

融合词性特征的深度迁移学习商品评论情感分析

张震¹

ZHANG Zhen

摘要

情感分析是文本分类的研究方向,深度迁移学习通过学习目标领域数据和已有领域数据之间的相关度,提高当目标数据不足时文本分类的精度。从基于网络迁移的角度设计算法,首先使用 Word2vec+ 词性特征词向量表示,然后进行卷积神经网络文本分类,再将训练好的模型共享网络参数,迁移至跨域商品评论数据,训练、分类评论数据。实验证明,在小样本数据集中算法精度有明显提升。

关键词

深度迁移学习;卷积神经网络;词性特征;商品评论分析

doi: 10.3969/j.issn.1672-9528.2024.03.027

0 引言

随着网购用户的逐年增加,消费者在网购前会浏览产品及评价信息。其中商品评论是网购前决策的重要参考,它真实反馈了店铺、商品和快递服务的各类满意度,同时也有助于商家及时查找不足,以提升销量。在信息冗杂时代,海量文本数据特征提取、表示和分类,可应用于电商、微商的商品评论方面^[1]。目前常见的三种方法为基于规则、传统机器学习和深度学习^[2]。深度学习兴起之后,将卷积神经网络等方法应用于情感分类,利用网络结构自动获取特征,实现端到端表示大规模文本^[3]。与传统机器学习文本分类方法相比,深度学习文本分类之所以精度高,是因为需要海量文本数据学习来理解文本的潜在语义,因此对训练集的依赖性很大,而且构建一个规模质量好、标准高的标定数据集是非常困难的,当数据不足时还会造成欠拟合或过拟合的情况^[4]。

张博等学者^[5]认为,迁移学习是将一种模型或者知识应用于其他相关任务的方法,尤其面对构建模型所需规模数据无法达到的情况,借助迁移学习,可以将源模型领域训练结果,直接应用于目标模型。Pan S J 等学者^[6]十分认可迁移学习方法,他们进一步证明了在迁移学习算法中,训练数据和测试数据不一定非得完全独立,两者之间不一定非得完全相同,目标域中的模型也可以不再从头开始训练,这样一来,训练的数据量以及训练的时间均是可以减少的,从而能解决训练数据不足的问题。深度学习方法虽然已经占据主导地位,但是将其与迁移学习相结合,能达到互补的效果,也使得众

多学者开始逐渐对深度迁移学习的研究产生兴趣。Yosinski J 等学者^[7]最初对深度迁移学习下了定义并受到了广泛的认可,他认为该方法能有效地迁移源模型领域的目标知识,运用于不同的任务当中,再加上能自动化端到端提取到更具有表征能力特征的深度学习方法,使其在应用中效果会更好。起初深度迁移学习用于解决计算机视觉和图像的问题,但深层次来讲,自然语言处理方向使用起来也更加方便,尤其是面对数据较少的文本。Tan C 等学者^[8]受到深度迁移学习启发后,认为在文本分类的任务中,一些预训练好的模型或网络参数是构成深度迁移学习的基础,除此之外,还可通过实例、映射、网络和对抗性迁移方式完成文本分类任务。

基于以上的梳理和研究,本文重点研究利用深度迁移学习算法解决情感分析问题的方法,将迁移学习引入到深度神经网络模型中,一方面可以提高模型对其他外界环境的适应性,另一方面也保证了当数据不足时,文本分类的准确性不会降低。本文的方法在提高分类精度的前提下,减小数据运行,使得研究者能够更容易开展相关实验,具有重要的理论意义。另外,精准的商品评论分析为用户精准推荐相关商品信息,提供应用参考,根据分析结果来提高服务质量,调整推荐策略等,本文的方法在这些方面均具有很大的实践应用意义。

1 研究思路

本文根据深度迁移学习技术的特点以及文本分类问题的基本需求,设计在训练数据不足情况下的文本分类算法。同时,为了验证算法的性能和效果,拟采用具体的实验对算法进行实证研究。从深度特征提取、网络参数和权重的迁移、数据实例的分类、训练测试部署这四个阶段开展。具体思路为:一是结合目标任务给定语料资源类型,这样能避开大量

1. 河南工业贸易职业学院就业处 河南郑州 450000

[基金项目] 2022 年度河南省高等学校重点科研项目指导计划项目“基于深度学习的企业科技需求的分析及协同创新服务平台设计”(项目编号: 22B520007)

标定数据集的制作；二是将源领域与目标领域分别进行组合，然后利用词向量和词性特征双通道的增强深度卷积神经网络模型，来训练本研究的分类器；三是进行模型网络的迁移学习，并将迁移完成后的模型对目标领域数据分类；四是得到词向量文本表示，来实现特征提取器的网络迁移，用于分类商品评论文本数据的挖掘，更好地服务于商家和用户。

基于词性标注的文本表示模型：文本均是由多个词语组成，即可以表示为词向量序列^[9]。词性是词向量的语法属性，一般表示在句中属于哪一种句子成分。词性标注方法可以将每个词的词性给予标注，其标准选择用北大词性标注集。通过 Python 程序语言解释器环境下的 jieba 分词工具，结合基于规则和统计的方法，利用前缀词典法构建有向无环图后，计算最大概率路径，对文本进行词语的拆分，然后执行词典比对和 HMM 共同标注^[10]。词性标注完成后再通过词与词性的拼接技术，形成关于词的“W-POS”序列，并作为词向量模型的输入。

基于词性特征的卷积神经网络分类模型（POS-Text CNN）：传统的 CNN 文本分类模型使用卷积操作处理词向量序列，按照本文的研究思路加以改进，主要是在 TextCNN^[11] 模型的输入层的基础上再添加一项词性的内容，即“W-POS”序列的表示矩阵，通过与词向量矩阵两两结合，拼接成双模式的输入。随着词性特征文本表示的加入，特征表示会更加准确。经过原有模型的池化层和全连接层，形成语料集文本的定长向量表示，然后通过 Softmax 输出分类结果。以上过程即为双通道输出的 POS-Text CNN 模型。

模型迁移：模型是需要大量标注数据来获取和学习潜在的数据特征的，按照迁移学习思想，本文采用基于网络的迁移学习策略，通过网络模型训练得到源领域和目标领域中的一些通用的权重和参数（这一过程属于预训练）。随后，这些模型权重和参数可以复用在目标领域的学习上，因此能实现通用特征迁移。

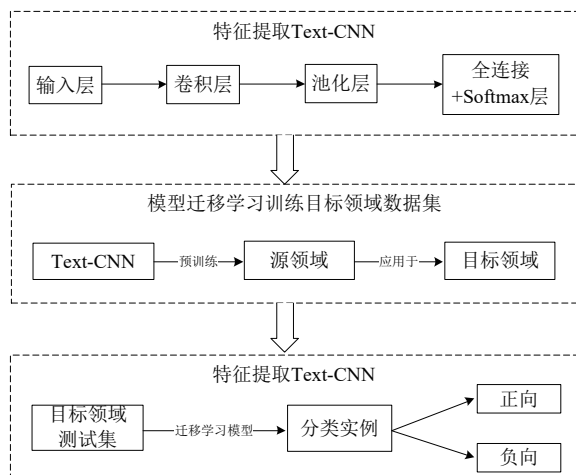


图1 研究思路

按照实验需要，现有规模并且标记的源领域数据集和小部分目标领域新数据集这两个数据集，有特征相同的地方，但不完全一致。下面将模型迁移方法列举如下。

一是由于预训练代价较大，执行一次预训练源领域数据得到一层模型的全部参数，即为预训练的结果。一次迭代即可能达到模型训练的性能和权重结果的收敛。

二是将除 softmax 层以外的预训练模型参数迁移到目标模型上，不包括 softmax 层是因为其通常包含特定域特征，如果参数也迁移，就会出现干扰或抑制的情况^[12]。

三是需要微调一些 softmax 层的特定参数，微调的依据为训练新目标任务数据集得到的结果。softmax 层参数固定后得到完整参数，再利用微调对权重进行测试。

2 实验分析

设计商品评论文本分类实验，基于此对情感分析方向的问题进一步展开挖掘。本研究借助卷积神经网络文本分类方法、Word2Vec 词向量模型、词性标注算法和迁移学习技术，实现深度学习特征提取器的网络迁移。研究的实验是将其应用于亚马逊商品评论数据上，按照情感正、负向的分类实验，来对比模型的精度是否有所提高。

2.1 实验环境

实验环境均在我校优胜校区 10 号楼大数据创新平台实训室开发完成，硬件设施齐全，配置先进。代码均采用 Python 程序设计语言进行开发，目前选择的框架是较为常用且完善的 Tensorflow^[13]。为更好地应对教学和产教融合需要，我校大数据创新平台实训室的集成环境是 Pycharm 专业版，足以应对本研究实验的实施和开展。

2.2 数据集说明

本文采用的是亚马逊商品评论的公开数据集。数据集挑选出来 Book (B)、Clothing (C)、Electronic (E)、Food (F) 四个方面的商品评价数据，并且这些评论信息分为积极和消极两个方向，便于开展商品评论正向和负向的分类。实验共分为两个方面，一是 POS-Text CNN 实验，二是 Tr-POS-Text CNN 实验。第一个实验是进行情感分类的，第二个实验是用商品名的英文首字母，构造 12 组迁移任务表示源领域数据到目标领域数据，对 12 组迁移任务进行跨域分类。

2.3 评价标准

当前在自然语言处理领域内，大多数研究者使用的依然是将精准率 Pre (Precision)、召回率 Rec (Recall)，二者综合值 F 值 (定量评价) 和准确率 (Accuracy) 几个指标来评价模型的效果^[14]。

2.4 对比实验说明

本文的第一个实验是经过 SVM 和 Skip Gram-CNN 两个模型的分类实验后，与加上词性特征后的 Skip Gram-CNN 进行对比。第二个实验是 Tr-POS-Text CNN 的实验，是采用对出全连接层以外的其它层预训练，得到稳定网络结构后迁移过来，实验中直接称为 Tr-POS-Text CNN。两个实验中，对网络中的参数和函数设置分别按照 256 个卷积核，窗口 (3,4,5) 的高度，通过 Relu 激活和最大池化的方法，被关闭神经元的比例按照 0.5 调参，正则化按照 L2 的范数，遍历 20 个周期。

2.5 基础实验

按照商品评论的情感分析要求，将评论数据在 SVM、Skip Gram-CNN 和加上词性特征后的 Skip Gram-CNN 模型上分别进行分析，判断评论的正负性。经过评价标准的计算公式得到计算结果，最后予以展示。

通过图 2 可以看出，四个方向数据情感分析上，深度学习的相关算法精准率明显高于传统的机器学习算法。另外，当加入词向量表示后精准率普遍得到提升，一旦通过词向量和词性双重输入后精准率提升的幅度更大。值得一提的是，图 2 得知，POS-Skip Gram-CNN 的综合评价 F_1 值提升在 1 个百分点以上。综上所述，深度学习相关算法对于传统机器学习来说是有提高的，而且在词性和词向量双重表示后，评论数据情感分类的效果更好。

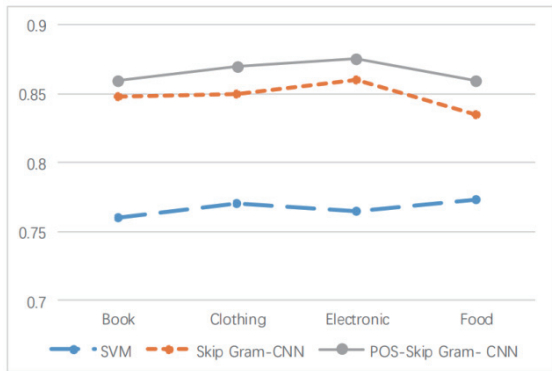


图 2 三种模型的 F_1 值

2.6 迁移学习实验

(1) 相似度计算

本文是通过文本之间的词频重叠性^[15]来计算相似度，假设定义两个域是 C_1 和 C_2 ，可以按照公式 (1) 的过程计算相似度。其中等式右边的分母指 C_1 、 C_2 中共有的文本数量，分子指的是两个域当中所有词的个数。

$$Sim_{overlap}(C_1, C_2) = \frac{2samewords(C_1, C_2)}{Len(C_1) + Len(C_2)} \quad (1)$$

通过计算四个方向商品的两两之间域相似度，得到的结果如表 1 所示。可以观察到以 B 为基准，C 的相似度最高；

以 E 为基准，F 的相似度最高，因此 B、C 互为一组，E、F 互为一组。

表 1 四种商品的域相似度

域相似度	B	C	E	F
B	1	0.564	0.481	0.495
C	0.564	1	0.474	0.477
E	0.481	0.474	1	0.532
F	0.495	0.477	0.532	1

(2) 实验分析

实验通过对比三种方法下 12 组跨域数据的情感分类准确率，通过结果来探究迁移效果。经过实验，得到的实验结果如表 2 所示。

表 2 文本分类准确率

Task	SVM	Tr-Text CNN	Tr-POS-Text CNN
B → C	0.779 5	0.805 5	0.827 5
E → C	0.704 5	0.753 5	0.774 5
F → C	0.740 5	0.767 0	0.799 8
B → E	0.724 0	0.766 0	0.789 7
C → E	0.805 0	0.840 5	0.868 8
F → E	0.736 0	0.785 5	0.804 7
F → B	0.697 5	0.769 0	0.769 0
E → B	0.681 5	0.730 5	0.754 2
C → B	0.731 0	0.762 2	0.819 5
E → F	0.813 0	0.857 5	0.890 0
B → F	0.745 5	0.804 5	0.832 5
C → F	0.724 5	0.791 0	0.815 0

一是与其它两种方法相比，本文的 Tr-POS-Text CNN 准确率提高了很多，尤其是与传统机器学习算法相比准确率提升高达 2.60% ~ 7.15%，因此验证了相关或相近领域知识迁移的有效性，词性表征商品评论数据后情感正负性分析的准确性提升的水平非常明显。

二是当两个域的相关程度或相似性越高的时候，不难发现，迁移之后商品情感分析的正负准确度会更高。比如以 Clothing 为目标域的迁移任务，这样形成的分组为 B → C，E → C，F → C。按照相似度计算，已经得出的是 B 和 C 域相似度最高，而经过表格对比发现，同组当中 B → C 的准确率也是最高的。同样，以 E 为目标域的迁移任务 F → E 的准确率最高，以 B 为目标域的 C → B 准确率最高，以此类推，以 F 为目标域 E → F 准确率最高。

三是运用迁移学习的方法，能够有效提高迁移后模型的时间复杂度和收敛程度。据实验记录记载，按照实验室服务器对算法复杂度的计算情况，本文算法相比 POS-Text CNN，同样的迭代次数，12 组实验的时间复杂度大约 13 min。

(3) 其他实验结果分析

为了更好地探究本实验参数设置的科学性和合理性,将参数敏感程度表达准确,主要涉及的参数包括卷积核、激活函数的使用和迭代的次数等。已知本算法的基准参数为:滑动窗口大小(3,4,5),数量为256,激活函数为Relu,Dropout rate为0.5。按照控制变量法的实验方法,通过学习和训练,最终得出,当滑动窗口训练到(6,6,8)这三个值,filter数量设置为512,激活函数为ReLU,Dropout值设置为0.6时,四种分类任务的准确率是最高的。

另一方面,目标数据量和准确率的变化情况呈纺锤型曲线,即算法准确率虽有所提高,但是如果目标域的数据量大于400就会出现平衡,准确率增加的加速度变得十分缓慢。如果到达1000条的量时,准确率不仅不再增长,反而还低于未迁移的商品评论情感分类模型。因此,迁移学习算法需要综合考虑目标数据量。

3 总结

深度迁移学习技术能解决情感分析问题,实现源数据域和目标数据域的特征提取,改善分类的适应性,解决文本训练数据不足、标记数据少而产生的分类不精确等问题^[16]。本文构建融合词性特征的深度迁移学习商品评论情感分析模型,针对当目标数据不足时情感正负性分类准确性不够高的问题,本文采用基于网络的迁移学习策略,通过迁移预训练模型中通用的权重和参数,然后复用到目标领域中,实现网络的通用特征迁移,使得特征表示更加准确,分类算法精度不断提高。

参考文献:

- [1] 李明, 胡吉霞, 侯琳娜, 等. 商品评论情感倾向性分析[J]. 计算机应用, 2019, 39(S2): 15-19.
- [2] 王紫薇. 基于文本挖掘的太原市电子政务投诉建议分析研究[D]. 太原: 中北大学, 2023.
- [3] DU C, HUANG L. Text classification research with attention-based recurrent neural networks[J]. International journal of computers communications & control, 2018, 13(1):50-61.
- [4] TIAN F, WU F, CHAO K M, et al. A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews[J]. Electronic commerce research and applications, 2016, 16:66-76.
- [5] 张博, 孙逸, 李孟颖, 等. 基于迁移学习和集成学习的医学短文本分类[J]. 山西大学学报(自然科学版), 2020, 43(4): 947-954.
- [6] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.
- [7] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[J]. Advances in neural information processing systems, 2014, 27: 27-35.
- [8] TAN C, SUN F, KONG T, et al. A survey on deep transfer learning[C]// International Conference on Artificial Neural Networks. Cham: Springer, 2018: 270-279.
- [9] ZHU W, YAO T, ZHANG W, et al. Part-of-speech-based long short-term memory network for learning sentence representations[J]. IEEE access, 2019, 7: 51810-51816.
- [10] 甘丽新, 万常选, 刘德喜. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2):284-302.
- [11] 黄泽民, 吴迎岗. 结合BERT和卷积双向简单循环网络的文本情感分析[J]. 计算机应用与软件, 2022, 39(12):213-218.
- [12] SEMWAL T, YENIGALLA P, MATHUR G, et al. A practitioners' guide to transfer learning for text classification using convolutional neural networks[C]// Proceedings of the 2018 SIAM International Conference on Data Mining. Philadelphia:Society for Industrial and Applied Mathematics, 2018: 513-521.
- [13] PANG B, NIJKAMP E, WU Y N. Deep learning with tensorflow: a review[J]. Journal of Educational and Behavioral Statistics, 2020, 45(2): 227-248.
- [14] 孔丽雅, 周治平. 基于ELMo的混合注意力网络的方面级情感分析研究[J]. 中文信息学报, 2023, 37(6):147-156.
- [15] 吴凤慧, 成颖, 郑彦宁, 等. 文本聚类中文本表示和相似度计算研究综述[J]. 情报科学, 2012, 30(4):622-627.
- [16] IMAN M, ARABNIA H R, RASHEED K. A review of deep transfer learning and recent advancements[J]. Technologies, 2023, 11(2): 40.

【作者简介】

张震(1994—), 男, 河南光山人, 硕士研究生, 助教, 研究方向: 自然语言处理、机器学习。

(收稿日期: 2023-12-15)