$$Recall=\frac{N_r}{N} \tag{15}$$

$$F\text{-}Score=\frac{Precision\times Recall\times 2}{Precision+Recall} \tag{16}$$

其中，N_r 是正确分类的样本数，N_c 是已经分类的样本数，N 是样本总数。实验的对比结果如表 1 所列。

表 1　模型 BFDF 与模型 DF 的对比结果

Table 1　Compared results between BFDF model and DF model

（单位：%）

模型	DF(正)	BFDF(正)	DF(负)	BFDF(负)
准确率	91.81	94.31	90.95	93.45
召回率	90.95	94.09	92.24	94.74
F-Score	91.38	94.20	91.95	94.09

表 1 中，在完成文本特征提取后，将特征分别输入到原始模型 Deep Forest 与本文的模型 BFDF 中，结果显示，BFDF 的实验效果均高于原始模型，这说明本文对深度森林级联层改进后将 AdaBoost 集成方法融入到深度森林中是有效的。

文献[19]以构建的酒店领域情感词典为基础进行特征选择，并结合支持向量机 SVM 进行文本情感分类。表 2 中，对比文献[19]可知，本文的不足在于所用的情感词典是在原有情感词典的基础上进行扩充，但是该词典并不完全是针对酒店领域构建的，因此对情感词的识别不比文献[19]完善，从而给文本特征的提取造成了影响。

表 2　模型 BFDF 与模型 SVM 的对比结果

Table 2　Compared results between BFDF model and SVM model

（单位：%）

模型	正向(准确率)	负向(准确率)
SVM	86.73	85.45
BFDF	94.31	93.45

表 3 和表 4 的对比文献中均采用了神经网络模型。表 3 中文献[20]改进了传统 CNN 在卷积层的操作，将分段池化策略与 Dropout 算法相结合来实现情感分类。表 4 中文献[21]提出在 C&W 模型的基础上融入情感信息和词性信息来训练词向量，并使用 LSTM 模型实现情感分类。不过本文提到，相比于深度森林，神经网络模型的复杂度更高，训练参数

更多，而且由实验结果对比可知，本文方法对深度森林进行改进后更具有效性。

表 3 模型 BFDF 与模型 PCNN+Dropout 的对比结果

Table 3 Compared results between BFDF model and PCNN+Dropout model

（单位：%）

模型	准确率
PCNN+Dropout	91.00
BFDF	93.23

表 4 模型 BFDF 与模型 LSTM-CSP-WE 的对比结果

Table 4 Compared results between BFDF model and LSTM-CSP-WE model

（单位：%）

模型	LSTM-CSP-WE		LSTM-CSP-WE	
	（正）	BFDF（正）	（负）	BFDF（负）
准确率	86.41	94.31	85.76	93.45
召回率	85.33	94.09	86.33	94.74
F-Score	85.87	94.20	86.04	94.09

表 5 中文献[22]通过计算词汇与情感种子词的距离给词汇赋予不同的权重，然后利用关键词判断情感倾向，该算法属于无监督类型。而本文为了突出文本特征的重要性，利用 AdaBoost 方法给特征赋予初始权重，并且在训练过程中根据每一层的分类结果调整特征权值，提高了模型的分类准确率。

表 5 模型 BFDF 与模型 WLDA 的对比结果

Table 5 Compared results between BFDF model and WLDA model

（单位：%）

模型	正向（准确率）	负向（准确率）
WLDA	86.80	92.60
BFDF	94.31	93.45

表 6 中文献[23]以情感词典为基础，对传统的评价词-评价对象抽取方法做出改进，并根据对二者不同搭配组合的定义来实现文本情感分类。对比文献[23]，本文提取的特征是评价对象-评价词极性特征对，在提取规则上实现了对缺省值以及隐式评价对象的提取，此外还提取了情感语义概率特征作为文本的最终特征。之后的工作将继续对特征提取进行改进，以更好地实现分类。

表 6 模型 BFDF 与其他模型的对比结果

Table 6 Compared results between BFDF model and another model

（单位：%）

模型	文献[21]（正）	BFDF（正）	文献[21]（负）	BFDF（负）
准确率	84.92	94.31	67.43	93.45
召回率	78.83	94.09	61.54	94.74
F-Score	81.76	94.20	64.65	94.09

结束语 为了实现评论文本的情感倾向性预测，本文首先提取文本二元特征和情感语义概率特征，然后将二者融合作为文本的最终特征。为了避免信息的削减与特征维度的增加，本文改进了深度森林模型的表征学习能力，将级联层每一级的输入向量改进为将原始向量与之前每一级输出类分布向量的平均值进行结合。为了强化特征，使深度森林每一级都注意到特征的重要性，本文将 AdaBoost 集成方法应用于深度森林模型，从而更好地提升分类性能。

虽然本文方法取得了一定的成果，但还存在不足之处。首先在特征提取阶段，由于本文所用情感词典并不是针对酒店领域构建的情感词典，因此会导致情感词识别误差较大；其次，在使用聚类算法对评价对象进行处理时，由于提取出的评价对象并不都是酒店的特征属性，其中会掺杂一些无效数据，因此会对聚类结果产生影响；最后，在对模型的改进方面还需进一步优化。因此，下一步的工作将针对这几方面展开，并有望将深度森林模型与网络模型相结合，以更好地学习文本的特征表示。

参 考 文 献

[1] XU J F，XU Y，XU Y C，et al. Hybrid Algorithm Framework for Sentiment Classification of Chinese Based on Semantic Comprehension and Machine Learning[J]. Computer Science，2015，42(6)：61-66.（in Chinese）
徐健锋，许园，许元辰，等. 基于语义理解和机器学习的混合的中文文本情感分类算法框架[J]. 计算机科学，2015，42(6)：61-66.

[2] ZHANG D，XU H，SU Z，et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications，2015，42(4)：1857-1863.

[3] WU Y J，ZHU F X，ZHOU J. Using Probabilistic Graphical Model for Text Sentiment Analysis[J]. Journal of Chinese Computer System，2015，36(7)：1421-1425.（in Chinese）
吴钰洁，朱福喜，周竞. 基于概率图模型的文本情感分析[J]. 小型微型计算机系统，2015，36(7)：1421-1425.

[4] JO Y，OH A H. Aspect and sentiment unification model for online review analysis[C]// ACM International Conference on Web Search and Data Mining. ACM，2011：815-824.

[5] ZHOU Z H，FENG J. Deep Forest：Towards An Alternative to Deep Neural Networks[J]. arXiv：1702.08835v1，2017：2-3.

[6] ZHAO Q，WANG H，LV P，et al. A Bootstrapping Based Refinement Framework for Mining Opinion Words and Targets[C]// Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. 2014：1995-1998.

[7] LIU K，XU L，ZHAO J. Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model[J]. IEEE Transactions on Knowledge & Data Engineering，2015，27(3)：636-650.

[8] GU Z J，YAO T. Extraction and Discrimination of the Evaluated Object and Its Orientation[J]. Journal of Chinese Information Processing，2012，26(4)：91-97.

[9] KAMAL A，ABULAISH M，ANWAR T. Mining feature-opinion pairs and their reliability scores from web opinion sources[C]// International Conference on Web Intelligence，Mining and Semantics. ACM，2012：15.

[10] WANG S G，WU S H. Feature-Opinion Extraction in Science Spots Reviews Based on Dependency Relation[J]. Journal of Chinese Information Processing，2012，26(3)：116-121.（in Chinese）
王素格，吴苏红. 基于依存关系的旅游景点评论的特征—观点对

抽取[J]. 中文信息学报,2012,26(3):116-121.

[11] SARU,KETKI B M. A new approach towards co-extracting opinion-targets and opinion words from online reviews[C]// International Conference on Computational Intelligence & Communication Technology. IEEE,2017:1-4.

[12] FENG S,FU Y C,YANG F,et al. Blog Sentiment Orientation Analysis Based on Dependency Parsing[J]. Journal of Chinese Information Processing,2012,49(11):2395-240 6. (in Chinese)
冯时,付永陈,阳锋,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.

[13] XIAO H,XU S H. Analysis on Web Public Opinion Orientation Based on Syntactic Parsing and Emotional Dictionary[J]. Journal of Chinese Computer Systems,2014,35(4):811-813.

[14] UTKIN L V,RYABININ M A. A Siamese Deep Forest [J]. Journal of Knowledge-Based Systems, arXiv: 1704. 08715v1, 2017:5-6.

[15] ZHU X Y. Application of Deep Forest Model for Flame Detection[D]. Wuxi:Jiangnan University,2017. (in Chinese)
朱晓妤. 应用深度森林模型的火焰检测[D]. 无锡:江南大学,2017.

[16] WANG H Y. Dense Adaptive Cascade Forest:A Densely Connected Deep Ensemble for Classification Problems[J]. arXiv: 1804. 10885v1,2018:6-9.

[17] UTKIN L V,RYABININ M A. Discriminative Metric Learning with Deep Forest[J]. arXiv:1705. 09620v1,2017:4-8.

[18] YANG F,XU Q,LI B,et al. Ship Detection From Thermal Remote Sensing Imagery Through Region-Based Deep Forest[J]. IEEE Geoscience & Remote Sensing Letters,2018,15(3):449-453.

[19] SHI X. Research on Sentiment Classification Based on Semantic Lexicon of Hotel Field[D]. Baoding:Hebei University,2014. (in Chinese)
石馨. 基于酒店领域情感词典的分类器研究[D]. 保定:河北大学,2014.

[20] DU C S,HUANG L. Sentiment Analysis with Piecewise Convolution Neural Network[J]. Computer Engineering & Science, 2017,39(1):173-179. (in Chinese)
杜昌顺,黄磊. 分段卷积神经网络在文本情感分析中的应用[J]. 计算机工程与科学,2017,39(1):173-179.

[21] CHEN N N. Text Sentiment Analysis based on Deep Learning Methods[D]. Hangzhou:Zhejiang Gongshang University,2017. (in Chinese)
陈南南. 基于深度学习的文本情感分析技术研究[D]. 杭州:浙江工商大学,2017.

[22] HAO J. Research on Text Sentiment Analysis Based on Topic Model[D]. Taiyuan: Taiyuan University of Technology,2017. (in Chinese)
郝洁. 基于主题模型的文本情感分析研究[D]. 太原:太原理工大学,2017.

[23] WANG N N. Research on Sentiment Orientation Technology for Review Texts[D]. Beijing:Beijing Jiaotong University,2017. (in Chinese)
王娜娜. 评论文本情感倾向性分析技术研究[D]. 北京:北京交通大学,2017.