第36卷 第15期 计 算 机 工 程 2010年8月
Vol.36 No.15 Computer Engineering August 2010
·博士论文· 文章编号：1000—3428(2010)15—0020—03 文献标识码：A 中图分类号：TP391.1

新闻报道文本的情感倾向性研究

昝红英[1]，郭 明[1]，柴玉梅[1]，吴云芳[2]
(1. 郑州大学信息工程学院，郑州 450001；2. 北京大学计算语言学研究所，北京 100871)

摘 要：将机器学习中的经典分类方法与规则方法相结合，用以分析新闻语音文本的情感倾向，并判断其强弱。通过支持向量机分类器来研究特征选择方法及特征权重计算方法的组合对实验结果的影响。在实验过程中发现适当的结合规则后，实验结果在不同程度上都有了提高，在KNN和Bayes分类器上做了对比实验，结果证实该方法的普适性。

关键词：机器学习；特征选择；特征权重；支持向量机

Research on News Report Text Sentiment Tendency

ZAN Hong-ying[1], GUO Ming[1], CHAI Yu-mei[1], WU Yun-fang[2]
(1. College of Information and Engineering, Zhengzhou University, Zhengzhou 450001;
2. Institute of Computational Linguistics, Peking University, Beijing 100871)

【Abstract】This paper uses machine learning techniques combined with the rules to solve sentiment classification of news text, and researches affection of feature selection and feature weights based on Support Vector Machine(SVM) classifier. Experiments show that combined with the rules, experimental result is improved. In order to test universality of the combined method, more experiments based on KNN and Bayes classifier are done. Results show that combined method does better than not combined ones.

【Key words】machine learning; feature selection; feature weight; Support Vector Machine(SVM)

1 概述

在较长一段时期内，情感一直位于认知科学研究者的视线之外。直到20世纪末期，情感作为认知过程重要组成部分的身份才得到了学术界的普遍认同。

随着互联网的发展而迅猛地发展着，大量的文本充斥着网络。数量已大大超出了手工能够处理的极限。因此，对文本的自动分析处理成为一项重要的工作。文本的情感分析成为自然语言处理的研究热点。

本文主要研究如何利用机器学习的方法结合规则的方法来对新闻文本的正负极性及其强弱进行分类。相比于文本，比如BLOG的文本，新闻文本有其自身的特点，本文就是抓住新闻文本的这些特点来抽取规则，并利用这些规则结合机器学习的方法进行分类。

2 相关研究

现阶段对文本情感的研究总体可分为2类：

(1)利用语言自身的特点，总结出一套规则，并利用这套规则来分析文本的情感[1-2]。

(2)利用统计学的方法，对文本中出现的词进行统计得出统计信息来进行情感分析[3-4]。

许多研究者从词、短语、句子，到篇章等多个角度来进行探索。例如文献[2]利用点对互信息(PMI)来对词的语义倾向进行判断。文献[3]则使用了贝叶斯、最大熵、SVM等算法对电影的评价进行分类，其结果显示SVM算法取得了比较好的效果。

在中文情感分析领域，文献[4]以句子为单位对用户评论进行特征标注。将具有多个特征的复杂特征句划分为简单特征句，分析评论句的语义极性和极性强度。文献[5]研究多种分类器在中文情感分类中的表现，并且针对不同词性的词做特征来测试词性在不同类别文本(如影视、教育、房产、电脑、手机)的情感分类中的重要性。文献[6]使用朴素贝叶斯和最大熵方法进行新闻及评论语料的情感分类研究，通过实验认为二值作为特征项权重正确率较好，但不足之处就是未用SVM分类器。

3 文本情感分析模型

本文结合使用统计的方法和简单的规则方法来对新闻语料进行情感分类，采用结合规则的方法是基于以下考虑：

(1)规则包含一些语意信息，这是统计所不具备的，且本文人工提取的规则又经过统计方法的筛选以提高规则覆盖语料的正确率。

(2)统计方法可以整体衡量训练语料中词语出现的频度及分布等一些统计信息，并且能评价这些词语对分类的贡献大小。

本文用规则的方法对测试语料进行第1次分类，分出规则能够覆盖的语料。第2次再将剩下的测试语料用统计的方法进行分类来提高分类效果。

基金项目：国家“863”计划基金资助项目(2007AA01Z198)；国家自然科学基金资助项目(60970083)；国家社会科学基金资助项目(08CYY016)

作者简介：昝红英(1966－)，女，副教授、博士，主研方向：自然语言处理，文本挖掘；郭 明，硕士；柴玉梅，副教授、硕士；吴云芳，副教授、博士

收稿日期：2010-02-25 **E-mail：**iehyzan@zzu.edu.cn

3.1 规则提取及其效果

用规则来进行情感分析的关键点在于找到一种可靠的区分文本情感的方法，尽量保证用规则的正确率。本文根据新闻事件的性质和给社会带来的影响区分文本的极性，通过对语料的分析发现新闻文本的极性一般可以由表明这篇新闻报道性质的句子来定性。所以，提取该句子中能表明本篇极性的短语对作为备选规则，统计这些搭配在训练集中的分布，去除在多个类别中出现的搭配，剩余部分作为本文分类的规则。

为了体现否定词对语意的影响，本文取这些规则前开 3 个词的窗口，如果出现否定词或者一些模棱两可的词如：是不是，是否等，则认为这些规则不适用。提取规则如表 1 所示。

表 1　提取规则

规则类别	规则
赞扬类规则	做出 突出 贡献、公仆 楷模、美不胜收、赢得 声誉、喜结良缘、汗马功劳、动人 故事、白衣天使、……
喜悦类规则	论坛 开幕、庆祝 周年、进展 顺利、圆满 结束、通过 验收、摧毁 团伙、获? 殊荣、打破 纪录、……
中性类规则	今日 天气
悲哀类规则	遭? 袭击、?船 沉没、导致 中断、台风 登陆、爆发 疫情、?气 泄露 陷入 瘫痪、引发 火灾、……
贬斥类规则	引起 不满、表示 愤慨、给予 罚款、参差不齐、带来 危害、天花乱坠、逮捕 嫌疑人、瘦肉精、……

规则中的“?”为通配符：“?气 泄露”中的“?气”可以是毒气、氯气等。

规则覆盖语料的实验正确率如表 2 所示，数据表明利用规则分离出来的各类语料的正确率除赞扬类外都在90%以上。

表 2　规则分离出来的语料的正确率

赞扬	喜悦	中性	悲哀	贬斥
0.667	0.915	1	0.933	1

通过观察语料发现，赞扬类正确率低的主要原因是赞扬类的一些规则覆盖了喜悦类语料。这种错误不存在极性的差别，仅是程度的区别，是可以接受的。

3.2 统计方法

3.2.1 特征选择

常见的特征选择方法有：基于文档频率的特征选择法，信息增益法，X^2 统计量等，国内外很多学者对各种特征选择方法进行研究。结果表明在英文文本分类中表现比较好的方法如：信息增益法(IG)和 X^2 统计量(CHI)在不加修正的情况下，并不适合中文文本分类。因此在本文中对所采用的特征选择方法做了些限制，例如：去掉停用词，去掉单字词等。

(1)信息增益法

信息增益：依据某特征项 t_i 为整个分类所能提供的信息量多少来衡量该特征项的重要性，从而决定该特征的取舍。信息量的大小由熵来衡量：

$$Gain(t_i) = Entropy(s) - ExpectedEntropy(S_{t_i})$$

(2)X^2 统计量

X^2 统计量：衡量的是特征项 t_i 和类别 C_j 之间的相关程度。

特征项 t_i 对于类别 C_j 的 X^2 值表示如下：

$$X^2 = \frac{N\times(A\times D - C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中，A、B、C、D 的含义如表 3 所示。

表 3　公式说明

是否包含特征	属于类 C_j	不属于类 C_j
包含特征 t_i	A	B
不包含特征 t_i	C	D

(3)特征维数的选择

对于大量文本都存在“维数灾难”问题，如何有效地降低维数并尽可能地减少噪声是文本特征提取的关键问题。在实验过程中发现，并非特征数量越多效果就越好。

从图 1 看出，特征数量在取 500 的情况下，所有方案的正确率都基本趋于稳定且正确率基本达到最大值。

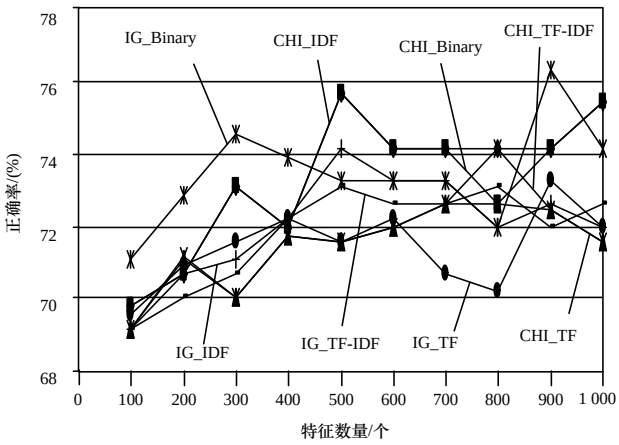


图 1　分类正确率与特征数量的关系图

3.2.2 特征权重计算

特征权重用于衡量特征项 t_i 在文档表示中的重要程度或者区分能力的强弱。本文所用特征权重计算方法如下：

布尔权重：如果文本中出现该特征则该特征权重为 1，否则为 0，表示为 $W_{ij}=\begin{cases}0 & tf_{ij}>0\\ 1 & \text{else}\end{cases}$。

绝对词频(TF)：特征项在文本中出现的次数。

倒排文档频度(IDF)：

$$W_{ij} = \mathrm{lb}\frac{N}{n_i}$$

其中，n_i 是训练集中出现特征项 t_i 的文档数；N 为训练集全部的文档数。

TF-IDF：

$$W_{ij} = tf_{ij}\times \mathrm{lb}\frac{N}{n_i}$$

其中，tf_{ij} 表示特征项 t_i 在训练文本 D_j 中出现的频度；W_{ij}、N、n_i 同上。

3.3 分类器选择

本文主要以经典的统计分类方法 SVM 分类器为主，然后再用 Bayes 和 KNN 分类器与规则的方法结合做对比实验，以验证这种结合的方法的普适性。

4 实验与结果分析

4.1 语料简介

实验所用的是从 2006 年 6 月 5 日-2008 年 6 月 26 日播报的新闻中的新闻稿做为语料，共计 2 287 篇，并对这些语

料在播音员播报新闻时的态度做人工标注为赞扬、喜悦、中性、悲哀、贬斥 5 类，见表 4。

表 4　人工校对后语料的统计分布情况

态度	篇数	比例/(%)
赞扬	190	8.3
喜悦	1 322	57.8
中性	392	17.1
悲哀	279	12.2
贬斥	104	4.5
总数	2 287	100.0

由表 4 可以看出，本文使用的语料分布极不平衡，其中，喜悦类占整个语料的 57.8%，而贬斥类仅占了 4.5%。

4.2　结果分析

实验结果如表 5 所示。

表 5　各分类器的正确率　(%)

特征选择	权重计算	SVM	RSVM
CHI	Binary	74.84	77.46
	TF	71.55	73.74
	IDF	74.84	77.46
	TF-IDF	71.55	74.40
IG	Binary	73.30	73.30
	TF	71.55	73.30
	IDF	73.30	75.24
	TF-IDF	71.55	75.24

结合表 5 可以看出，SVM 及在 SVM 算法上的改进方法在新闻语音文本情感倾向计算方面都取得基本令人满意的效果，在以上实验中，CHI 结合 IDF 和 CHI 结合 Binary 的效果最好，说明文本的情感与一些情感特征词是否出现的关联性更大，而与该特征词在文本中的词频关系不大。

仅从正确率来评价实验效果似乎有些片面。表 6 为取实验效果最好的 X^2 和 IDF 的组合和效果最差的信息增益和 TF 组合来对比分析每次实验的召回率及 $F1$ 指标。

表 6　各指标对比

类别	算法	X^2_IDF		IG_TF	
		召回率	$F1$ 值	召回率	$F1$ 值
赞扬	SVM	0.632	0.706	0.421	0.508
	RSVM	0.541	0.645	0.459	0.566
喜悦	SVM	0.955	0.848	0.951	0.834
	RSVM	0.963	0.864	0.959	0.827
中性	SVM	0.416	0.522	0.372	0.424
	RSVM	0.468	0.563	0.390	0.488
悲哀	SVM	0.537	0.617	0.491	0.574
	RSVM	0.611	0.683	0.491	0.621
贬斥	SVM	0.333	0.482	0.142	0.230
	RSVM	0.333	0.466	0.190	0.319

由表 6 可以看出，加入规则后(RSVM)的实验效果均要优于 SVM 算法。语料中喜悦类在数量上占有优势造成喜悦类的召回率非常高，而贬斥类则相反。贬斥类的 $F1$ 值最低，而且结合规则方法对贬斥类基本没有影响，这和贬斥类的语料所占总语料比例太小有很大关系。

为了检验规则和统计结合方法的普适性，本文在 SVM 的基础上做了推广，分别在 KNN 和 Bayes 分类器上做实验，结果如图 2 所示。

从图 2 中可以看出，结合规则以后各类的 $F1$ 指标都有了不同程度的提高。SVM 算法效果要优于 Bayes 和 KNN 算法，同样地与规则结合后的 SVM 算法效果也优于与规则结合后的 Bayes 和 KNN 算法。

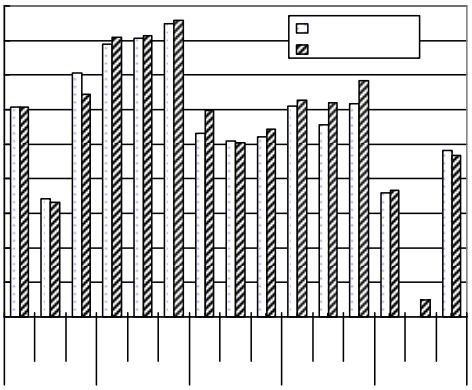


图 2　Bayes KNN SVM 分类器结合规则后的 $F1$ 指标对比

5　结束语

SVM 算法在文本情感分析中取得了基本令人满意的效果，将规则和 SVM 分类算法结合后的实验效果又有了进一步的提升。通过本文的实验可以发现，这种语料分布的极不平衡性对稀有类的实验效果有很大的影响。如何提高稀有类的实验效果，并抑制优势类对其他类的影响将是另一个难题。

参考文献

[1] Liu Hugo, Lieberman H, SelKer T. A Model of Textual Affect Sensing Using Real-world Knowledge[C]//Proc. of International Conference on Intelligent User Interfaces. Miami, Florida, USA: [s. n.], 2003: 125-132.

[2] Turney P D, Littman M L. Measuring Praise and Critism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[3] Pang Bo, Lee Lilian, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques[C]//Proc. of Conferenee on Empirieal Methods in Natural Language Processing. Morristown, NJ, USA: [s. n.], 2002: 79-86.

[4] 潘　宇, 林鸿飞. 基于语义极性分析的餐馆评论挖掘[J]. 计算机工程, 2008, 34(17): 208-210.

[5] 唐惠丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术的比较研究[J]. 中文信息学报, 2007, 21(6): 88-94.

[6] 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.

编辑　任吉慧