文章编号：1003-0077(2012)04-0003-06

基于单层标注级联模型的篇章情感倾向分析

李本阳，关 毅，董喜双，李 生

（哈尔滨工业大学 计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘 要：情感分类是目前篇章情感分析的主要方法，但该方法存在难以融入中文结构特征的问题。针对此问题，采用级联模型对篇章情感倾向进行分析，将篇章情感倾向分析分为两层：小句级和篇章级，对篇章情感倾向分析引入小句级的情感分析。该文使用最大熵模型处理小句级情感分类，小句级的输出作为上层篇章级的输入，并结合句型特征和句子位置等信息作为特征，采用支持向量机模型进行篇章级情感分类。同时对于级联模型中双层标注问题，基于交叉验证的思想提出了单层标注级联模型，避免了多层标注工作以及错误。实验结果表明，该方法的准确率较传统情感分类方法提高了 2.53%。

关键词：情感倾向分析；情感分类；级联模型；最大熵；支持向量机

中图分类号：TP391 **文献标识码**：A

Single-label Cascaded Model for Document Sentiment Analysis

LI Benyang, GUAN Yi, DONG Xishuang, LI Sheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Classification is the main method to analyze the document sentiment polarity, but it is defected in its deficiency in integrating the structure features. A cascaded model for sentiment polarity analysis is proposed to address this issue, which consists of two levels: the clause level and the document level. The document is first segmented into clauses which are classified into positive and negative categories by an Maximum Entropy model. Afterwards, these categories are combined with types and positions of clauses as features for document classification via the Support Vector Machine model. Meanwhile, a Single-label Cascade Model based on cross-validation is proposed. Experimental results prove that the accuracy of the proposed method is improved by 2.53 compared with traditional methods of sentiment classification.

Key words: sentiment analysis, sentiment classification; cascade model; ME; SVM

1 引言

情感分析目的是确定文本所表达的态度或观点[1]，近几年来已经成为信息检索和自然语言处理领域的一个热点问题。情感分析分为两个方面：情感(emotion)和情感倾向(sentiment/opinion)。这两个方面都是人物主观意愿的反映，情感表达人物自身的情绪起伏，例如，快乐、悲伤等；情感倾向则表达人物对外界事物的态度或者喜爱的程度，例如，赞成、反对等[2-3]。情感分析一般分为四个层级，即词级，短语级，句子级和篇章级。本文是针对篇章情感倾向的研究，篇章情感倾向分为两类：支持和反对。

情感分析应用十分广泛，在舆情分析、有害信息过滤、影视评价、产品调查等方面都有广阔的应用前景，但是目前公开的情感分析语料还比较少，给研究带来很大的困难。作者参加了中文倾向性分析评测会议 2009 词级与句子级情感分析的评测，并且在句

收稿日期：2011-04-06 定稿日期：2011-07-11

基金项目：国家自然科学基金资助项目(60975077,90924015)

作者简介：李本阳(1985—)，男，硕士研究生，主要研究方向为自然语言处理；关毅(1970—)，男，教授、博士生导师，主要研究方向为智能化信息检索、网络挖掘、自然语言处理、认知语言学；董喜双(1981—)，男，博士研究生，主要研究方向为自然语言处理。

子情感分析中获得第一名[4]，但是从评测效果上看，目前情感分析还是比较初步，要使情感分析达到可应用的程度，还有很长的路要走。

2 相关工作

2.1 情感倾向分析相关工作

国内外许多学者对篇章情感倾向分析做了相关研究，这些研究大致可以分为三类：情感分类方法对比研究、情感分类特征研究和其他方法。在情感分类方法对比研究方面：Pang[5]通过实验比较了特征表示为二值或者词频以及分类模型选择NB(Naïve Bayes)、ME(Maximum Entropy)、SVM(Support Vector Machines)的效果，实验结果表明选择二值的特征表示和SVM分类器效果最好。唐慧丰[6]对比了n-gram作为文本特征，以互信息、信息增益、CHI统计量和文档频率作为不同的特征选择方法，以中心向量法、KNN、Winnow、NB和SVM作为不同的文本分类方法的效果，得到的结论是，在足够大训练集和选择适当数量特征的情况下，采用Bigrams特征表示方法、信息增益特征选择方法和SVM的情感分类取得较好的效果。徐军[7]对词频和二值的特征表示、NB和ME作为分类模型在新闻情感分类的效果做了探讨，结果表明使用二值的特征表示和ME做分类模型可以取得更好的效果。在情感分类特征研究方面：Matsumoto[8]在情感分类中，提出频繁词序列和频繁子树作为情感分类特征提高情感分类效果的方法。陈锦禾[9]把识别的情感词作为情感分类的重要特征，提出了先提取情感词，再利用情感词分类的方法。在其他方法方面：Turney[10]提出了基于语义学信息的无监督方法，进行篇章级情感倾向分析。Pang又在文献[1]中提出了一种使用最小割的方法，区分情感句和非情感句，并对区分后的句子进行分析，获得篇章的情感倾向。McDonald[11]提出了由细到粗使用结构模型的方法对多个层级进行情感分析，取得了比任意单层情感分类更好的效果。李钝[12]研究了基于一个或多个中心的短语模式，并通过短语模式进行情感倾向分析。

邢福义[13]在《小句中枢说》指出小句是“最小的具有表述性和独立性的语法单位”。因此获得小句级情感倾向对于篇章情感倾向分析有很大帮助。小句是指单句、复句中的分句、相当于充当句子成分的主谓短语、相当于在思维中可完形为小句的其他单位[14]。单句和复句中的分句是小句主要组成。本文经过两层分析获得篇章情感倾向，即小句级和篇章级。

McDonald注意到了不同层级之间相互影响，提出了使用联合结构模型，对句子级和篇章级同时进行情感分析的方法，所使用的模型是Collins[15]提出的感知器模型。使用当前句、前一句子以及篇章的类别联合作为特征，对于每个篇章类别，句子类别序列可以使用维特比进行解码，在取得概率最大的句子类别序列的同时，也确定了篇章的类别。联合模型使得篇章级的类别和句子类别相互影响，在两个层级都产生了更好的效果，句子级的提高更明显。

本文方法与McDonald的方法虽然都关注层次模型，但有本质的不同。

(1) McDonald关注的是句子和篇章的相互影响，一同产生两个层次的情感类别，而本文是针对篇章级的情感分析，关注的是细化到小句并融入小句信息对篇章级情感分析的提高。因此本文提出单层标注的级联模型，只需要篇章级的标注。

(2) McDonald使用的是基于感知器的联合结构模型，而本文使用的是单层标注级联模型。联合模型使用多层的类别联合作为特征，使特征数量激增，而单层标注级联模型将划分下层特征和上层特征，下层的输出作为上层输入，特征数量将少于联合模型，因此在时间上比联合模型有优势。

(3) 本文根据“小句中枢理论”，在模型的下层为小句，而非句子，这样更容易加入句型等信息。同时本文引入了小句句型以及位置等信息，使得篇章级的情感倾向分析效果得到提高。

本文使用级联模型对篇章情感倾向进行分析。首先使用ME获得小句情感分类结果，再使用小句级分类结果与句型、句子的位置等信息相结合作为上层篇章级的输入特征，进而用SVM模型对篇章级进行情感分类。这一方法有三个特点：1)处理层面细化到小句，使得小句结构更加明确；2)级联模型使得加入句型及位置等句子信息更容易；3)本文提出的单层标注的级联模型，克服了级联模型对多层次分类需要每个层次的标注的问题。本文情感倾向分析方法结果准确率，由使用ME的结果0.849提高到了0.874。

2.2　相关模型简介

2.2.1　最大熵模型

文献[16]在 1957 年基于香农信息熵理论建立了最大熵模型。信息熵代表信息系统的测度和可信度。在一定的限制条件下，选择一个系统的最优分布时，如果这些限制条件无法确定唯一的系统分布，那么最好的分布就是在满足所有限定条件下，系统的信息熵最大的分布。经过公式转化得到式(1)：

$$P(t \mid h) = \frac{\exp\left(\sum_i \lambda_i f_i(t,h)\right)}{Z(h)} \qquad (1)$$

其中 f_i 为模型的第 i 个特征函数，λ_i 是特征 f_i 的权重，$Z(h) = \sum_t \exp\left(\sum_i \lambda_i f_i(t,h)\right)$ 为归一化因子，t 代表一个特定的状态，h 代表该状态上的上下文观测值。训练的过程就是用数值算法求得 λ_i 值的过程。

2.2.2　支持向量机模型

支持向量机模型(SVM)是基于统计学习理论(SLT)发展起来的通用分类模型，其核心内容是在 1995 年提出的[17]，目前还在不断发展阶段。较好地解决了小样本、非线性、高维数、局部极小点等实际问题[18]。

支持向量机模型在进行分类时分两种情况：线性可分和线性不可分。对于线性可分的情况，寻找支持平面，并选择最优分类平面。最优的分类平面就是要求分类平面不但能将两类正确分开(训练错误率为 0)，而且使分类间隔最大。对于线性不可分的情况，支持向量机采用不同的核函数将其映射到高维空间成为线性可分问题，这样就很好地解决线性不可分的问题。

3　基于单层标注级联模型的情感分析

3.1　单层标注级联模型

本文将篇章情感倾向分析分为两层：小句级和篇章级。如果使用级联模型进行情感分析，两个层级都使用有监督学习模型，需要对每个层级标注。但是，大多数应用只需要篇章一个层级的情感倾向。例如，通过电影评论获得某部电影评价，不需要小句级的电影评价，只需对多篇评论进行情感分析就可以知道大家对这部电影的喜爱程度。如果仍然使用多层标注，就会带来更多工作量。因此本文提出了单层标注的级联模型，使用上一个层级篇章级的类别标注篇章内的所有小句，再使用小句级模型对小句的倾向进行重新的预测，篇章级模型会利用小句预测结果作为输入进行篇章分析。

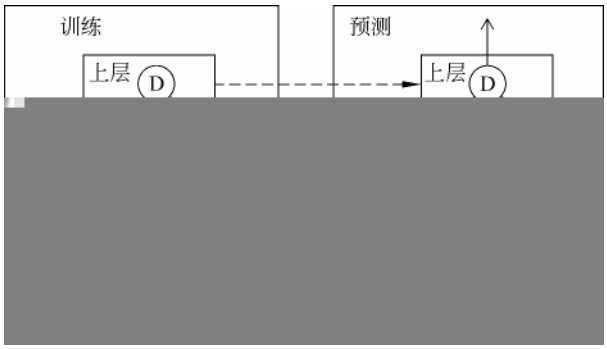

图 1　级联模型结构图

如图 1 所示，级联模型分为训练和预测两部分。在单层标注级联模型中，下层使用篇章类别标注的小句级训练语料训练小句级模型，再利用该模型对小句级语料进行重新预测分类。之后，使用小句级的预测结果作为上层输入特征训练篇章级分类模型。为了获得训练部分所有小句的预测结果，本文使用了交叉验证方法的思想，如下所示。

```
input:D=d1,d2...dn;
(1) predict_all=NULL;
(2) Fold=splitFold(D,v);
(3) for i=1 to v
(4)     trainClause(D-Fold[i],model);
(5)     result=predictClause(Fold[i],model);
(6)     predict_all+=result;
(7) endfor
(8) return predict_all。
```

其中，D 表示篇章级训练语料，由篇章 $d_1, d_2 \ldots d_n$ 组成，而每个篇章由若干小句组成，predict_all 保存全部训练语料的预测结果，Fold 保存分份后的语料结果。为了获得训练数据的预测结果，交叉验证方法首先使用将篇章级训练语料分成 v 份，i 从 1 循环到 v，做如下操作：除了第 i 份以外其他部分语料作为训练语料，用以训练小句级分类模型；第 i 份语料作为预测部分；将分类产生小句预测结果保存在 predict_all 中。最终返回 predict_all。

此处使用交叉验证方法是为了获得所有训练语料中小句预测类别。所选取的份数 v 同实验中训练和预测所分的总份数相同，这样尽量保持训练语料和预测语料小句预测结果分布是相近的。本文实验中将 v 置为 4，三份做训练一份做预测。这使得上层训练和预测的小句特征都是下层输出预测类别，

就解决了训练和预测语料上层输入特征分布不一致以及训练语料上下层标注一致起不到分类作用的问题。

通过上面的过程可以看到，本文直接使用篇章级情感标注小句级的情感，然后选取部分标注小句作为训练语料，对另一部分小句进行预测，重复上一过程，获得所有小句的预测结果，最终篇章级情感预测使用的是小句级的预测结果做特征。这样就使得细化到小句的篇章情感倾向分析只需要篇章单层的标注结果。

使用篇章级的情感直接标注小句，本身是会有一些偏差的，因为在一个倾向篇章中可能含有另一个倾向的小句，实验中发现经过一次预测可以有效地纠正这种错误，同时被误判的句子很多表现出言外的反向倾向意义，例如，“虽然房间很整洁”，这个小句更容易出现在反对情感的篇章中。小句标注错误对篇章倾向预测的影响在本文实验中表明是有限的。

3.2 基于单层标注级联模型的情感倾向分析方法

篇章情感倾向分析大多使用有监督分类模型（如最大熵，支持向量机等）进行情感分类，这类方法重点是提取合适的特征，而且能够达到较高的准确率。但是情感分类方法存在着难以细化结构并融入结构特征的问题。例如，在篇章情感分类中，多个词汇出现的小句范围、句子的句型和小句在篇章中的位置等信息都难以反映出来并融入模型。如果把这些信息与相应的上下文特征联合作为特征，会使特征空间增大，而引起上下文特征稀疏问题。然而细化篇章结构可以使篇章的结构更加清晰，同时更容易加入相应的结构特征，有助于分析整个篇章的情感倾向。本文使用单层标注级联模型将篇章情感倾向分析细化到小句级，再由小句级结果融入更多的句子信息过渡到篇章级情感倾向分析。

如图2所示，本文使用单层标注级联模型进行情感倾向分析过程如下。

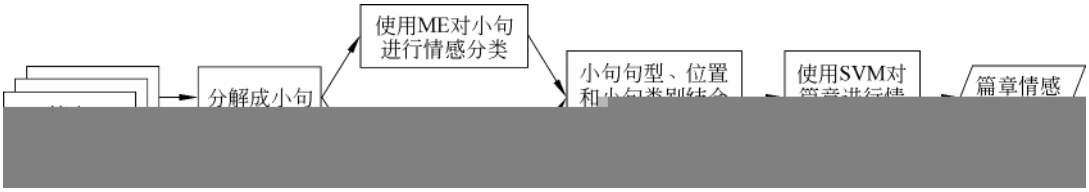

图2 级联模型情感分析过程

（1）分句。首先根据篇章中的标点符号以及一些纠正规则将篇章级语料分解成小句。本文中划分小句的标点符号有：“……”、“！”、“!”、“？”、“?”、“。”、“.”、“；”、“;”、“：”、“:”、“，”、“,”。纠正规则包括重复标点的去除等。

（2）小句级情感倾向分析。去掉不含有情感特征的小句，使用最大熵模型对含有特征的小句进行分类，获得相应的情感倾向类别。

（3）结合句子特征。由于不同句型将会对篇章情感产生影响，同时小句所在位置不同对篇章情感的影响程度也不同，将小句的预测结果与小句句型、位置等特征相结合，作为上一层篇章级情感倾向分类的输入特征。

（4）篇章级情感倾向分析。使用支持向量机模型利用结合后的特征对篇章进行分类，得到了最终的篇章情感倾向分类结果。

本文选取情感词等作为小句级特征对小句进行情感分类，选取句子句型、小句位置等信息与小句类别结合作为篇章级特征进行篇章级情感分类。这样的特征结合比句型等信息和小句级特征结合更自然，同时避免了句型等信息与小句级特征相结合带来的特征激增和小句级特征稀疏问题，使得模型训练和预测的时间都会减少，同时也降低了过多特征一起使用发生过拟合的可能性。

3.3 分类特征

如表1所示，本文将所有的特征分为两类：小句级特征和篇章级特征。在小句级，使用的是二值特征表示形式，即特征出现或不出现。在篇章级，使用特征频率的特征表示，即不同倾向的小句分别做累计，将累计结果作为篇章倾向分析特征。

情感词、否定词和程度副词作为小句级特征，这三类词汇都是情感倾向分析重要词汇，其中情感词是情感分析的基础，分为支持和反对两类。否定词会改变原本所表达的情感倾向，程度副词能改变所表达情感倾向的程度。

表 1　特征

特征	特征层级	例子
情感词	小句级	喜欢
程度副词	小句级	非常

模型进行分类，准确率提高到了 0.852，说明级联模型对比单一分类模型存在优势。当在上一个实验的基础上加上了句型信息和位置信息等特征时，准确率提高到了 0.874，比基准实验提高了 2.53%。

4.2 分析

本文方法准确率较基准实验有了比较大提高，作者认为主要有以下几点原因。

（1）细化结构。本文将篇章划分到小句级，并使用级联模型对小句级和篇章级进行情感倾向分析。使得结构更加细化，并且可以明确不同的特征所属小句范围以及小句的位置。这样对于篇章情感倾向分析起到了重要作用。

（2）两种分类模型互补。本文在不同的层级使用了不同的分类模型：最大熵和支持向量机。这两种分类模型原理不同，在不同的层级用于情感倾向分析，可能会对情感倾向分析结果产生一些互补，而使效果有所提高。

（3）级联模型能融入更多的句子特征。使用级联模型使得句子结构得以清晰刻画，因此句子的特征能够更方便地融合进来。例如，句子位置、句型信息、句子数量等，这些特征对实验结果的提高也起了很大作用。通过统计可以看出句型信息在实验语料中出现的次数并不多，如果语料中句型信息更丰富，可能会发挥更大作用。

5 结论与展望

本文提出单层标注级联模型对篇章情感倾向进行分析，将篇章情感倾向分析分为两级，即小句级和篇章级。通过实验结果可以看出，使用最大熵模型且情感词、否定词、程度副词作为特征的情感分类准确率为 0.849，本文方法为 0.874，提高了 2.53%。这表明了细化结构到小句级对篇章级情感倾向分析是非常必要的，同时使用单层标注级联模型进行情感分类以及句型、句子位置等句子特征是非常有效的。

在下一步的研究里，对单层标注级联模型与多层标注级联模型在情感分类中的分类效果，仍需要做更多对比研究，同时对于单层标注级联模型不同层次之间的影响做进一步的探讨。

另外王根、赵军[19]和刘康、赵军[20]利用 CRFs 模型在句子主客观和褒贬分析方面做了有益的工作，由于本文侧重利用小句分析提高篇章情感分析，未作对比，后续将结合 CRFs 模型和本文的方法进行对比研究。

致谢 首先感谢中国科学院计算技术研究所的谭松波博士提供的情感分析语料，给我们的实验带来很大帮助，同时感谢台湾大学林智仁先生公开了高效易用的支持向量机模型。还要向对本文工作给予支持的同学表示感谢，他们是吕新波、孙慧和薛璐影，以及实验室最大熵模型的编写者陈志杰、李赟和阎于闻。

参考文献

[1] Bo Pang, Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts [C]//Proceedings of the ACL, Barcelona, Spain: 2004, 201-208.

[2] Mike Thelwall, David Wilkinson, Sukhvinder Uppal. Data mining emotion in social network communication: Gender differences in MySpace [J]. Journal of the American Society for Information Science and Technology, 2010, 1(64): 190-199.

[3] 许洪波，姚天昉，黄萱菁. 第二届中文倾向性分析评测技术报告[C]//第二届中文倾向性分析评测. 上海：2009，1-23.

[4] 董喜双，关毅，李本阳. 基于最大熵模型的中文词与句情感分析研究[C]//第二届中文倾向性分析评测. 上海：2009，50-58.

[5] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques [C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language processing. 2002: 79-86.

[6] 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.

[7] 徐军，丁宇新，王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报，2007，21(6)：95-100.

[8] Shotaro Matsumoto, Hiroya Takamura, Manabu Okumura. Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees[C]//Proceedings of PAKDD. 2005: 301-311.

[9] 陈锦禾，范新，沈闻，等. 基于情感词识别的 BBS 情感分类研究[J]. 计算机技术与发展，2009，7(19)：120-123.

[10] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of

（下转第 20 页）

Improving MapReduce Performance in Heterogeneous Environments[C]//Proceedings of San Diego, CA: Proc. OSDI, 2008: 29-42.

[61] W Zhao, H Ma, Q He. Parallel K-Means Clustering Based on MapReduce[J]. Lecture Notes in Computer Science, 2009, 5931: 674-679.

[62] Bacchiani M, Beaufays F, Schalkwyk J, et al. Deploying GOOG-411: Early Lessons in Data, Measurement, and Testing[C]//Proceedings of Acoustics, Speech and Signal Processing, ICASSP 2008: 5260-5263.

(上接第8页)

reviews[C]//Proceedings of the ACL. 2002: 417-424.

[11] Ryan McDonald, Kerry Hannan, Tyler Neylon, et al. Structured Models for Fine-to-Coarse Sentiment Analysis[C]//Proceedings of the ACL. 2007: 432-439.

[12] **李钝,曹付元,曹元大,等. 基于短语模式的文本情感分类研究[J]. 计算机科学**, 2008, 4: 132-134.

[13] **邢福义. 小句中枢说[J]. 中国语文**,1995, 6.

[14] **黄忠廉. 小句中枢全译说[J]. 汉语学报**, 2005, 2.

[15] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms[C]//Proceedings of EMNLP. Philadelphia, PA: 2002: 1-8.

[16] T. Jaynes. Information Theory and Statistical Mechanics [J]. Physics Reviews. 1957, 106: 620-630.

[17] C Cortes, V Vapnik. Support vector networks [J]. Machine Learning, 1995, 20: 273-297.

[18] **阎威武,邵惠鹤. 支持向量机分类器在医疗诊断中的应用研究[J]. 计算机仿真**,2003, 20(2): 69-70.

[19] **王根,赵军. 基于多重冗余标记** CRFs **的句子情感分析研究[J]. 中文信息学报**,2007,21(5): 51-55,86.

[20] **刘康,赵军. 基于层叠** CRFs **模型的句子褒贬度分析研究[J]. 中文信息学报**,2008,22(1): 123-128.

[21] **王素格. 基于** Web **的评论文本情感分类问题研究[D]. 上海: 上海大学**,2008.

[22] **王国胜. 支持向量机的理论与算法研究[D]. 北京: 北京邮电大学**,2008.

[23] Jiang Wenbin, Huang Liang, Liu Qun, et al. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging[C]//Proceedings of the ACL. 2008: 897-904.