DOI:10.3969/j.issn.1000-1565.2015.06.016

基于多特征融合的中文评论情感分类算法

陈昀,毕海岩

(国网天津市电力公司 城东供电公司,天津 300010)

摘 要:为解决情感分类中词间的语义关系难以表达和分析的问题,提出了一种基于词向量(word representation)和支持向量机(support vector machine)的情感分类算法,对电子商务在线评论的情感分类问题进行研究.首先使用 word2vec 聚类相似特征,然后使用 word2vec 和 SVM 对情感数据进行训练和分类,并分别使用基于词特征和基于词性标注的方法进行特征选择.在京东评论数据上进行的实验结果表明,与现有方法相比,分类准确率和召回率得到了提高.

关键词:情感分类;词向量;支持向量机;机器学习

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-1565(2015)06-0651-06

A sentiment classification algorithm of Chinese comments based on multi features fusion

CHEN Yun, BI Haiyan

(Chengdong Electric Power Supply Company, State Grid Tianjin Electric Power Company, Tianjin 300010, China)

Abstract: To solve the problem that semantic relationships between words can not be well analyzed in sentiment classification, a method for sentiment classification based on word2vec and SVM is proposed to carry out the study of sentiment classification of E-commerce online reviews. First of all, we use word2vec to cluster the similar features. And then, we train and classify the comment texts using word2vec again and SVM. In the process, the lexicon-based and part-of-speech based feature selection methods are respectively adopted to generate the training file. We conduct the experiments on the data set of Chinese comments of jingdong. The experimental result indicates that the precision and recall of sentiment classification of using word2vec again and SVM are superior to those of using the traditional optimalization method.

Key words: sentiment classification; word2vec; SVM; machine learning

---

随着网络技术的迅速发展,电子商务在消费构成中变得尤为重要.为了吸引顾客和提升顾客的购物体验,电子商务网站引入了用户评论机制,一方面,评论内容中包含了非常有价值的商品信息,另一方面,用户又很难从海量评论数据中完整地了解商品的全貌,同时商品制造商也很难根据这些评论信息来改进商品的

收稿日期:2015-02-20

基金项目:国家自然科学基金资助项目(61375075);河北省自然科学基金资助项目(F2013201064)

第一作者:陈昀(1977-),男,天津市人,国网天津市电力公司工程师,主要从事电力工程技术方面的研究.

E-mail:20951518@qq.com

生产和设计.与基于主题的分类不同(通过关键词进行识别),情感分类技术可以自动地将评论信息分为正类和负类,并帮助消费者和生产商从海量评论数据中获得有用信息,受到了很多电子商务公司的追捧和很多研究者的关注.

情感分类的研究大致分为2类,分别是基于情感词典及规则的方法和基于监督和半监督的机器学习方法.Turney等[1]针对情感字典的不足,使用PMI方法对基准字典进行了扩充;李寿山等[2]利用标签传播算法构建覆盖领域语境的中文情感词典用于文本情感分析;唐慧丰等[3]利用不同的特征选择方法组合多种机器学习算法验证情感分类的精度;杨经等[4]通过提取分析情感词的相关特征,使用SVM对句子进行情感识别及分类;李素科等[5]针对监督学习分类的不足,对情感特征进行聚类并提出了一种半监督的情感分类算法.然而,语义特征在情感分类中却很少被考虑到,事实上,语义特征可以揭示词间的深层次和隐含语义关系,从而提升情感分类效果.

1 相关工作

1.1 相似特征聚类

特征聚类的目的是将描述商品特征的同义词划分到同一组中,Zhai等[6]通过使用基于词共现和词间相似度的方法,使用半监督的EM算法来解决此问题.通过允许标注实例改变类别的方式来提高准确率,但是还是无法达到实用系统的目的.也有一些算法从评论文本中抽取商品特征,并对相似特征进行聚类,但在中文情感分类领域,相关工作还较少.

1.2 基于监督式机器学习的情感分类

监督式机器学习的情感分类算法希望通过标注语料来训练出情感分类模型.Pang等[7]第1次将这种方法应用于情感分类领域,他们尝试使用n-grams模型和SVM分类模型并选择unigrams作为特征来获取最佳分类结果.近些年提出了多种多特征选择方法和分类模型,Yao等[8]使用统计机器学习方法进行特征选取和降维来完成中文宾馆评论数据的在线情感分类;Moraes等[9]在词袋模型中采用标准评估上下文和监督式方法进行特征选取和分配权重;Wang等[10]综合使用文档频率、信息增益、卡方分布和互信息来进行特征选取,并应用布尔权重方法来分配权值从而构造向量空间模型.

1.3 word2vec和SVM^{perf}

word2vec是谷歌于2013年开发的深度学习工具包,该工具包主要采用2种模型架构,continuous bag-of-words (CBOW)和continuous skip-gram model来学习获得词向量.CBOW通过上下文来预测当前词汇,skip-gram则通过当前词汇来预测周围词汇[11].

SVM^{perf}是支持向量机的工程化实现版本,SVM^{perf}实现了SVM二值分类优化问题的替代结构化公式,更为重要的是SVM^{perf}利用cutting-plane subspace pursuit (CPSP)算法来训练稀疏核SVM,从而提高预测的速度和准确率.

2 多特征融合的中文评论情感分类算法

Wordvec在中文分类和英文文本聚类上表现出了优异的性能,但是目前还没有研究表明Wordvec在中文文本分类上同样具有良好的性能,因此,本研究首先使用Wordvec在同一特征组中对同义词进行聚类,然后联合使用Wordvec和SVM^{perf}对评论文本进行正类和负类的划分,图1展示了本研究的主要框架.

2.1 相似特征聚类

用户可能使用很多不同的词汇加以描述同一个商品特征,为了生成有效的评论摘要,这些近义词需要聚到同一个特征组中,使用word2vec来对相似特征进行聚类,分为如下几个步骤.

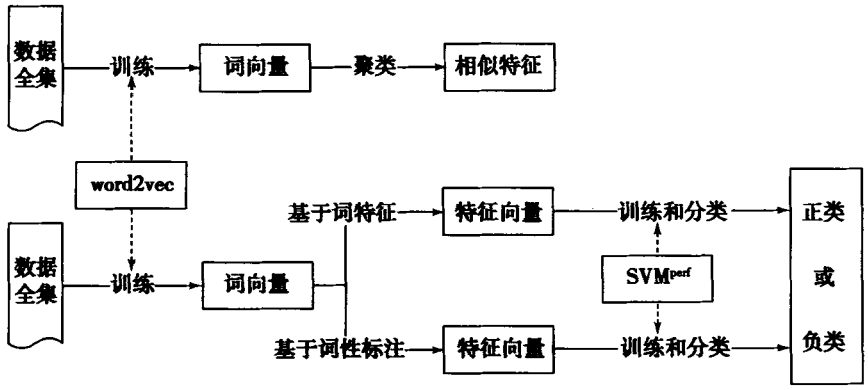

图 1　情感分类框架

Fig. 1　Framework of sentiment classification

1)预处理:通过使用中科院计算所的 ICTCLAS 的分词系统对中文评论文本进行分词和词性标注,去除停用词和标点符号后,生成所需的训练文件;

2)模型训练:使用 word2vec 训练模型文件,表 1 中给出了模型训练中所用参数和它们的解释. word2vec 以训练文件作为输入,并输入 1 份模型文件,首先从训练文件生成词表字典,然后学习生成词的高维词向量表示;

3)聚类:训练完成后每个词的词向量都存储在模型文件中,word2vec 提供了叫做"distance"的命令,该命令通过 2 个词汇词向量间的余弦相似度对它们间的语义距离进行计算,从而达到对近义词进行聚类的目的,余弦相似度的值越高,则 2 个词汇在语义层面的距离越近. 通过对结果进行降序排列,就会获得输入词最近似的词的列表.

表 1　模型训练的主要参数

Tab. 1　Main parameters of model training

参数	释义	默认值
−Train	输入文件的名称	Train. txt
−Otput	输出文件的名称	Vectors. bin
−cbow	训练模型的类型:0:Skip−gram model;1:CBOW model	0
−Size	向量维度的大小	200
−Window	上下文的大小	5
−Negative	训练方式的类型:0:Hierarchical Softmax method; 1:Negative Sampling method	0
−hs	训练方式的类型:0:Negative Sampling method; 1:Hierarchical Softmax method	1
−Sample	采样阈值	1e−3
−Threads	线程数量	12
−Binary	存储格式:0:普通格式;1:二进制格式	1

2.2　情感分类

与传统情感分类方法不同,主要采用 word2vec 和 SVM^{perf}作为分类工具. 首先,使用 word2vec 去除训练语料中词频低于 5 的词汇,其余频繁词汇则作为候选特征集合,通过 word2vec 训练出包含频繁词汇及其对应的特征模型文件,并使用基于词汇的特征选择方法和基于词性标注的特征选择方法来获取最优候选特征集.

2.2.1　基于词汇的特征选择方法

该方法需要 1 份情感词汇词典,词典包含情感词汇(肯定和否定)及其对应的权重[12],本工作选择从 HowNet 在线知识库中抽取的词集作为情感词典. 首先从词典中选择权重最高的是个情感词汇作为初始输

入,使用 word2vec 的 distance 命令来获得更多的情感词汇,通过该方法,对初始词典进行扩充.

选择同时出现在候选特征集和扩充词典中的特征作为最终的训练特征,特征选取过程如下,其中,feature_set 表示最终的训练特征集:

①*word_set*←frequent words
②dic_set←opinion words in lexicon
③**for** each w in dic_set **do**
④**if** w is in word_set then
⑤add w to feature_set
⑥**else**
⑦continue
⑨end **if**
⑩end **for**

2.2.2 基于词性标注的特征选择方法

该方法根据词性标注进行特征选取,不同标注的选取会直接影响特征选择结果[13].例如,只选取形容词作为特征的结果就不如同时选择副词、动词和形容词的作为特征的结果,这是由于多种词性标注的词汇会成为情感标签.

在该方法中,经过词性标注后,选择形容词、副词、动词和名词作为特征,并将它们之间的不同组合作为训练特征.

2.2.3 训练和分类

在该步骤中,被选取的特征向量被用来训练分类器,从而预测测试文档的情感极性(肯定和否定).之前的很多研究表明,与其他分类系统相比,SVM 在分类性能和系统健壮性上都表现出了很大的优势,基于此,本工作选择 SVM 作为分类工具.

SVM^{perf}是SVM^{light}的优化版本,总体架构沿袭了SVM^{light},但是升级了核算法,也使其具备了更快速和更准确的分类速度,因此,采用SVM^{light}作为训练和测试工具集.

3 实验结果与分析

3.1 实验数据集

实验从京东上爬取了 110 000 中文衣物商品评论信息,经过去除重复和无意义数据后,有效数据共 96 548条.本研究中,基于 word2vec 的相似特征聚类并不需要确定文本极性,语料库越大,训练越充分,聚类效果也越好.所以采集的所有评论语料都用来进行特征聚类.

本文的主要工作是进行基于 word2vec 和SVM^{perf}的监督式情感分类,采集的评论语料都是根据五星打分的,并将五星评价的语料作为正例,一星评价的语料作为负例.为了进行实验,将数据集分为 2 份,选取 2 500正例和2 500负例作为训练集,其余作为测试集.

3.2 实验结果

采用准确率(precision)、召回率(recall)和 F1 值作为实验结果的评价标准,通过实验对相似特征聚类和情感分类 2 项任务进行评估.

3.2.1 相似特征聚类结果

对于中文商品评论,选取在中文衣物评论中出现频率最高的"价格"、"面料"、"尺码"和"款式"作为典型特征,在获取同义词列表后,只保留排序最靠前的 5 个词汇,通过 3 个不同维度向量来训练 word2vec.

表 2 展示了相似特征聚类的结果,对每个典型特征,它的相似特征具有和它相似或相同的中文语义,不同维度的聚类结果并没有很大差别,只是在次序上稍有不同,该结果展现了 word2vec 较为强大的在中文文本聚类中获取深层次语义的能力.

表 2　典型特征聚类结果

Tab. 2　Main parameters of model training

典型特征	词向量维度	相似特征				
价格	200	价位	价钱	价码	价值	天价
	500	价位	价钱	价码	价值	天价
	1000	价位	价钱	价码	价值	天价
面料	200	料子	布料	质地	材质	材料
	500	料子	布料	材料	质地	材质
	1000	料子	布料	质地	材质	衣料
尺码	200	尺寸	号码	型号	码号	码子
	500	尺寸	号码	码子	码号	型号
	1000	尺寸	号码	型号	码号	码子
款式	200	样式	款型	外观	样子	外形
	500	样式	款型	样子	外形	外观
	1000	样式	款型	外观	样子	外形

3.2.2　情感分类结果

本研究采用基于 word2vec 和 SVM^{perf} 的情感分类算法,其中使用了 2 种特征选择方法,分别是基于词特征和基于词性标注. 表 3 列出了基于词特征的特征选择方法的性能,选取 HowNet 作为特征词的来源,并对其中的已标注特征词进行分类预测,结果如表 3 所示.

表 3　基于词特征的情感分类结果

Tab. 3　Sentiment classification results based on lexicon

特征	肯定			否定		
	准确率/%	召回率/%	$F1$/%	准确率/%	召回率/%	$F1$/%
HowNet	83.29	84.57	83.92	84.62	83.98	84.29

表 4 列出了基于词性标注的特征选择方法的性能,由数据可以看出,选择形容词、副词和动词作为特征的实验结果明显优于其他组合. 只选择形容词和副词的结果最差,选择全部特征的方法在正例中取得了最高的准确率以及负例中最高的召回率;但是较低的正例召回率和负例准确率拉低了整体 $F1$ 值. 其余 2 种策略获得了相近的结果.

表 4　基于词性标注的情感分类结果

Tab. 4　Sentiment classification results based on part-of-speech

特征	肯定			否定		
	准确率/%	召回率/%	$F1$/%	准确率/%	召回率/%	$F1$/%
形+副	86.19	82.78	84.45	82.89	86.54	84.68
形+副+动	91.23	88.87	90.03	88.43	91.51	89.94
形+副+名	90.63	84.64	87.53	86.28	91.53	88.82
形+副+动+名	91.12	83.73	87.17	85.69	91.42	88.46
全部	92.32	83.73	8782	84.32	92.57	88.25

从上述实验结果可以看出,基于词特征和基于词性标注的情感分类方法都可以取得较好的分类效果,这主要基于如下原因:首先,word2vec 的词向量表示方法可以学习到词间的深层语义,从而可以提升分类效果;其次,基于 SVM^{light} 的 SVM^{perf} 在大规模数据集上也表现出了更好的准确性和更快的处理速度,基于此,所提情感分类方法才取得了较好的实验结果.

4　结束语

与传统情感分类方法关注词特征和句法特征不同,本研究主要关注词间的语义特征,主要使用了

word2vec 和 SVM^{perf}2 种工具来对中文评论文本进行分类,首先使用 word2vec 对相似特征进行聚类,结果表明 word2vec 同样适用于中文特征选择,不管采用基于词特征的方法还是基于词性标注的方法,所提方法都取得了较好的实验结果.

即使本研究取得了较好的实验结果,但距离最好的结果还有很大的距离,为了训练出可用于 SVM^{perf}的文件格式,牺牲了 word2vec 的向量维度,如何将高纬度 word2vec 文件使用来 SVM^{perf}模型进行训练,还有待研究.另外文本所使用的 2 种词特征选择方法还不足以找出句子中的所有情感特征,词特征的抽取方法也是下一步研究的重点方向.

参 考 文 献:

[1] TURNEY P D, LITTMAN M L. Measuring praise and criticism inference of semantic orientation from association[J]. ACM Transon Information Systems, 2003, 21(4): 315 - 346.

[2] 李寿山,李逸薇,黄居仁,等. 基于双语信息和标签传播算法的中文褒贬词典构建方法[J]. 中文信息学报,2013,27(6): 75 - 80.
LI Shoushan, LI Yiwei, HUANG Juren, et al. Construction of Chinese sentiment lexicon using bilingual information and label propagation algorithm[J]. Journal of Chinese Information Processing, 2013, 27(6): 75 - 80.

[3] 唐慧丰,谭松波,程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报,2007,21(6):88 - 94.
TANG Huifeng, TAN Songbo, CHENG Xueqi. Research on sentiment classification of chinese reviews based on supervised machine learning techniques[J]. Journal of Chinese Information Processing, 2007, 21(6): 88 - 94.

[4] 杨经,林世平. 基于 SVM 的文本词句情感分析[J]. 计算机应用与软件,2011,28(9):225 - 228.
YANG Jing, LIN Shiping. Emotion analysis on text words and sentences based on SVM[J]. Computer Applications and Software, 2011, 28(9): 225 - 228.

[5] 李素科,蒋严冰. 基于情感特征聚类的半监督情感分类[J]. 计算机研究与发展,2013,50(12):2070 - 2577.
LI Suke, JIANG Yanbing. Semi-supervised sentiment classification based on sentiment feature clustering[J]. Journal of Computer Research and Development, 2013, 50(12): 2070 - 2577.

[6] ZHAI Zhougwu, LIU Bing, XU Hua, et al. Grouping product features semi-supervised learning with soft-constraints [Z]. The 23rd International Conference on Computational Linguistics: Association for Computational Linguistics, Beijing, China, 2010.

[7] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up: sentiment classification using machine learning techniques [Z]. The ACL-02 conference on Empirical methods in natural language processing-: Association for Computational Linguistics, Pennsylvania, USA, 2002.

[8] YAO Jiani, WANG Hongwei, YIN Pei. Sentiment feature identification from Chinese online reviews: Advances in Information Technology and Education[M]. Berlin: Springer, 2011: 315 - 322.

[9] MORRAES R, VALIATI J F, NETO W P G. Document-level sentiment classification: An empirical comparison between SVM and ANN[J]. Expert Systems with Applications, 2013, 40(2): 621 - 633.

[10] WANG Hongwei, YIN Pei, ZHENG Lijian, et al. Sentiment classification of online reviews: using sentence-based language mode[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2014, 26(1): 13 - 31.

[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv, 2013, 23(1):1301 - 1306.

[12] LIU Bing. Sentiment analysis and opinion mining[M]. San Rafael: Morgan & Claypool Publishers, 2012: 1 - 167.

[13] LIU Bing, ZHANG Lei. A survey of opinion mining and sentiment analysis[J]. Mining Text Data, 2012, 5(2):415 - 463.

(责任编辑:孟素兰)