一种基于特征选择的主观性文本分析方法①

田卫新[1]，郑 胜[2]

[1](三峡大学 计算机与信息学院，宜昌 443002)
[2](三峡大学 理学院，宜昌 443002)

摘 要：提出了一种主观性文本分析方法。方法采用多种不同策略表示文本，使用特征选择算法消除不相关特征及冗余特征后，训练 SVM 对文本按主观性和客观性进行分类。采用的特征选择算法以 Simba 为基础，通过实验对其迭代和相似度计算方法进行了改进，克服了在实际应用中出现的不稳定性问题。分别在中英文语料上进行了实验，结果表明该方法在实验语料上的性能优于已有方法。

关键词：观点挖掘；情感分析；主观性分析；特征选择

Approach to Analyzing Subjective Text Based on Feature Selection Algorithm

TIAN Wei-Xin[1], ZHENG Sheng[2]

[1](College of Computer and Information Technology, Three Gorges University, Yichang 443000, China)
[2](College of Science, Three Gorges University, Yichang 443000, China)

Abstract: This paper proposed a method to analyzing subjective text. The method uses various strategies to stand for text with feature vectors, and uses SVM to classify text according to the property of subjectivity and objectivity after eliminating the rundant and irrelevant features using feature selection algorithm. The feature selection algorithm in the paper bases on SIMBA. We improve the original SIMBA on the way of iteration and the measure of similarity through experiment, and overcome the instability when putting into application. In the experiment done on English and Chinese corpus respectively, the accuracy overperforms that by SVM algorithm alone and the F-MEASURE is better than that by the baseline method on same corpus.

Key words: opinion mining; sentiment analysis; subjectivity analysis; feature selection

1 引言

随着互联网应用技术的不断发展，各类信息以前所未有的速度在互联网上积累和传播，极大地促进了社会的发展，同时给信息处理能力提出了新的挑战。主观性文本信息是其中的一类重要信息，广泛出现在博客、社区以及电子商务等网站上，表达作者的观点、偏好、倾向等具有较强感情色彩的主观意愿，在产品使用反馈、民意调查、消费心理等方面具有很大的应用价值。在这种背景下，意见挖掘和情感分析（Opinion Mining and Sentiment Analysis）等技术以处理主观性文本，实现文本中各种对象的提取、倾向性分析等的自动处理功能，成为当前国内外的研究热点之一。

主观性分析技术旨在将主观性文本和客观性文本分离开来，根据分析的粒度可以分为篇章、段落以及句子等级别。将主观性文本抽取出来有助于提高人工分析的效率，同时也有利于提高自动分析的准确率。影响文本主观性和客观性的因素比较复杂。首先，判断标准不统一，在文本中，主观性内容和客观性内容往往同时出现，或者混合在一起很难分开，使得判断标准带有主观性；其次，主观性和客观性是语义层概念，为了准确区分，应该先分析句子语义，然后在语义分析的基础上再做主客观判断。由于一则语义分析技术尚不成熟，二则主客观分析和语义分析技术的应用背景不同，因此目前只能以词法、句法、语法或浅层语义等层面的特征来反映其语义上的区别；最后，文本的主客观属性具有局部性，意见或情感是和对象

① 收稿时间:2010-12-13;收到修改稿时间:2011-03-03

密切联系的，文本中的意见或情感的关联对象经常发生改变，因此需要预先确定范围并在限定的范围内进行分析。

本文采用基于数据驱动的方法。首先，尽可能多地收集影响文本主客观性质的因素，并以向量方式构造成特征；然后使用特征选择算法消除特征中的冗余或不相关特征;最后使用得到的特征训练 SVM 分类器并对文本按照主客观属性进行分类。在英汉语料上的实验表明该方法是有效的。

本文后面部分内容安排如下：第 2 节介绍和本文相关的研究工作；第 3 节提出主观性分析问题的数据驱动的解决方法及采用特征；第 4 节介绍特征选择算法及其改进；第 5 节介绍实验过程及结果分析；第 6 节为结论及后续研究展望。

2 相关研究

主观性文本分析在上世纪九十年代后期开始随着 Internet 的普及逐渐受到了研究者的重视。目前在国际上知名的文本和自然语言处理方面的会议均设置了主观性文本分析的议题和评测：如针对英语文本分析的有 TREC、EMNLP 等；针对亚洲语言和中文文本方面的有日本国家科技信息中心（NACSIS）举办的 NTCIR 和中国中文信息学会信息检索专业委员会举办的 COAE 等。当前对主客观分析技术的研究可以粗略地分为二个方面，一是从语言学角度出发，对主观性和客观性的区别进行深层地定义；二是以具体应用为背景，在分类方法上开展研究。前一方向的研究又可分为概念、线索和模型等三个内容。在概念方面，提出私人状态和其它一些可用于情感计算方面的基本概念[1]；在线索方面，建立了形容词和主客观属性之间的正相关关系;研究了具有语义或分级形容词的词法语义特性对主客观分析的影响[2,3]。另外一些研究使用信息抽取模板表示主观表达式，使用经过标注的句子训练抽取模板学习算法，然后通过算法得到模板[4]；使用语言信息量大的名词表示主观性文本[5]。在模型方面，文献[1]提出了标注方法根据句子中形容词的出现情况决定其主客观性；文献[6]提出了在文本中根据其它特征出现的密度区分主观性的方法。在后一方面，根据不同的分析级别采用了不同的方法。在文档级别中，文献[7]利用 Naive Bayes 等分类器，文献[8]利用数据自身的属性或人工标注的方法获得训练数据后进行分析；句子级别的主客观分析是其中实用性最大的。除了采用上面文档级别分析方法以外，其它方法如使用 WordNet、采用模糊集以及其它新型的分类器等被采用。如文献[9]使用 Mini-Cut 分类器利用成对交互信息分类数据；Kim 等首先收集一些情感词和非情感词的集合作为主观性线索，然后定义二种模型组合判断句子的主客观性。文献[10]在分类前使用模糊集表示属性。通常在表达观点、倾向时，都需要使用主观词，所以主观词通常作为主客观性判断的线索。主观词一般是形容词，也有一些副词或表达感情色彩的名词或短语。获取主观词的方法概括起来有使用手工标注、使用词典或使用自举学习方法。

本文的工作与文献[1]较为接近，区别在于本文中在训练和使用分类器前采用了改进的特征选择算法消除评价向量中存在的冗余和不相关特征。

3 方法和特征

用数据驱动方法解文本主客观性分析问题，首先准备与待处理文本集合独立同分布的训练数据集，然后设计特征表示策略对待处理文本集合和训练数据集进行向量化，选择分类算法在训练数据向量集上学习一个分类器，分类待处理文本向量集。最后由向量集的类别标识得到待处理文本的主客观性类别。该问题的形式化表述如下：

设 $S_u=\{s_1,s_2,...s_n\}$ 为待分析的句子集合，目标集合 $Y=\{-1,1\}$ 表示句子所属类别，为客观性或主观性。S_l 为主观性文档集合，与 S_u 独立且具有相同分布函数，已知 $Tr=\{<s_i,y_i>\},1\le i\le m$， $s_i\in S_l$。

求 $z:S_u\to Y$。

为了求 Z，定义文档特征集 $F=\{f_1,f_2...f_l\}$，设 H 为由 F 张成的 1 维特征向量空间，$g:S_u+S_l\to H, v:H\to Y$ 分别为 $S(S_u+S_l)$ 到 H，H 到 Y 的映射。

则 $z=v\bullet g(S_u)$。

在求解特征向量集合到目标集合之间的映射时，可转换为求解如下的最优化问题：

$$\begin{aligned}&\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_jK(h_i,h_j)-\sum_{j=1}^{m}\alpha_j\\&s.t.\quad \sum_{i=1}^{m}y_i\alpha_i=0,\\&\qquad 0\le\alpha_i\le C,i=1,...m.\end{aligned}\tag{1}$$

求解得到最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,...\alpha_m^*)^T$。

选择α^*的一个小于C的正分量α_j^*，计算$b^* = y_j - \sum_{i=1}^{m} y_i \alpha_i^* K(h_i, h_j)$。

则决策函数为：

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{m} y_i \alpha_i^* K(x, h_i) + b^*) \quad (2)$$

使用 SVM 算法求解，根据文本分类问题的研究经验，在算法中选择线性核函数。

从句子到特征集之间的映射，可以选择字序列、词序列、语言模型等多种特征表示方案，各种表示方案性能对领域和文本特点较为敏感。在假定没有任何领域相关和文本类型的信息，我们采用了尽量多的特征表示策略，在后面这些特征将会被综合起来，下面是采用的特征。

词。词是文本分析中常采用的特征。在英文文本中获取词很容易；在中文文本中，首先对文本进行分词，然后再统计其中的词。我们采用词现和词频二种特征。

短语。短语是由二个或以上词组成的固定搭配。短语中的词可能相邻也有可能被文本中的其它词分开。由于短语特征的稀疏性，采用其出现与否作为特征。获取短语特征的方法是预先定义短语表，根据短语表来判断短语在句子中出现的情况。

句法。句法是指语态、句型、时态等语法层面的语言特征。我们利用句子的形态和特定词的位置来定义句法特征。

情感词。情感词是判断句子主客观性的重要特征，我们通过查询预先定义情感词典来确定句子包含情感词的情况。应用中忽略了情感词的级别，并且所有的情感词被看成同一个词，采用情感词出现的频率作为特征。

4 特征选择算法

特征选择是指从给定的特征集中抽选一个小的子集，使得该子集能完整表达特征全集的对数据类别的预测特性。由于文本处理领域数据集的维数通常很高，造成了数据稀疏和运算量大等问题，因此采用特征选择算法对数据进行处理通常是十分必要的。

特征选择算法大致可分为包装模型和过滤模型二种。包装模型和分类器结合起来，直接优化分类器的性能，在算法的每一轮估算抽选的子集对分类器的预测效果；过滤模型采用预处理方式，定义特定的评价函数，通过搜索特征全集中的子集使得评价函数达到最大值。

在单纯使用的 SVM 算法时，通常做法是对文本数据建立特征向量表示，训练和分类时，不同类型特征分量所起的作用是相同的。这样避免了对语言经验知识的依赖，但增加了冗余或不相关特征，加大了算法的计算量，同时降低了算法的准确率。为了过滤特征向量中冗余特征和不相关特征，我们通过对 Simba 算法进行改进后用于对特征向量进行选择，去掉其中的冗余和不相关特征。

Simba 是一种基于边界度量的过滤型特征选择算法，在以往实验中表现出较好的性能[11]。原始的 Simba 算法在用于文本倾向性分析时存在两个问题：一是原始的 Simba 算法采用固定的迭代次数，需要根据以往经验进行选择，在应用时通常很难选择合适的迭代次数。迭代次数直接影响最终输出的特征权值向量 w，迭代次数太小，则 w 取不到极值点；迭代次数过大，则算法的效率会降低。二是对于实例的选择，原始的 Simba 算法采用随机选择的方法，不能避免选择的实例之间相似度高的问题，当选择的实例之间的相似度高则特征权值向量 w 会陷入局部极值点，使得最终选择的特征向量不能很好的表示目标文本。

针对上面的两个问题，以下对 Simba 算法作如下改进：一是使用本轮输出的 w 和上轮输出 w 之间的差值控制迭代次数，当两者差值小于某一固定值时，终止迭代过程；二是在随机选择实例以后，通过判断该实例和已有实例之间的相似度，排除相似度过高的实例，从而避免输出陷入局部极值的 w。改进后的 Simba 算法在算法 1 中列出。

方法首先按照特征集的维数定义每一维的权值向量，并将其初始化为单位向量，表示各维具有相同的重要性。定义基于边界的度量函数，对训练集中的实例计算该函数值，用来对 w 进行迭代，当该函数达到最大时的权值向量每一维的值即反映了对应特征的重要性，按照阈值即可选取所求的特征子集。

边界度量选择。在机器学习算法中，边界对算法的设计和评价有重要意义，一个好的学习算法必须保证分类的结果中不同类别之间的边界距离最大。本文采用的特征选择算法也将边界距离作为判断标准。边界度量有二种方式，一种是样本边界，通过由分类器划分的样本在空间上的距离计算；另一种是假设边界，

通过实例与其最近的不同类别的实例之间距离计算。算法采用第二种方法，计算公式为：

$$\theta_P^w = \frac{1}{2}\left(\|x - nearmiss(x)\|_w - \|x - nearhit(x)\|_w\right)$$

其中 $\|z\|_w = \sqrt{\sum_i w_i^2 z_i^2}$ (3)

度量函数。通过对训练集中每一个样例求其假设边界，并将这些值累加，得到一个特征权值向量的度量值，具体算式为：

$$e(w) = \sum_{x \in S} \theta_{s \setminus x}^w (x) \quad (4)$$

计算过程。由于 $e(w)$ 是平滑的，通过计算梯度迭代来使 $e(w)$ 最大化。计算公式为：

$$(\nabla e(w))_i = \frac{\partial e(w)}{\partial w_i} = \sum_{x \in S} \frac{\partial e(w)}{\partial w_i}$$

$$= \frac{1}{2}\left(\frac{\left(x_i - nearmiss(x)_i\right)^2}{\|x - nearmiss(x)\|_w} - \frac{\left(x_i - nearhit(x)_i\right)^2}{\|x - nearhit(x)\|_w} \right) w_i \quad (5)$$

每一轮计算 $(\nabla e(w))_i$，归一化并更新 w，直到梯度值满足指定阈值为止。

I-Simba 算法

1.初始化 l 维行向量 $w = (1,1,...1), \Delta = w$，初始化文档集合 $S_{sel} = Empty$ 0

2. while $\max|\Delta_i| > \varepsilon$ *do* 3,4

3. 从 Tr 中随机选择文档

4. If $Sim(S_{sel}, s_i) < 0$ *then* 5,6,7

5. $S_{sel} = S_{sel} \cup \{s_i\}$

6.寻找 $g(s_i)$ 与 Tr 中剩余文档特征向量 $g(Tr - s_i)$ 的同类最近邻 $g(s_i)$ 与异类最近邻 nearmiss($g(s_i)$)。

7. *for* $i = 1$ *to* l *do*

$w = w + \Delta$ *△按照公式（5）计算

8. $w \leftarrow w / \|w^2\|_\infty$

其中 $Sim(S_{sel}, s_i)$ 比较 s_i 和 S_{sel} 中的每一个文档 s，当全部相似度低于特定阈值值则返回-1，否则返回 1。两篇文档之间的相似度按照文档中字的出现与否定义向量空间模型后，通过计算向量之间的内积得到，通过预先对若干篇已知相似度的文档进行比较实验后确定阈值。

5 实验结果

我们在 TREC 数据集和 MP3 评价语料上对本文提出的方法分别进行测试。TREC 数据集是一种用于评价意见识别任务的英文语料，该任务给定一个 TREC 主题和一个按照对该主题相关性排好序的文档集，查找所有的具有观点的句子。数据集中共有 22 个主题包含 21115 个句子。我们首先通过词线索选择和主题相关的句子， MP3 评价语料是我们在产品评价检索任务中收集的，该语料包含针对蓝魔、纽曼、OPPO 以及 IPOD 等四种不同品牌的 MP3 的介绍、评论文章 1426 篇。我们从该语料中选择 400 篇文档，共 5638 个句子，按照 TREC 数据集的方法手工标注了每个句子的主客观属性。最后我们将二种语料中的句子分别划为 2 组和 4 轮用于训练和测试。

特征选择。在 Matlab 上实现了改进的特征选择算法，得到针对原始特征集的权值向量。特征选取的阈值设定为 0.01。从原始的特征向量中滤除权值小于 0.01 的特征后得到新的评价表示向量，并以此作为训练和分类的依据。

训练和测试。使用 SVMLight 来完成 SVM 训练和测试过程。将特征向量按照规定格式的数据文件整理后，生成模型文件，然后由模型文件和测试数据生成最终的类别预测文件。

实验结果及分析。我们在二个数据集上分别进行了二类实验测试本文方法。一类实验是在数据集上使用 I-SIMBA 方法与不使用特征选择选择算法得到的准确率进行比较，不使用特征选择算法时，单独用词、短语、句法以及情感词作为特征，使用 SVM 进行分类，结果分别记为（W-SVM, P-SVM, SYN-SVM, SEN-SVM），表 1 列出了在二种数据集上 I-SIMBA 方法与不使用特征选择方法得到准确率的对比；第二类是在相同数据集上，将 I-SIMBA 方法同基准方法得到的 F-Measure 值进行比较，采用的基准方法是由 Kim 提出的[12]。表 2 列出了在不同数据集上二种方法得到的结果。表格中的数据是 4 轮结果的平均值

表 1　不同数据集上方法的准确率对比

SCHEMA	TREC-ACCURACY	MP3VAL-ACCURACY
W-SVM	0.7824	0.7954
P-SVM	0.7511	0.7424
SYN-SVM	0.7213	0.7141
SENT-SVM	0.7983	0.8013
ISIMBA-SVM	0.8137	0.8321

表 2 不同数据集上的 F-MEASURE 值对比

METHOD	TREC F-MEASURE	MP3VAL F-MEASURE
I-SIMBA	0.532	0.561
BASELINE	0.514	0.502

从表 1 可以看出，五组方案的准确率均超过随机选择的结果，且在应用了特征选择的算法后的结果均优于相同数据下单独使用 SVM 算法。其中基于词序列的特征和基于情感词特征的准确率比短语和语法特征表示的高，而应用特征选择后的方法得到的准确率要高于同类数据的其他特征表示方案。表 2 显示了 I-SIMBA 方法分别在 TREC 数据集和 MP3VAL 数据集上与基准方法得到结果的比较情况，在二种数据集上，ISIMBA 方法的 F-MEASURE 值均高于基准方法。反映了通过对候选特征进行筛选，以及对多种特征综合后，更能准确的表示类别信息。表 1，表 2 中 I-SIMBA 方法在汉语语料上的结果好于在英文语料中得到的结果，基准方法和使用其它特征表示的方法在语言上的差异不明显。除了偶然因素外，语言的结构特征上的区别可能是原因之一。

6 结论及展望

主观性文本包含大量表明观点、态度、评价等信息，在经济、社会领域具有较大意义。然而和基于事实的文本不同，主观性文本中评价对象、观点通常和字词之间没有明确的对应关系，增加了分析难度。本文采用数据驱动的方法，在以 SVM 作为基本分类器的基础上，使用改进的 Simba 特征选择算法应用于文本主观性分析中消除冗余和不相关特征，在与单独特征表示的方法以及基准方法的比较实验中，取得了较好的结果。在后续的研究中，将进一步分析算法选择出来的特征同情感词之间的匹配情况，采用自学习算法或附加启发式信息提高方法的准确率。

参考文献

1 Wiebe JM, Wilson T, Bruce R, Bell M, Martin M. Learning subjective language. Computational Linguistics, September 2004,30(3):277-308.

2 Bruce R, Wiebe J. Recognizing subjectivity: A case study of manual tagging. Natural Language Engineering, 2000, 6(2).

3 Hatzivassiloglou V, Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity. Proc. of the International Conference on Computational Linguistics (COLING). 2000.

4 Riloff E, Wiebe J. Learning extraction patterns for subjective expressions. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2003.

5 Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. Proc. of the Conference on Natural Language Learning (CoNLL). 2003. 25-32.

6 Wiebe J, Wilson T. Learning to disambiguate potentially subjective expressions. Proc. of the Conference on Natural Language Learning (CoNLL). 2002. 112-118.

7 Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

8 Ni XC, Xue GR, Ling X, Yu Y, Yang Q. Exploring in the weblog spaceby detecting informative and affective articles. Proc. of WWW, 2007. Industrial practice and experience track.

9 Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proc. of the Association for Computational Linguistics (ACL), 2004. 271-278,

10 Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: Sentiment tag extractionfrom WordNet glosses. Proc. of the European Chapter of the Association for Computational Linguistics (EACL). 2006.

11 Gilad-Bachrach R, Navot A, Tishby N. Margin Based Feature Selection: Theory and Algorithms. International Conference on Machine Learning (ICML). 2004.

12 Kim SM, Hovy E. Automatic detection of opinion bearing words and sentences. Companion Volume to the Proc. of the International Joint Conference on Natural Language Processing (IJCNLP). 2005.