文章编号:1000-5862(2013)02-0120-05

中文情绪识别方法研究

刘欢欢[1],李寿山[1],周国栋[1*],李逸薇[2]
(1. 苏州大学计算机科学与技术学院,江苏 苏州 215006;
2. 香港理工大学中文及双语学系,香港 999077)

摘要:以中文情绪语料库(Ren-CECps)为基础,重点研究了句子级情绪识别方法. 比较了不同特征以及不同机器学习分类方法(NB,SVM,ME)对情绪识别的影响. 此外,针对情绪文本和非情绪文本在语料中的分布非常不平衡问题,通过集成学习的算法来实现不平衡情绪识别,用以提高情绪识别的整体性能. 实验结果表明:使用基于样本的集成学习方法能够有效解决不平衡问题,明显提高情绪识别的分类性能.

关键词:情绪识别;特征工程;分类方法;不平衡分类;集成学习
中图分类号:TP 391 **文献标志码**:A

0 引言

随着 Web 2.0 技术的高速发展,互联网已经成为社会各种信息的重要载体,成为人们生活中不可或缺的重要信息来源. 特别是近年来,随着博客、电子商务、社交网站以及微博的兴起,互联网给广大用户提供了丰富的发表自己观点的平台. 为了处理和分析这些海量的评论信息,情绪分析正渐渐发展成为自然语言处理中一项越来越受关注的研究课题. 情绪是指人内在的心理反应与感受,如喜、怒、哀、乐等. 在人类的活动中,人对事或物的态度(观点、看法)往往与人的情绪紧密相联,也就是说可以从人的情绪观察到人对事物的观点倾向. 而情绪具有潜在的领域、主题与时期独立性的特点. 借助于这些特点,可以从情绪的角度去研究文本分类问题.

一般而言,情绪分析有 2 项基本任务:情绪识别和情绪分类[1]. 前者通过对大粒度文本(如句子级、篇章级文本)进行分析,判断其是否含有情绪;后者是在情绪识别的基础上对含有情绪的文本进行分析,进行文本情绪具体类别的判别,如属于“高兴”类别还是属于“生气”类别等.

在情绪分析研究中,由于关键词比较少,而且在一般情况下情绪都是隐藏的,手工操作需要大量的人力和物力. 因此,通常使用机器学习的方法,即通过对少量的标注样本进行学习得到分类器,从而实现对大量未标注样本的分类. 然而现有分类器的设计是基于类分布大致平衡这一假设的,即通常假定训练数据集中各类所含的样本数大致相当. 但是,通过分析情感识别语料发现,这一假设并不成立. 数据集中某个类别的样本数可能会远远少于其它类别. 例如,在新闻类文本中,不含情绪的文本明显多于含有情绪的文本,而在微博语料中,含有情绪的文本通常远远多于不含情绪的文本. 当传统的机器学习方法用于解决这类不平衡分类问题时,往往会出现分类器性能的大幅度下降,得到的分类器具有较大的偏向性,最常见的表现为少量样本类的识别率远远低于大量样本类[2-4]. 因此,情绪分析中对不平衡数据集的分类问题的研究需要寻求新的分类方法.

本文重点研究中文情绪的识别方法,并通过样本的集成学习方法解决不平衡问题,用以提高情绪识别方法的性能. 实验结果表明:不平衡问题的解决能够有效提高情绪识别的分类性能.

1 相关工作

随着网络上具有主观性评价的文本不断增多,文本情绪分析渐渐成为自然语言处理领域中的一个研究热点. B. Pang 等[5]首次将机器学习方法(贝叶斯、最大熵、支持向量机)用于情感分析,使用关键词作为特征来识别电影评论中的积极和消极情感. A. Neviarouskaya 等[6]研究了文本通信中的情感识

收稿日期:2012-11-15
基金项目:国家自然科学基金(61003155,60873150)和模式识别国家重点实验室开发课题基金资助项目.
通信作者:周国栋(1967-),男,江苏常州人,教授,博士,主要从事自然语言理解,机器翻译,信息提取,信息检索和机器学习等方面的研究.

别和分析. 先前的研究大多是基于篇章级的情绪分析,这在很多应用方面存在不足. 通过分析句子间的情绪关联,可以提高基于文本的情绪分析性能. 因此,基于句子级的情绪分析就显得十分重要. S. Aman 等[7]通过一种基于知识的方法实现句子级的情绪识别. Quan Changqin 等研究了基于情绪词的句子级情绪分析,使用的语料是中文情绪语料库(Ren-CECps)[8],文中实现了通过使用情绪词对句子的情绪进行识别,F 值达到了 65.0%.

本文研究基于句子级的中文情绪识别方法研究,这是情绪分析的基础性工作. 实验中提取了 3 种特征,通过分别构建空间向量进行情绪识别,来比较这 3 种特征的性能,还比较了 3 种分类方法的分类效果. 考虑到样本的不平衡对分类结果的影响,采取基于样本集成学习的策略来解决样本的不平衡问题,取得的分类效果优于 Quan Changqin 等的结果.

2　中文情绪识别方法

2.1　特征提取

为了便于处理文档中的信息,首先要对文本进行科学的抽象,建立数学模型,将文档表示成计算机能够处理的形式,用以描述和代替文本. 文本表示首先要确定的问题是如何表示文本的基本单位,即文本的特征或特征项. 特征项[9]必须具备一定的特性:(ⅰ)特征项能够标识文本内容;(ⅱ)特征项分离要比较容易实现;(ⅲ)特征项具有将目标文本与其他文本相区分的能力.

本文使用的语料是中文情绪语料(Ren-CECps). 对于特征,提取了词、词 + 词性、词 + 词 3 种特征,由于语料中已经实现了文本分词并标注了词性,可以方便地提取出这 3 种特征,并统计出词频. 在此基础上,分别构建了 3 种空间向量模型对文本进行情绪识别,例如:

世界/n　因/c　你/r　而/c　精彩/a　,/w
中国/ns　因/c　你/r　而/c　震撼/v　!

此例句是含有情绪的,提取的词特征有: 世界、因、你、而、精彩、,、中国、震撼; 词 + 词性特征是在词特征的基础上添加以下特征: 世界_n、因_c、你_r、而_c、精彩_a、,_w、中国_ns、震撼_v; 词 + 词特征为在词特征的基础上添加以下特征: 世界_因、因_你、你_而、而_精彩、精彩_,、,_中国、中国_因、而_震撼.

2.2　分类方法简介

在后续实验中所使用的相关机器学习方法,分别为朴素贝叶斯(NB)分类方法、最大熵(ME)分类方法和支持向量机(SVM)分类方法.

2.2.1　朴素贝叶斯分类方法

朴素贝叶斯是一种概率统计学方法,其基本思想是: 对于给定的待分类文档,利用特征项和分类的联合概率来估计文档的分类概率. 该方法有一个假设前提,在给定的文档中,文档的特征项是相互独立的. 朴素贝叶斯方法一般采用 DF 向量表示法表示文档,文档向量的每一个分量都是一个布尔值,用 1 表示相对应的词在该文档中出现,0 表示相对应的词未在该文档中出现. 在此方法中,对于一个给定的文档 d,它属于 c 类的概率为

$$P(c\mid d)=\frac{P(c)\prod_{t\in v}P(d(t)\mid c)}{\sum_{c}P(c)\prod_{t\in v}P(d(t)\mid c)},$$

其中

$$P(d(t)\mid c)=\frac{1+N(d(t)\mid c)}{2+\mid dc\mid},$$

$P(c)$为 c 类文档的概率,$P(d(t)\mid c)$是对 c 类文档中特征 t 出现的条件概率的拉普拉斯估计,$N(d(t)\mid c)$为在 c 类文档中 t 出现的文档数,$c\in\{-1,1\}$,1 表示情绪文本,-1 表示非情绪文本.

2.2.2　最大熵分类方法

最大熵分类方法是基于最大熵信息理论,其基本思想是为所有已知的因素建立模型,而把所有未知的因素排除在外. 也就是说,要找到一种概率分布,满足所有已知的事实,但是让未知的因素最随机化[10]. 在最大熵模型下,预测条件概率 $P(c\mid d)$的计算公式为

$$P(c\mid d)=\frac{1}{Z(d)}\exp\left(\sum_{i}\lambda_{i,c}F_{i,c}(d,c)\right),$$

其中 $Z(d)$为归一化因子,$F_{i,c}$是特征函数且

$$F_{i,c}(d,c')=\begin{cases}1, & n_i(d)>0 \quad \text{且 } c'=c,\\ 0, & \text{其他.}\end{cases}$$

2.2.3　支持向量机分类方法

支持向量机[11]是近些年来在统计学习理论的基础上发展起来的新的模式识别方法,它在解决小样本、非线性以及高维模式识别问题中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中. 最初是作为 2 类分类问题提出来的,其基本思想是在向量空间中找到一个决策平面(Decision surface),这个平面能"最好"地分割 2 个分类中的数据点[12],分类边界值是从决策平面分别向 2 个类的点平移,直到遇到第 1 个数据点. 2 个类的分类边界的距离就是分类间隔,如图 1 所示

在图 1 中,正方形和圆圈分别代表正负两类样本,H 为分类线,H_1,H_2 代表两类样本中距离分类线最近且平行于分类线 H 的样本,将它们之间的距离成为分类间隔(*Margin*). 能将两类正确分开且 *Margin* 最大的分类线为最优分类线. 可以看到,当分类间隔

为 $1/\|w\|$ 时刚好能最好地分割两边的数据点.

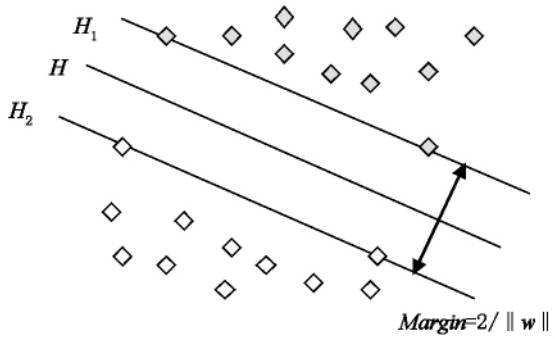

图 1 SVM 基本原理图

2.3 基于样本集成学习的不平衡分类方法

由于情绪语料中情绪和非情绪分布不平衡,使分类器有较大的偏向性. 为了充分利用多类样本,使用一种集成学习的方法[13]解决情绪识别中的不平衡问题. 基于样本的集成学习策略是在多类样本集合中进行多次欠采样和少类样本构建多个训练集合,并在每个训练集合上训练一个基分类器,最终融合每个基分类器的结果,从而达到充分利用多类样本的效果.

针对不平衡分类问题,对于少类样本集合 S_{MI} 和多类样本集合 S_{MA},通过随机欠采样在多类样本集合中选择出一个子集 S'_{MA},并且 $|S'_{MA}| = |S_{MI}|$. 多个多类样本的子集 $S'_{MA_1}, S'_{MA_2}, \cdots, S'_{MA_n}$ 独立地从多类样本集合 S_{MA} 中获得,并且保证 $|S'_{MA_i}| = |S_{MI}|$. 通过对每个多类样本的子集 S'_{MA_i} 和少类样本集 S_{MI} 构建基分类器 C_i,最终多个基分类器通过集成学习的方式融合结果. 具体步骤为:(ⅰ)通过随机欠采样在多类样本集合 S_{MA} 中选择多个样本子集 S'_{MA_i},而且 $|S'_{MA_i}| = |S_{MI}|$;(ⅱ)基于每个多类样本的子集 S'_{MA_i} 和少类样本集 S_{MI} 构建基分类器 C_i;(ⅲ)对每个基分类器 C_i 的分类结果进行融合. 整个集成学习的流程如图 2 所示.

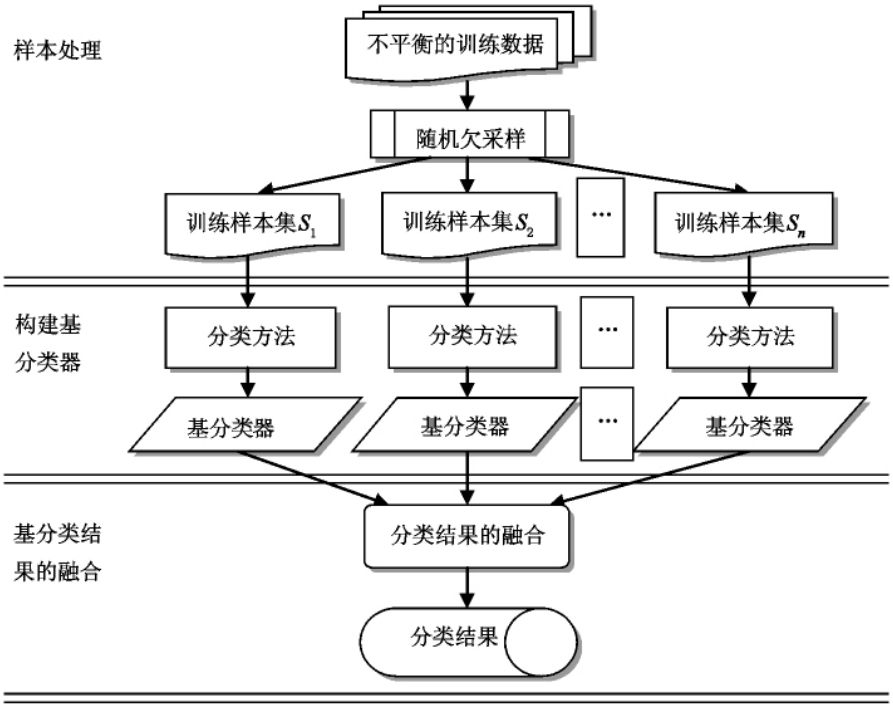

图 2 集成学习的流程图

本文使用 3 种分类方法(朴素贝叶斯、最大熵和支持向量机)分别实现集成学习方法,在对分类结果的融合中,使用加法规则对各个基分类的结果进行融合.

3 实验设计与分析

3.1 实验设置

实验使用的语料是中文情绪语料(Ren-CECps). 该语料为 Ren-lab 实验室基于一个相对细粒度标注规则而创建的语料库,以 XML 文档形式组织,包括句子的分词标签和属性标签. 语料的标注分别从 3 个层次标注文本情绪:文档级、段落级、句子级. 该语料库主要来源于新浪微博、百度微博、腾讯微博等,包含了 1 487 个博客文章,35 096 个句子. 句子级是情绪标注的基本层次,也是文本情绪分析的基础. 语料中有 8 种基本的情绪:高兴(Joy)、喜爱(Love)、厌恶(Hate)、悲伤(Sorrow)、焦虑(Anxie-

ty)、惊奇(Surprise)、生气(Angry)、期待(Expect)，每句均按照句中所表达的情绪对这 8 种基本情绪进行赋值，值越大表达的情绪越强烈.

在获取情绪和非情绪两类文本时，从文档中提取出句子级的语料，当 8 种基本情绪的值不均为 0 时，将其视为有情绪; 当 8 种基本情绪的值均为 0 时，将其视为无情绪.

实验中使用了 3 种分类方法: 朴素贝叶斯分类方法、基于 Mallet 工具包的最大熵分类方法和支持向量机分类方法. 针对衡量识别的性能，本文采用准确率(*Acc*)作为分类效果的衡量标准，准确率的计算公式为

$$Acc = \frac{\text{Number of correctly classified samples}}{\text{Total number of all samples}}.$$

3.2　实验结果分析

3.2.1　特征及分类方法的比较研究　在本实验中，从语料中分别选取正负样本 1 000 句、1 500 句、2 000 句作为训练数据，选取 400 句作为测试数据. 通过基于词(Unigram)、词 + 词性(Unigram_POS)、词 + 词(Bigram) 3 种特征，分别采用朴素贝叶斯、最大熵以及支持向量机分类方法进行情绪识别.

图 3 ~ 图 5 为分别使用贝叶斯、最大熵、支持向量机分类方法对 3 种特征在不同数量的训练数据中所得到的分类结果. 从图 3 ~ 图 5 中可以看出随着训练数据中正负样本量的增加，分类结果呈上升趋势，而且基于 3 种特征在最大熵和支持向量机分类方法中得到的分类结果相差不大，并且这 3 种分类方法在同一种特征下的分类结果基本相近. 这说明这 3 种特征的选取对情绪识别的效果影响不明显，而且 3 种分类方法的分类性能相差不是很大.

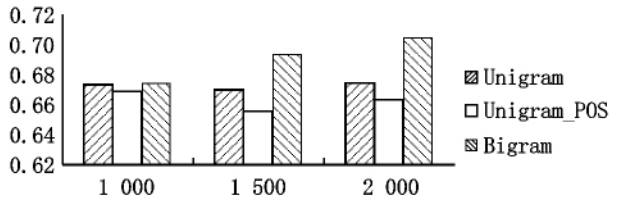

图 3　使用 NB 分类方法得到的分类结果比较

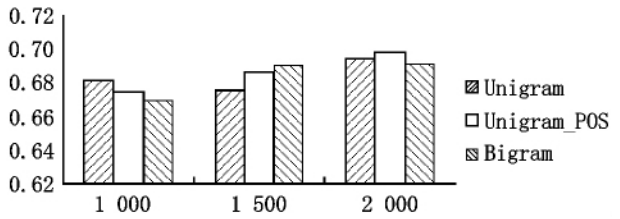

图 4　使用 ME 分类方法得到的分类结果比较

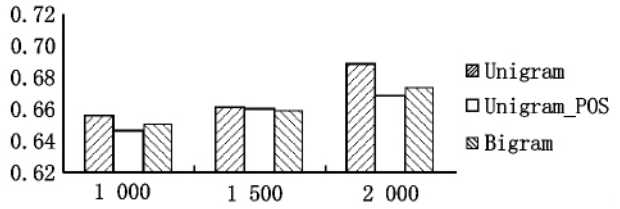

图 5　使用 SVM 分类方法得到的分类结果比较

3.2.2　利用样本集成学习的方法处理不平衡问题　为方便起见，下面的实验中只使用词特性. 接下来在训练数据中采用完全训练方法(FullT)进行实验，测试数据保持不变. 最后利用集成学习的方法(Ensemble)来处理样本，其中不平衡比为训练样本中含有情绪的句子数量与不含情绪的句子数量之比. 本文使用的语料中不平衡比约为 12∶1. 训练数据中不含情绪的句子仍选取 2 000 句，而含有情绪的样本按照不平衡比在标注样本中选取，测试数据同上面的实验设置. 在 3.2.1 节中特征与分类方法的比较研究中，为了更清楚地描述实验结果，将针对使用词特征和在训练数据中正负样本均为 2 000 句的一组实验称为欠采样方法(UnderS).

图 6 为在测试数据中选择不同数量的正负样本，并通过基于样本的集成学习方法得到不同的分类结果. 从图 6 中数据可以看出，采用完全训练方法(FullT)，即当测试数据中正负样本数量不相等时，得到的分类结果明显下降，而且对朴素贝叶斯分类方法和支持向量机分类方法的分类性能影响非常明显，说明这 2 种分类方法对正负样本平衡与否比较敏感. 而最大熵分类方法的分类效果下降幅度比较小，这可能和分类方法所使用的分类原理不同有关. 从图 6 还可以看出，样本集成学习方法表现明显优于其他两种方法(FullT 和 UnderS)，提高幅度都超过了 5 个百分点. 此结果说明基于样本的集成学习方法能够很好地处理情绪识别任务中的数据不平衡问题. 由于语料标注本身比较困难，同时又有很大的主观性，因此在对文本的标注过程中会存在不合理的地方，这些都可能对分类结果造成影响.

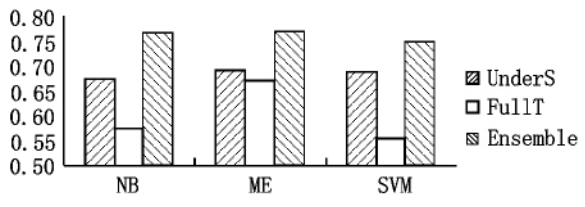

图 6　基于样本集成学习的分类效果的比较

4 总结与展望

本文主要研究中文情绪识别方法,通过对词、词+词性、词+词3种特征使用不同的分类方法分别进行实验,实验结果表明这3种特征的选取对情绪识别的效果影响不是很大.同时,这3种分类方法在同一种特征下都取得了较好的分类结果,它们的分类效果相差不大.由于语料中含有情绪的样本数量远远多于不含情绪的样本数量,无法充分利用多类的标注样本.因此,本文提出了基于样本的集成学习方法,充分使用了标注样本,解决了样本的不平衡问题.实验结果表明该方法能够明显提高情绪识别的效果.在下一步工作中,将在情绪识别的基础上,进行情绪分类的相关工作.

5 参考文献

[1] Aman S,Szpakowiczm S. Identifying expressions of emotion in text [EB/OL]. [2012-04-12]. http://nlp.ipipan.waw.pl/NLP-SEMINAR/071029b.pdf.

[2] Castillo M,Serrano I. A muhistrategy approach for digital text categorization from imbalanced documents [EB/OL]. [2012-03-19]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.6892&rep=rep1&type=pdf.

[3] Murphey Y,Wang H,Ou G et al. An effective algorithm for multi-class learning from imbalance data [EB/OL]. [2012-03-23]. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04370991.

[4] Neviarouskaya A,Prendinger H,Ishizuka M. Textual affect sensing for sociable and expressive online communication [EB/OL]. [2012-04-19]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.6182&rep=rep1&type=pdf.

[5] Pang B,Lee L,Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [EB/OL]. [2012-04-22]. http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

[6] Quan Changqin,Ren Fuji. Construction of a blog emotion corpus for chinese emotional expression analysis [EB/OL]. [2012-06-17]. http://www.aclweb.org/anthology/D09-1150.

[7] Quan Changqin,Ren Fuji. Sentence emotion analysis and recognition based on emotion words using Ren-CECps [J]. International Journal of Advanced Intelligence,2010,2(1):105-117.

[8] Shahshahani B,Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon [J]. Journals & Magazines,1994,32(5):1087-1095.

[9] Zheng Zhaohui,Wu Xiaoyun,Srihari R. Feature selection for text categorization on imbalanced data [J]. SIGKDD Explorations,2004,6(1):80-89.

[10] 胡燕,吴虎子,钟珞. 中文文本分类中基于词性的特征提取方法研究 [J]. 武汉理工大学学报,2007,29(4):132-135.

[11] 刘挺,车万翔,李生. 基于最大熵分类器的语义角色标注 [J]. 软件学报,2007,18(3):565-573.

[12] 王中卿,李寿山,朱巧明,等. 基于不平衡数据的中文情感分类 [J]. 中文信息学报,2012(3):33-37,64.

[13] 叶志飞,文益民,吕宝粮. 不平衡分类问题研究综述 [J]. 智能系统学报,2009,4(2):148-156.

A Study on Chinese Emotion Recognition Method

LIU Huan-huan[1],LI Shou-shan[1],ZHOU Guo-dong[1*],LI Yi-wei[2]

(1. School of Computer Science and Technology,Soochow University,Suzhou Jiangsu 215006,China;
2. Department of Chinese & Bilingual Studies,The Hong Kong Polytechnic University,Hongkong 999077,China)

Abstract: The emotion recognition method at the sentence level is studied with a Chinese emotion corpus(Ren-CECps). Specifically has been investigated,the impact of different linguistic features as well as different classification methods(NB,SVM,ME) on the emotion recognition and classification has been compared. Moreover,they has proposed an ensemble learning approach to tackle the problem imbalanced data distribution of the emotion and non-emotion text. Experimental results have shown that the approach effectively enhances the performance of emotion recognition when the data distribution has been imbalanced.

Key words: emotion recognition; feature engineering; classification method; imbalanced classification; ensemble learning

(责任编辑:冉小晓)