基于语义理解和机器学习的混合的中文文本情感分类算法框架

徐健锋[1] 许 园[1] 许元辰[2] 张远健[1] 刘 清[2]

(南昌大学软件学院 南昌 330047)[1] (南昌大学信息工程学院 南昌 330031)[2]

摘 要 快速、准确和全面地从大量互联网文本信息中定位情感倾向是当前大数据技术领域面临的一大挑战。文本情感分类方法大致分为基于语义理解和基于有监督的机器学习两类。语义理解处理情感分类的优势在于其对不同领域的文本都可以进行情感分类,但容易受到中文存在的不同句式及搭配的影响,分类精度不高。有监督的机器学习虽然能够达到比较高的情感分类精度,但在一个领域方面得到较高分类能力的分类器不适应新领域的情感分类。在使用信息增益对高维文本做特征降维的基础上,将优化的语义理解和机器学习相结合,设计了一种新的混合语义理解的机器学习中文情感分类算法框架。基于该框架的多组对比实验验证了文本信息在不同领域中高且稳定的分类精度。

关键词 情感分类,语义,机器学习

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.014

Hybrid Algorithm Framework for Sentiment Classification of Chinese Based on Semantic Comprehension and Machine Learning

XU Jian-feng[1] XU Yuan[1] XU Yuan-chen[2] ZHANG Yuan-jian[1] LIU Qing[2]

(Software College, Nanchang University, Nanchang 330047, China)[1]

(School of Information Engineering, Nanchang University, Nanchang 330031, China)[2]

Abstract In the background of big data, it is a major challenge to distinguish sentiment orientation from a large number of Internet text information quickly, accurately and comprehensively. The main sentiment classification methods of text information are roughly divided into two categories: one is semantic comprehension and the other is supervised machine learning. The advantage of dealing with sentiment classification by using semantic comprehension method is that it can classify the text in different fields. However, the performance can be greatly affected by avariety of word collocations and sentence patterns. The supervised machine learning method can achieve higher classification accuracy, however, a satisfying classification classifier in a field may not be suitable for a new field. This paper proposed a new hybrid algorithm framework for Chinese sentiment classification combining optimized semantic comprehension and machine learning based on the features extracted by information gain. Experimental results on two separate fields show that this framework has both high classification accuracy and satisfying portability.

Keywords Sentiment classification, Semantic, Machine learning

1 引言

伴随着计算机和互联网技术的蓬勃发展,网络已经成为人们发布信息和获取信息的重要场所,它使得信息的广泛传播不再依赖于传统的广播电视。据CNNIC近期发布的第33次中国互联网发展状况统计报告显示,截至2013年12月,中国网民规模达6.18亿,全年新增网民5358万人,互联网普及率为45.8%[1]。越来越多的人愿意通过互联网渠道来表达自己的观点和看法。近几年来,国内外发生的重大事件,几乎是第一时间在网络上引发剧烈的反响和激烈的辩论,并且这些言论能够起到引导舆论的作用。正因为如此,其中一些虚假的信息和言论会引起大众的恐慌,造成恶劣的社会影响,危害人民安全和社会稳定。快速、准确和全面地从网上海量信息中定位情感倾向是当前信息科学与技术领域面临的一大挑战。虽然网络信息的形式多种多样,但是它们的主要载体是文本,因此从网络上及时获取相应的文本信息有着巨大的现实意义和研究价值。由于人工进行分析需要消耗大量的人力、物力和财力,因此如何采用计算机去自动分析这些文本的情感倾向并加以分类就成为当前学术界研究的一个热点。

文本情感分类的研究在近几年逐步成为学术界的热点课题,也取得了相对不错的研究成果。国外对于英文文本的情感倾向分类研究比较多,COLING、ACL、AAAI等涉及自然语言处理、人工智能、数据挖掘等领域的国际顶级会议都收录了关于文本情感倾向分类的论文[2]。为了推动国内关于文本

到稿日期:2014-04-28 返修日期:2014-05-30 本文受本体学习与粒计算基金(61070139)资助。

徐健锋(1973-),男,副教授,硕士生导师,主要研究方向为粒计算、人工智能,E-mail:jianfeng_x@ncu.edu.cn;许 园(1992-),女,硕士生,主要研究方向为数据挖掘、Web智能;许元辰(1991-),男,硕士,主要研究方向为数据挖掘、Web智能;张远健(1990-),男,硕士生,主要研究方向为粒计算、机器学习;刘 清(1938-),男,教授,博士生导师,主要研究方向为人工智能、Rough逻辑。

情感倾向分类研究的发展以及情感语料库的建设，中国中文信息学会信息检索专业委员会于2008年推出了中文倾向性分析评测(Chinese Opinion Analysis Evaluation)[3]，该评测不仅推动了国内的研究，同时还产生了大量的优秀论文。如今，该评测已经举办了5次，很好地推进了国内研究发展的进度。

中文文本情感倾向分类研究在近几年蓬勃发展，在此基础上，基于语义理解的文本倾向分类的优点在于无监督、不需要训练语料、可移植性能好，缺点在于精度不高；而有监督的机器学习方法的优点在于精度高，但是在分类前需要人工标注语料，移植性不够好[4]。本文充分发挥两种方法的优势，将优化的语义理解方法与机器学习相结合，构建一个全新的文本情感分类算法框架。该框架首先使用优化语义理解方法标注类别隶属信任度高的样本，之后采用这些标注的样本进行机器学习，不仅提高了模型的可移植性，而且在训练文本的选取方面提供了一定的理论指导，从而代替了盲目随机选择训练样本集。实验表明，与传统方法相比，本方法具有一定的优势。

2 相关理论研究

从研究技术的角度上来看，目前文本情感倾向分类研究大致分为两类，即基于语义理解的文本情感分类研究和基于机器学习的文本情感分类研究。

2.1 基于文本情感分类的主要研究方法

2.1.1 基于语义理解的文本情感分类的研究方法

基于语义理解的文本情感分类研究方法主要分为两类。

第一类是基于短语模板，这类方法首先需要从待分类的文本中抽取出能体现情感倾向性或者主观色彩的名词短语或者形容词短语，然后依据一些事先定义好的规则判断这些短语是否具有倾向性，并且赋予一定的倾向性程度，最后累加所有的倾向性值，得到文本的总体情感倾向性。Hatzivassiloglou和McKeown等人[5]通过判断连接不同形容词或名词的连词的语言学性质来判断所连接的这些词的倾向性是否一致，然后通过聚类的方法来获得这些具有倾向性的词语。Recchia G等人[6]使用点互信息(Pointwise Mutual Information)技术来评估词语或者短语与表示褒贬两个极性的基准词的相似度，通过这个相似度来判断这些词语或者短语是属于褒贬哪一类的，从而判断出其倾向性，甚至可以通过这个相似度来制定出这些词语或者短语的倾向性程度。针对褒贬两个极性的基准词，很多学者提出使用本体知识库(比如WordNet[7]和知网HowNet[8])，并从这些语料库中抽取出基准词，利用这些语料库中词与词之间的各种依赖关系，更加准确地判断待评估的词语与基准词的相似度，最后判断出待估词的倾向性以及倾向性程度。Kamps等人[9]就使用了WordNet，通过WordNet中的同义词结构来计算待评估的词语与选定的褒贬基准词的近似程度，从而确定待评估词语的倾向程度。朱嫣岚和闵锦[10]则利用知网HowNet提供的语义相似度计算功能，来计算待评估的词语与选定的褒贬基准词组的相似度，以确定待评估词语的倾向程度。

第二类是基于语义模式库。首先由语言专家建立一个倾向性语义模式库，同时他们会建立一些辅助的情感词典、程度词典等相关辅助信息，然后待分析的文本依据语义模式库进行模式匹配，最后将所有匹配的模式累计，计算该文本的倾向性及强烈程度。Jeonghee Yi等人[11]所采用的方法就是使用一个已经备好的具有倾向性词汇的表格，和一个已经提炼好的倾向性的模式库，通过制定的规则对句子和短语进行语义分析，进而获取句子和短语的倾向性。刘永丹和曾海泉[12]通过语法和语义框架描述评论文本中的语义关系，利用已有的语义分析技术，判断文本的倾向性。何凤英[13]则是采用了一个倾向性的词典和特制的语义匹配规则来对博客评论文本做倾向性分析。

2.1.2 基于机器学习的文本情感分类研究方法

基于机器学习的文本情感分类是基于传统的文本分类技术，主要思想是先通过专家事先标注一些文本的倾向性，并将这些文本作为训练语料，通过机器学习的方法按照特殊的倾向性要求构造一个分类器，最后使用这个分类器对实验语料进行分类，识别出该文本的具体分类。Bo Pang等人[14]早在2002年就分别使用支持向量机(Support Vector Machines)方法、最大熵(Maximum Entropy)方法和朴素贝叶斯(Native Bayes)方法对倾向性文本进行了对比实验，分析之后，发现支持向量机方法的效果是最好的。徐琳宏等人[15]在实验中选取程度比较强烈的具有褒义或者贬义的词作为特征项，构造了一个支持向量机的褒义、贬义两类分类器，该分类器取得了相对可观的分类效果。唐慧丰等人[16]根据特征表示、选择等文本分类中的关键技术也进行了对比实验，采用二元特征表示方法、支持向量机分类方法和信息增益特征选择方法也取得了相对不错的分类效果。

除此之外，一些用于商业领域的文本情感分类系统也相继问世，如商用产品信息反馈系统Opinion Observer[17]，它利用网上大量的顾客评论，进行了产品的市场反馈分析，为消费者和生产者提供针对商品各个特性的评价报告。

2.2 文本情感分类流程

2.2.1 基于语义理解的文本情感分类流程

基于语义理解的文本情感分类方法，大致分为以下几个步骤。

步骤1 构造初始情感词词典。情感词主要来源于知网等相关知识库，人工进行初步筛选后构成初始情感词词典。

步骤2 情感词权值的确定。词汇不同，其表达的褒贬程度也是所有差别的，使用合理的方法如投票、PMI、语义相似度等计算情感词权值。

步骤3 文本预处理。将待分类的文本进行中文分词以及词性标注，之后根据规则筛选出情感词句。

步骤4 文本情感分类。根据情感词权值以及情感句句法分析最终计算出文本的情感权值，从而判定文本的褒贬类别。

2.2.2 基于机器学习的文本情感分类流程

基于机器学习的文本情感分类方法通过学习领域专家所确定的已分类的文本集合，构建出相应的分类器，然后利用学习得到的知识规则指导新的测试文本的学习，并将测试的数据集划分到对应的类别中。该方法大致分为以下几个步骤。

步骤1 训练集的选择。训练集来源于领域专家确定的已分好类别的文本。

步骤2 文本预处理。针对确定类别的语料进行分词、去停用词、特征选择与特征提取，构造特征向量。

步骤3 训练分类器。对训练集中的每个文本使用某种

统一的形式进行表示，之后将其作为输入数据使用分类算法进行学习，最终得到合适的分类器。

步骤4　使用分类器。选择测试集语料，通过文本预处理，将测试文本构成的特征向量作为输入数据，测试分类器的准确率等相关指标。

3　混合语义理解的机器学习中文情感分类框架

利用语义理解进行情感分类，主要依靠对情感词、情感语句的句法分析最终得到文本的褒贬分类。该方法最大的优势在于对不同领域的文本都可以进行情感分类，移植性比较好；但由于中文表达的复杂多变，不同的修饰、句式都会影响到文本的情感表达，因此单纯地使用语义理解来进行情感分类，分类精度会受到限制。

有监督的机器学习，例如SVM，在情感分类精度上能够达到比较高的标准。该方法的不足在于可移植性比较差，同一个分类器在一个领域方面达到较高的分类能力，当换到一个新的领域后，其效果会大打折扣。同时SVM是被动地随机选择训练样本，被动地接受这些样本的信息进行学习，由于训练样本都是经过人为标注的，既耗费了人力、物力，又含有一定的主观性，对分类器的精度造成一定的影响。

针对两种方法的优势，在综合前两类研究方法的基础上将优化的语义理解和机器学习相结合，设计一种新的文本情感分类模型，如图1所示。其主要思想为：

第一，在训练样本集已标注的领域，首先使用优化的语义理解对训练样本进行分类，然后根据分类后文本情感倾向值的不同，筛选出情感程度明显的样本作为机器学习的训练样本，即在训练文本上的选取不再是随机进行，而是择优选择；

第二，在训练样本集类别未知的领域，充分发挥语义理解方法良好的移植性能，将训练样本分类后进行标注，之后作为机器学习的训练样本，从而使本文设计的情感分类器具有良好的可移植性。

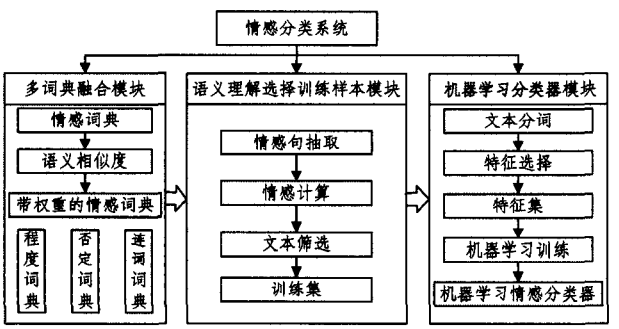

图1　文本情感分类算法框架

(1)多词典融合模块主要功能

1)情感词典：从知网提供的情感词库中选择词汇并人工去除掉不常用的词汇，汇总成初始情感词典。

2)语义相似度：构建褒贬情感基准词典，在选择过程中遵循词频高、情感倾向明显、数量具有一定规模3个方向，之后采用知网语义相似度计算公式计算初始情感词典中情感词的权重。

3)多词典融合：去除情感词典中一些错误的以及情感倾向不明显的词汇，之后结合程度词典、否定词典以及连词词典，构建成多词典模块。

(2)语义理解选择训练样本模块主要功能

1)情感句抽取：根据情感句的常用搭配模式，结合情感词典，从文本中抽取出情感表达语句。

2)情感计算：结合多词典模块，针对词语之间的句式搭配，使用情感句倾向计算算法进行判别。

3)文本筛选：将训练样本根据情感倾向值大小进行排序，之后选取情感倾向值靠前的文本作为机器学习的择优训练样本。

(3)机器学习分类器模块主要功能

1)文本分词：对语义理解选择出来的训练文本进行分词处理，在使用过程中使用中科院汉语词法分析系统(ICTCLAS)进行分词与词语标注，之后去除停用词，形成最初的特征集合。

2)特征选择：文本分词后，利用特征选择方法在原始特征集合中选择出具有代表性的特征词汇构成分类特征子集。

3)特征权重计算：将特征子集中的每个特征使用TF-IDF(Term Frequency Inverse Document Frequency)方法得到权重，之后将每个文本按照特征子集表示为特征向量。

4)机器学习训练：对输入的数据进行归一化，之后进行机器学习训练，最终构建成一个情感分类器模型。

5)情感预测：对未知的测试数据样本进行分词处理，使用分类特征集合、特征权重计算，对测试数据样本进行文本表示，之后使用机器学习训练出来的情感分类器模型进行预测，得到情感分类结果，之后根据分类的评判标准衡量分类器的可行性。

4　实验分析

4.1　实验数据集

为了验证本算法框架在文本情感分类上的有效性，需要一个经过专家标注的文本集合作为基准测试数据集进行测试。在中文情感文本语料上，目前国内使用得最多的数据集是COAE中提供的40000份文本和由中科院谭松波博士的团队近几年收集和整理的语料。本文选取谭松波团队整理的关于酒店评论[19]的语料来进行实验。从4000份标注的酒店评论文本中选取300个文本，含150个褒义文本和150个贬义文本，随机选取3次构成3组不同的实验数据集，如表1所列。

表1　3组随机实验数据集文本分布

	文本总数	正面文本数量	负面文本数量
第一组	300	150	150
第二组	300	150	150
第三组	300	150	150

4.2　文本分类评估标准

传统意义上一般使用查全率和查准率来度量文本分类的质量，其中查全率定义为正确判别为该类的测试样本占该类总测试样本的比例，查准率定义为正确判别为该类的测试样本占判别为该类测试样本的比例[20]。然而，同时达到高查准率和高查全率是困难的。在综合考虑查准率和查全率的基础上提出了一种综合评估标准，即F-测量值。一般的分类器是靠这3个指标来衡量的，它们的公式表述见式(1)－式(3)。

$$\text{查准率：}P=\frac{\text{某类中判断正确的文本数}}{\text{判断为该类的文本总数}} \tag{1}$$

$$\text{查全率：}R=\frac{\text{某类中判断正确的文本数}}{\text{实际为该类的文本总数}} \tag{2}$$

F-测量值：$F=\frac{2\times P\times R}{P+R}$ (3)

4.3 实验结果与分析

4.3.1 基于文本情感分类算法框架的多机器学习分类器的性能分析

在实验过程中，分别使用决策树(Decision Tree，DT)、朴素贝叶斯(Naive Bayes，NB)、K 最近邻(K-Nearest Neighbour，KNN)以及支持向量机(Support Vector Machine，SVM)4 种方法进行对比。首先使用信息增益方法从原始的 4000 个特征中选取出 400 个特征，分别计算其 TF-IDF 值，归一在后分别进行训练；然后用测试语料进行验证，实验结果如表 2 所列。

表 2 不同的机器学习算法在情感分类中的性能评价

		第一组	第二组	第三组	平均值
DT	查准率(%)	78.7%	77.2%	78.55%	78.15%
	查全率(%)	68%	68.5%	72%	69.5%
	F-测量(%)	72.96%	72.59%	75.13%	73.56%
NB	查准率(%)	79.8%	78.08%	78.82%	78.9%
	查全率(%)	78.5%	77.56%	78.5%	78.19%
	F-测量(%)	79.14%	77.82%	78.66%	78.54%
KNN	查准率(%)	71.11%	75.09%	77.63%	74.61%
	查全率(%)	70.56%	74.52%	77.5%	74.19%
	F-测量(%)	70.83%	74.80%	77.56%	74.40%
SVM	查准率(%)	80.02%	79.29%	80.86%	80.06%
	查全率(%)	77.5%	79%	80.5%	79%
	F-测量(%)	78.74%	79.14%	80.68%	79.52%

由表 2 可以看出，SVM 和朴素贝叶斯的分类结果较好，其中 SVM 的查准率、查全率以及 F-测量值都达到了最高，所以本文的文本情感分类器最终选择 SVM 方法。

4.3.2 语义理解方法与 SVM 相结合的文本情感分类器性能分析

本实验的主要流程如下：通过语义理解对训练样本集中的 600 篇文本进行文本分类，然后选择出情感倾向明显的前 300 篇作为 SVM 的训练文本，之后使用测试语料集中的 3 组数据验证分类器的性能。在特征选择过程中分别使用文档频率(DF)、χ^2 统计法、信息增益(IG)和语义理解方法[18]进行特征选择，分别对比每组得到的实验结果。详细结果如表 3 所列。

表 3 本算法框架与经典方法性能分析比较

		第一组	第二组	第三组	平均值
文档频率(DF)	查准率(%)	82.90%	81.53%	84.92%	83.12%
	查全率(%)	82.5%	81.5%	84.5%	82.83%
	F-测量(%)	82.70%	81.51%	84.71%	82.97%
χ^2 统计法	查准率(%)	80.54%	81.5%	83.66%	81.9%
	查全率(%)	80.2%	81%	83.5%	81.57%
	F-测量(%)	80.37%	81.25%	83.58%	81.73%
信息增益(IG)	查准率(%)	83.5%	83%	87.96%	84.82%
	查全率(%)	83%	82.6%	87.5%	84.37%
	F-测量(%)	83.25%	82.8%	87.73%	84.59%
语义理解方法	查准率(%)	84%	84.55%	88.86%	85.8%
	查全率(%)	83.5%	84.2%	88.5%	85.4%
	F-测量(%)	83.75%	84.37%	88.68%	85.6%

由表 3 可以看出，虽然使用了少量的情感表现明显的样本对 SVM 进行训练，但是得到的分类器的分类性能超过了全量训练样本训练后得到的结果。原因在于：在全部的训练文本中，可能存在一些分类标注正确但是存在一定噪声的样本，这些样本的出现在一定程度上降低了分类器的准确度。所以在已标注的训练文本集中，使用语义理解方法择优选取训练样本不仅能够减少训练的时间，而且对准确率也有一定的提高。

4.3.3 基于语义理解选取训练样本对混合框架情感分类的准确率的分析

本实验的目的是衡量不同训练样本选择方式(随机选取样本、传统语义理解以及本文的语义理解选取训练样本)对分类准确率的影响。训练样本来自训练语料集中的 600 个(300 对)文本，训练文本的数目分别选取 50 对、100 对、150 对、200 对，实验结果如图 2 所示。

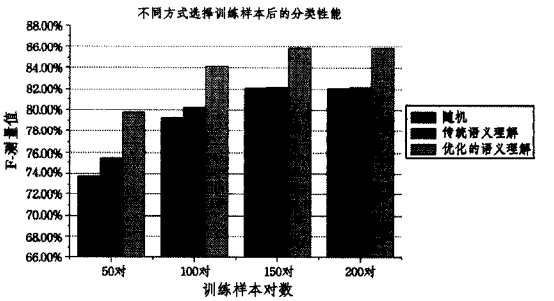

图 2 不同选择训练样本方式的分类性能对比结果

由图 2 可以看出，本算法框架的语义理解择优选择出的训练样本在进行 SVM 训练后得到的分类器的分类性能最好。这是因为传统的语义理解由于其准确率的问题，在少量的样本选取中比随机的效果好，当样本数目增多时其效果和随机选取的非常接近，而经过优化语义理解方法选择出来的样本在数量较少时就可以达到较高的准确率。

在分类精度优于传统分类器的前提下，构建分类器的时间复杂度可以用式(4)来度量：

$$O(Mm^2)+O(n\times m)+O(n_1^{2.2}) \quad (4)$$

其中，n 代表训练样本的个数，m 代表特征的个数，M 代表信息增益选择前的原始特征数，n_1 代表具有典型情感倾向的文本数量，通常 $n_1\ll n$，$m<n$。$O(n\times m)$代表筛选出具有典型情感倾向的文本(对训练文本使用信息增益提取得到的特征匹配)所需要的时间复杂度。$O(Mm^2)$[21]代表筛选特征需要的计算量。在筛选出的特征数量较少的前提下，本混合框架的时间复杂度显然低于经典的 SVM 时间复杂度 $O(n^{2.2})$[22]。

通过 4.3.2 和 4.3.3 节的对比实验可以看出，选择情感程度明显的少量训练样本，无论在训练时间还是分类精度方面都是有一定意义的。

4.3.4 混合框架移植性能分析

基于 SVM 的文本情感分类方法通常具有较高的精度，但前提是每个领域都需要大量带有标签的训练样本，而这往往需要大量的收集成本，所以 SVM 不具有良好的移植性；而语义理解方法在情感分类过程中精度不会很高，但是它不需要训练集，所以在一个没有训练样本标注的领域，该方法一样可以使用。本实验的目的在于验证基于优化的语义理解和 SVM 相结合的方法的可移植性能，语料集选取谭松波团队收集的笔记本电脑评论语料[19]，该语料包含 2000 篇褒义文本以及 2000 篇贬义文本。随机选取 1000 对褒贬文本作为语料进行实验，步骤如下：

(1)打乱 1000 对已标注好的情感语料模拟未知语料，随后使用优化的语义理解对这些语料进行情感分类，选取情感

程度靠前的 K 对训练样本，然后使用 SVM 进行训练，得到分类器 1；

(2)在未打乱的 1000 对训练语料中，随机选取 K 对训练样本，使用 SVM 进行训练，得到分类器 2；

(3)在另外的 1000 对语料中随机选取 100 对文本(100 篇褒义文本，100 篇贬义文本)作为测试语料，根据不同的 K 值，通过对比 F-测量值来衡量本文方法的性能，实验效果如图 3 所示。

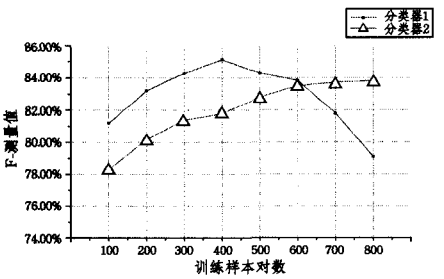

图 3　本文情感分类器的移植性能分析

由图 3 可以看出，本文的方法具有良好的可移植性，在训练样本情感类别未标记的情况下，首先使用优化的语义理解对语料进行标记，之后根据标记的样本训练 SVM，在样本数目较少的情况下达到了较高的准确率，最后随着训练样本对的增加，准确率降低。原因是基于语义理解的方法精度不高，随着样本数目的增加，在样本的标注上存在误标的可能性越来越大，从而降低了分类器的分类性能。而对于普通的 SVM 分类器，随着训练样本数目的增加，准确率在逐步上升，因为随着样本数目的增加，带有噪声的样本对分类器的影响会越来越小。

结束语　通过实验结果分析可知：本文提出的方法在训练文本的选择上有据可循，在使用情感程度明显的文本作为训练文本后，在少量训练集的情况下得到的分类器也具有很好的分类性能。这是因为，一方面，对传统的语义理解判断文本情感分类的方法做了优化，首先在情感词典的构建过程中选择了较为合适的褒贬基准词；其次对语义相似度公式进行微调，在保持较高准确率的前提下，解决了由褒贬基准词选取偏差带来的问题；之后采用情感词语搭配模式结合情感词典抽取情感句，并将程度词典、否定词词典、连词词典进行融合来判断文本的情感倾向，经过改进的算法的准确率得到了较大的提高。另一方面，对传统的信息增益特征选择方法作了优化，传统的信息增益是一种全局的特征选择方法，在特征的选择过程中偏重低频词的选取，这一定程度上降低了分类的准确性，经过改进的信息增益方法融入了词频和词语情感度的因素，在一定程度上增加了对中高频以及情感明显特征的词汇的选择。

综上分析，本文方法在训练文本类别已知的情况下，利用优化语义理解择优选取训练样本集，减少了训练时间并过滤掉了一些带有噪声的文本，最终得到的分类器具有较高的分类性能；在未知的领域即训练文本类别未知的情况下，利用优化语义理解的良好移植性选取少量的情感程度明显的文本作为标注文本，在训练样本数目选取较少的情况下达到了较高的准确率。

参 考 文 献

[1]　中国互联网信息中心. 第 33 次中国互联网络发展状况统计报告[EB/OL]. http://wenku.baidu.com/view/0d595bd0551810a6f5248694.html
China Internet Network Information Center. The 33rd Statistic Report for the development of China Internet Network [EB/OL]. http://wenku.baidu.com/view/0d595bd0551810a6f5248694.html

[2]　赵志伟. 中文文本倾向性分析[D]. 合肥：安徽大学，2012
Zhao Z W. Chinese text Orientation Analysis [D]. Hefei: Anhui University, 2012

[3]　赵军，许洪波，黄萱菁，等. 中文文本情感倾向性分析[J]. 中国计算机学会通讯，2008，4(2)：41-46
Zhao J, Xu H B, Huang X J, et al. Chinese text sentiment tendency analysis [J] Communication of China Computer Federation, 2008, 4(2): 41-46

[4]　赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848
Zhao Y Y, Qin B, Liu T. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848

[5]　Filatova E, Hatzivassiloglou V. A formal model for information selection in multi-sentence text extraction[C]// Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 397

[6]　Recchia G, Jones M N. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis[J]. Behavior research methods, 2009, 41(3): 647-656

[7]　WordNet[EB/OL]. http://wordnet.princeton.edu

[8]　知网[EB/OL]. http://www.keenage.com
HowNet [EB/OL]. http://www.keenage.com

[9]　Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientation of Adjectives[C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, 2004: 1115-1118

[10]　朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报，2006，20(1)：14-20
Zhu Y L, Min J, Zhou Y, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20

[11]　Yi J, Nasukawa T, Bunescu R, et al. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques[C]// the Third IEEE International Conference on Data Mining, November 2003. IEEE Computer Society Press, Los Alamitos, 2003: 427-434

[12]　刘永丹，曾海泉，李荣陆，等. 基于语义分析的倾向性文本过滤[J]. 通信学报，2004，25(7)：78-85
Liu Y D, Zeng H Q, Li R L, et al. Polarity text filtering based on semantic analysis [J]. Journal on Communications, 2004, 25(7): 78-85

[13]　何凤英. 基于语义理解的中文博文倾向性分析[J]. 计算机应用，2011，31(8)：2130-2133
He F Y. Orientation analysis for Chinese blog text based on semantic comprehension[J]. Journal of Computer Applications, 2011, 31(8): 2130-2133

[14]　Pang Bo, Lee L. Shivakumar Vaithyanathan. Sentiment Classification using Machine Learning Techniques[C]// the 2002 Conference on Empirical Methods in Natural Language Processing.

2002:79-86

[15] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):96-100
Xu L H, Lin H F, Yang Z H. Text Orientation Identification Based on Semantic Comprehension [J]. Journal of Chinese Information Processing, 2007, 21(1): 96-100

[16] 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 H F, Tan S B, Cheng X Q. Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques[J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94

[17] Liu Bing, Hu Min-qing, Cheng Jun-sheng. Opinion Observer: Analyzing and Comparing Opinions on the web[C]// the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 342-351

[18] 周城,葛斌,唐九阳,等. 基于相关性和冗余度的联合特征选择方法[J]. 计算机科学,2012,39(4):181-184
Zhou C, Ge B, Tang J Y, et al. Joint Feature Selection Method Based on Relevance and Redundancy[J]. 2012, 39(4): 181-184

[19] 情感评论语料[EB/OL]. http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp
Semantic Comment Corpus [EB/OL]. http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp

[20] 张启蕊,董守斌,张凌. 文本分类的性能评估指标[J]. 广西师范大学学报:自然科学版,2007,25(2):119-122
Zhang Q R, Dong S B, Zhang L. Performance Evaluation Metric for Text Classifiers[J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2): 119-122

[21] 王卫玲,刘培玉,初建崇. 一种改进的基于条件互信息的特征选择算法[J]. 计算机应用,2007,27(2):433-435
Wang W L, Liu P Y, Chu J C. Improved feature selection algorithm with conditional mutual information[J]. Journal of Computer Applications, 2007, 27(2): 433-435

[22] Platt J C. Fast Training of Support Vector Machines Using Sequential Minimal Optimization[M]// Schoelkopf B, Burges C, Smola A. Advances in Kernel Methods. Cambridge, USA: MIT Press, 1999: 185-208

(上接第 40 页)

并且,不是所有基因微阵列数据都适用于距离邻域的粗糙集模型,有些数据集在相交邻域下的分类效果更加出色。另外,如何对每一个基因设定一个合适的阈值,以及相容关系粗糙集与其它生物知识的结合应用,将是今后的研究重点。

参 考 文 献

[1] Piao Y, Piao M, Park K, et al. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data[J]. Bioinformatics, 2012, 28(24): 3306-3315

[2] Wang Shu-lin, Li Xue-ling, Zhang Shan-wen, et al. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction[J]. Computers in Biology and Medicine, 2010, 40(2): 179-189

[3] Tong Mu-chen-xuan, Liu Kun-hong, Xu Chun-gui, et al. An ensemble of SVM classifiers based on gene pairs[J]. Computers in Biology and Medicine, 2013, 43(6): 729-737

[4] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97(1/2): 273-324

[5] Wang Li, Zhu Ji, Zou Hui. Hybrid huberized support vector machines for microarray classification and gene selection[J]. Bioinformatics, 2008, 24(3): 412-419

[6] Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification[J]. Pattern Recognition, 2012, 45(1): 531-539

[7] Jiao Na, Miao Duo-qian. An efficient gene selection algorithm based on tolerance rough set theory[J]. Data Mining and Granular Computing, 2009, 5908: 176-183

[8] Pawlak Z. Rough sets[J]. Computer and Information Science, 1982, 11(5): 341-356

[9] Jensen R, Shen Q. Fuzzy-rough attribute reduction with application to web categorization[J]. Fuzzy Sets and Systems, 2004, 141(3): 469-485

[10] Paul S, Maji P. Rough set based gene selection algorithm for microarray sample classification[C]// International Conference on Methods and Models in Computer Science. New Delhi, 2010: 7-13

[11] Lu Zheng-cai, Qin Zheng, Zhang Yong-qiang, et al. A fast feature selection approach based on rough set boundary regions [J]. Pattern Recognition Letters, 2014, 36(15): 81-88

[12] 胡清华,于达仁. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报,2008,19(3):640-649
Hu Qing-hua, Yu Da-ren. Numerical Attrrbute Reduction Based on Neighborhood Granulation and Rough Approximation[J]. Journal of Software, 2008, 19(3): 640-649

[13] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. 1991

[14] Meng Jun, Wang Xiu-kui, Wang Peng, et al. Knowledge Dependency and Rule Induction on Tolerance Rough Sets [J]. Journal of Multiple-Valued Logic and Soft Computing, 2013, 20(3/4): 401-421

[15] Orr S J, Morgan N M, Elliott J, et al. CD33 Responses are Blocked by SOCS3 through Accelerated Proteasomal-mediated Turnover[J]. Blood, 2007, 109(3): 1061-1068

[16] Mark P K, Comeils M, Sun X H, et al. A new Homeobox Gene Contributes the DNA Binding Domain of the t(1;19) Translocation Protein in pre-B ALL[J]. Cell, 1990, 60(4): 547-555

[17] Sicinska E, Aifantis I, Laurent L C, et al. Requirement for Cyclin D3 in Lymphocyte Development and T Cell Leukemias [J]. Cancer Cell, 2003, 4(6): 451-461

[18] Mertelsmann R, Steven G, Steinmann G, et al. T-cell Growth Factor (Interleukin 2) and Terminal Transferase Activity in Human Leukemias and Lymphoblastic Cell Lines [J]. Blut, 1981, 43(2): 99-103

[19] Min Fan, William Z. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48-67