文章编号:1001 - 9081(2011)08 - 02130 - 04
doi:10.3724/SP.J.1087.2011.02130

基于语义理解的中文博文倾向性分析

何凤英
(福州大学 数学与计算机科学学院,福州 350002)
(fengyhe@163.com)

摘　要:博客作为一种大众化的信息及文化载体被越来越多的人所接受,博客文本的情感倾向性分析也逐渐成为信息挖掘领域的热点。目前,文本倾向性分析的研究大都围绕普通文本、新闻评论进行,针对博客文本的特点,提出一种基于语义理解的博客文本倾向性分类方法。首先以 HowNet 情感词语集为基准,构建中文基础情感词典,并用中文词语相似度方法计算词语的情感权值,同时分析语义层副词的出现规律及其对文本倾向性判断的影响,最后利用博主的语言风格因素对倾向性结果进行修正实现博文的情感分类。实验表明,该方法能有效地判定博客文本情感倾向性。

关键词:倾向性分析;语义理解;情感词典;博主语言风格;副词
中图分类号: TP391.1　　**文献标志码**:A

Orientation analysis for Chinese blog text based on semantic comprehension

HE Feng-ying
(*College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian 350002, China*)

Abstract: Blog has been accepted by more and more people as a popular information and cultural carrier. Orientation analysis for blog text also has become a hot spot in the field of information mining. The previous researches of text orientation mainly focus on plain text or news comments. A method of orientation analysis for blog text based on semantic comprehension was proposed according to the characteristics of blog text. Firstly, a Chinese basic emotional lexicon dictionary based on the HowNet emotional word set was constructed and the emotional value of Chinese emotional words was calculated on the basis of the similarity of Chinese words. Then, the adverbs and its influence on identification of text orientation in the semantic level were analyzed. Finally, the results were amended by using bloggers' language style factors and then the sentimental classification for blog text was realized. The experimental results show that the proposed method can effectively judge the blog text sentimental preference.

Key words: orientation analysis; semantic comprehension; sentiment lexicon; blogger language style; adverb

0　引言

近年来,随着互联网的快速发展,越来越多的用户开通并使用博客,博客已成为互联网上一种重要的知识交互工具。统计表明[1],截止 2008 年底,在中国 2.98 亿网民中,拥有博客的网民比例达到 54.3%,用户规模为 1.62 亿人,其中64.8%的博客为最近半年内进行过更新的活跃博客。人们利用博客交流思想、共享资源,博客之间通过互相引用,互相推荐形成一个巨大的博客空间。在博客空间中,人们可以自由发表对现实生活各种问题的观点,表达自己的情感,讨论热点事件等。由于博客空间信息丰富,对博客中的信息进行倾向性分析,准确检索出博客空间中人们对重要话题、热点事件的观点看法,对市场调研、网络舆情发现与预警等应用有重要意义。

目前,国内外对于文本倾向性的研究大都围绕普通文本或是新闻评论进行,主要分为两大类。

1)基于语义的文本倾向性研究方法。主要是由已有的电子词典或词语知识库扩展生成情感倾向词典,如朱嫣岚等人[2]利用 HowNet 提供的语义相似度和语义相关场的计算功能,计算待估词与预先选好的褒贬基准词对组的相关性,从而得到该词的倾向性;或是预先建立一个倾向性语义模式库,如 Yi 等人[3]就使用一个倾向性词汇表和一个倾向性模式库来对抽取出来的句子和短语进行语义关系分析,进而得到产品评论的文本倾向性;刘永丹等人[4]则将已有的语义分析技术用于倾向性判断,用精简的格语法和语义框架表达文本中的语义关系并进行倾向性分析。

2)基于机器学习的传统文本分类技术。如 Pang 等人[5]分别使用朴素贝叶斯、最大熵及支持向量机等方法进行文本倾向性研究。徐琳宏等人[6]选取褒贬倾向性比较强烈的词作为特征项,构造了一个支持向量机(Support Vector Machine, SVM)褒贬两类分类器来进行文本倾向性分析。但是和普通文本或新闻评论相比,博文有其自身的特点,博文内容通常是博主的自我表达,它是私有的,代表博主的兴趣、观点以及博主与他人的互动情况,博主的个性风格在很大程度上影响着博文倾向性强度,并且,博文的内容往往都很短小精练,观点性明确,情感词汇运用广泛,更新频率也更高。

本文在众多研究的基础上,针对博文的特点,提出了一种基于语义的博文倾向性分析的研究框架,实验结果表明,该系统可以有效地分析博文内容的倾向性。

1　语义理解的博客文本倾向性分析框架设计

基于语义理解的博客文本倾向性分析方法首先对博文进行预处理,得到标注词性的文本,然后对文本分句,进行句子

收稿日期:2011 - 02 - 18;**修回日期**:2011 - 04 - 14。　**基金项目**:福建省教育厅科研资助项目(JB09003)。
作者简介:何凤英(1979 -),女,福建武平人,讲师,硕士,主要研究方向:数据网格、分布计算。

的修辞分析,并建立情感词词典,利用词典及程度副词和否定副词词典计算情感词的极性,最后通过计算文档句子情感值之和获取博文的情感倾向性并对其结果给予统计分析。基于上述过程的博文倾向性分析系统框架在功能上主要由文本预处理、文本倾向性分析以及图形统计分析三个模块组成,如图1所示。

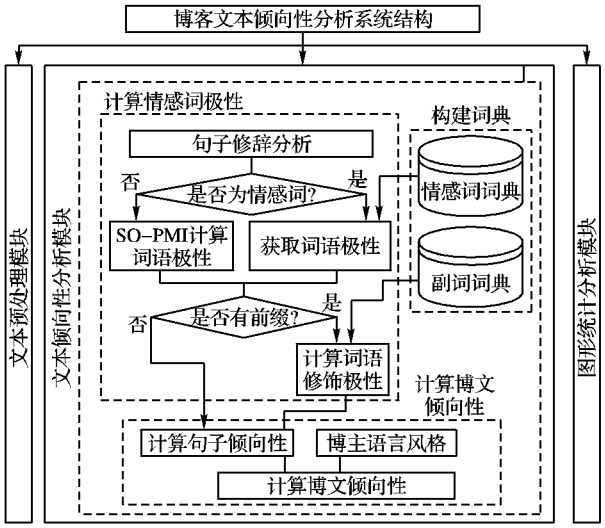

图1　系统框架设计

1)文本预处理模块。

文本预处理模块的主要功能是对输入的待分析文档进行断句、分词以及 pos 标注。本文下载使用中科院计算技术研究所研制的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS),实现文本的中文分词及词性标注工作,并以句子为单位进行存储。

2)文本倾向性分析模块。

文本倾向性分析模块是框架的核心模块,它的主要功能是根据文本所表达内容的情感倾向性将文本分成正面和负面两类文档。

本模块的基本处理单元是每一个单句,在对句子分词的基础上逐词检索情感词词典获取其极性和强度,若不在情感词词典中,则用 SO-PMI(Semantic Orientation-Pointwise Mutual Information)[7]公式计算极性,然后对句子进行修辞分析,判断其前缀是否出现程度副词或否定词,若有则根据副词词典计算情感词的修饰极性,然后统计每个句子中包含的情感词语及情感短语个数,并将其倾向值求和获取句子的极性,最后累加句子的倾向值,同时考虑博主语言风格因素的影响得到博文的情感倾向值。

3)图形统计分析模块。

图形统计分析模块的功能是根据文本倾向性分析模块的结果进行统计,提供交互信息,并显示最终的分析结果。

2　基于语义理解的博客文本情感分类方法

2.1　情感词典的创建

目前,国内最权威的情感词词库是 HowNet 提供的"情感分析用词语集",但缺少情感强度,尚需完善。本文以 HowNet 发布的情感词语集为基础,通过人工挑选,去掉一些非常用及情感倾向性不明显的词语,得到6196个情感词,然后采用词义相似度方法计算出每个词的情感倾向权值,去掉分类不正确的词以及权值过低的中性词,最后得到5281个基础情感词,构成一个基础情感词词典,处理过程包括以下4个方面。

1)选择基础情感词。

本文使用的基础情感词以 HowNet 发布的情感词语集为基础,通过人工挑选去掉一些不太常用或者情感倾向不很明显的词语,如:"假傥"、"云谲波诡"、"无可訾议"等,共650个。最终得到6196个基础情感词,其中褒义词3219个,贬义词2905个。

2)选择种子词。

将产生的基础情感词按 Google 搜索返回的 Hits 数进行排序,选择 Hits 数最高的词为种子词。在种子词数量的确定上,文献[8]研究表明,种子词数量为总情感词数量的15%左右时情感倾向性判断准确率达到峰值90%左右,并且在达到峰值后准确率都趋向稳定。因此,本文共确定出了种子词900个,其中正面评价种子词493个,负面评价种子词407个。选取部分正面种子词有:漂亮、完美、健康、真实、时尚、喜欢、快乐、流行、不错、精彩、方便、开心、稳定、感谢、积极、丰富、优秀、满意、精选、正确、美丽、轻松、文明、新鲜、聪明、舒适、热情、高档、和平、可靠、良好、顺利、干净、一流、美味、甜蜜、优雅、幸运、喜爱、齐全。选取部分负面种子词有:差、错误、恐怖、旧、辛苦、黄色、无聊、孤独、凄凉、恶劣、郁闷、黑暗、非法、愤怒、可怕、不利、烂、错误、背后、麻烦、讨厌、可怜、盲目、抱怨、失礼、脏、艰难、变态、恶心、虚假、不幸、否定、混乱、废、疯狂、毒、困难、弱、无奈、臭。

3)计算基础情感词的情感倾向权值。

本文采用 HowNet 的语义相似度计算公式计算两词语之间的相似度。HowNet 中词语相似度的计算以词的义原为基础。对于两个中文词语 w_1、w_2,假设它们分别有多个义原,w_1 有 $p_{11},\cdots,p_{1n}$ 个义原,w_2 有 $p_{21},\cdots,p_{2m}$ 个义原,则词语 w_1 和 w_2 的相似度计算公式如式(1)所示:

$$sim(w_1,w_2) = \max(sim(p_{1i},p_{2j})); 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m \tag{1}$$

其中:$sim(p_{1i},p_{2j}) = \dfrac{a}{d+a}$ 为两个义原的语义距离,d 为两个义原在知网中的路径距离,a 是一个可调节的参数。

而情感词 w 的情感权值大小由这个词与种子集中每个词的语义倾向相似程度来决定。假设种子集为 $seedset = \{PP, PN\}$,其中:PP 指褒义种子词集,PN 为贬义种子词集,则词 w 的情感权值定义如下:

$$o(w) = \frac{1}{M}\sum_{i=1}^{M} sim(w,pp_i) - \frac{1}{N}\sum_{j=1}^{N} sim(w,pn_j) \tag{2}$$

其中:$pp_i \in PP$,$pn_j \in PN$,M 和 N 分别为褒义种子集和贬义种子集中种子词的个数。设置0为阈值,$o(w) > 0$ 表明该词为褒义,$o(w) < 0$ 表明该词为贬义,$o(w)$ 数值大小则代表该词 w 褒贬倾向强度。

4)建立基础情感词词典。

根据约束域原理,情感词的极性最好约束在一个闭区间内,便于量化分析。Sentiwordnet[9]就是把情感词的极性定义在[-1,+1]。[-1,+1]这样一个以零为中心的对称区间,能很好地识别情感词的情感倾向以及极性强度。"0"代表中立情感,正区间代表褒义倾向,负区间代表贬义倾向,绝对值越大,则情感强度越强。本文使用线性的方法,对情感权值进

行了重新的规划。具体的计算公式如式(3)所示：

$$d' = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{3}$$

其中：d' 是重新规划后的情感权值，d 是根据式(2)计算出来的情感权值，$d_{\min}$ 是表示根据式(2)计算出来的所有情感权值中的最小值，$d_{\max}$ 是最大值。

将所有词语的情感权值按式(3)计算处理后，去掉分类不正确的词语和一些对文本的分类没有太大帮助的偏中性词语，如“甘愿0.038232 6”、“凝神0.047556”、“回礼0.044343 1”等，最后得到的情感词典包含正面情感词语2 873个，负面情感词语2 408个。表1为词典中部分词语的情感权值。

表1 部分情感词语权值

正面评价词语	正面词所占权值	负面评价词语	负面词所占权值
漂亮	0.988 285	乱	-0.981 231
优越	0.918 977	冷淡	-0.927 579
完美	0.913 708	粗鲁	-0.864 193
一流	0.830 725	暗	-0.846 732
精致	0.824 964	肮脏	-0.826 600
快乐	0.821 865	傲慢	-0.796 218
舒服	0.819 457	恶劣	-0.700 102
优雅	0.813 103	糟糕	-0.693 954
温馨	0.795 786	差劲	-0.677 165
辉煌	0.881 735	讨厌	-0.858 167

2.2 情感词极性的判断

情感词极性的判断，本文采用上海交通大学姚天昉等人[10]提出的算法，该算法混合了Hatzivassiloglou等人[11]提出的方法和Turney等人[12]提出的完全基于统计学的方法。具体过程如下：

1)对于每一个候选情感词，首先查找情感词字典。如果存在，则获取其极性和强度；

2)如果未查找到，则分别向前和向后查找情感词，并分别找到与前后情感词之间的关联词；

3)如果没有关联词出现，则利用SO-PMI公式计算候选情感词的极性；

4)如果该候选情感词与其前面或后面的情感词之间出现了关联词，首先判断关联词的类型，然后根据关联词类型计算其极性及强度。

实验表明，该方法的查全率和查准率均比前两种独立方法要好，分别达到了88.5%和94.4%，因此，本文采用这种方法来计算情感词的原极性。

但是，本文发现在博客中有大量的副词修饰情感词汇，比如“更为重要”、“很不公平”。“更为”修饰“重要”，表达的情感明显比“重要”要强烈；“不”修饰“公平”，表达了否定的意向，用“很”修饰“不公平”，表达的情感明显比“不公平”强烈。这些否定词和程度副词通常会改变情感词原极性的方向或强度，为了更好地分析情感词的语义倾向，必须计算情感词的修饰极性，为此，本文首先定义两个副词字典。1)否定词字典。本文在HowNet中选取有否定意义的义原，并从中抽取包含否定义原的概念，经人工过滤得到“不”、“不是”、“没有”等18个否定词。2)强调词字典。本文参考文献[13]的程度副词分类表，从中抽取了最常用的一些程度副词，并分别赋予不同的强度，具体设置如表2所列。然后从句子的修辞结构分析入手，参考文献[14]定义如下规则计算情感词的修饰极性。为了便于描述，本文用 N 代表否定词，D 代表程度副词，W 代表情感词，$O(PP)$ 为情感词的原极性，$O(MP)$ 为情感词的修饰极性，$I(D)$ 为程度副词 D 的强度值。

1) 短语 W：$O(MP) = O(PP)$。

2) 短语 $N + W$：$O(MP) = -2/3 * O(PP)$。

3) 短语 $D + W$：若 $I(D) < 0$，则 $O(MP) = (1 + I(D)) * O(PP)$；若 $I(D) > 0$ 同时 $O(PP) > 0$，则 $O(MP) = O(PP) + (1 - O(PP)) * I(D)$；若 $I(D) > 0$ 同时 $O(PP) < 0$，则 $O(MP) = O(PP) - (1 + O(PP)) * I(D)$。

4) 短语 $N + N + W$：$O(MP) = O(PP)$。

5) 短语 $N + D + W$：$O(MP) = -1/2 * O(PP)$。

6) 短语 $D + N + W$：若 $I(D) < 0$，则 $O(MP) = -(1 + I(D)) * O(PP)$；若 $I(D) > 0$ 同时 $O(PP) > 0$，则 $O(MP) = -O(PP) - (1 - O(PP)) * I(D)$；若 $I(D) > 0$ 同时 $O(PP) < 0$，则 $O(MP) = -O(PP) + (1 - O(PP)) * I(D)$。

7) 短语 $D + D + W$：按规则3)递归计算两次得出 $O(MP)$ 的值。

下面用几个句子来验证该算法的可行性。

首先，查得“漂亮”的原极性 $O(PP) = 0.988\,285$。

“她长得很漂亮”：$O(\text{“很漂亮”}) = O(\text{“漂亮”}) + (1 - O(\text{“漂亮”})) * I(\text{“很”}) = 0.995$。

“她长得有点漂亮”：$O(\text{“有点漂亮”}) = (1 + I(\text{“有点”})) * O(\text{“漂亮”}) = 0.494$。

“她长得不很漂亮”：$O(\text{“不很漂亮”}) = -1/2 * O(\text{“漂亮”}) = -0.494$。

“她长得不漂亮”：$O(\text{“不漂亮”}) = -2/3 * O(\text{“漂亮”}) = -0.659$。

“她长得很不漂亮”：$O(\text{“很不漂亮”}) = -O(\text{“漂亮”}) - (1 - O(\text{“漂亮”})) * I(\text{“很”}) = -0.995$。

由此可见，根据这种算法利用程度副词强度表计算得出的情感词极性基本合理，且都约束到域[-1,1]，便于情感词的比较。

表2 常用程度副词强度表

程度副词 D	强度值	强度增减
最	1.0	增强
极其	0.9	
非常、太、真、实在、完全	0.8	
相当、很、粉、多、得多	0.6	
挺、蛮、一点也、一点都	0.4	
比……更	0.2	
稍微、小小的	-0.7	减弱
点、有点、有些、偏	-0.5	
较、比较、较为、还、还算	-0.3	

2.3 博文倾向性的判断

博文倾向性的计算以每个句子为单元，利用情感词典查找句中带有感情色彩的词语 w，并记录其倾向值 $o(w)$，如果情感词前面出现过程度副词或否定词，则按2.2节定义的规则计算其倾向值 $o(w)$，最后对所有情感词的倾向值进行累加得到博文的情感倾向值。假设博文分成 n 个句子 $sen_1, sen_2, \cdots, sen_n$，而句子 sen_m 中包括 k 个情感词，记为 $w_{m1}, w_{m2}, \cdots,$

w_{mk},则整篇博文的倾向性如式(4) 所示:

$$D = \sum_{i=1}^{n}\sum_{j=1}^{k} o(w_{ij}) \tag{4}$$

但是,博客是博主表达自己观点情感的媒介,博主的语言习惯、个性风格能够很大程度上影响着倾向性强度。例如,乐观的博主往往用褒义程度比较强的倾向词来赞美某一事件(比如,"很好,非常棒"等),而悲观的博主则往往用褒贬程度比较弱的倾向词(比如,"一般,还可以")。因此,同样一个倾向词对于不同的博主所表达的倾向性强弱是不一样的,如果忽略博主因素,而仅仅使用情感词的语义极性,会给博文的倾向值带来偏差。文献[15] 针对博主书写博文的倾向性用语风格因素建立博主背景模型,认为博主所有博文的倾向性评分符合"平均值 $\mu = \bar{\chi}$ 和方差 $\sigma = \sigma_i$" 的正态分布,且有:

$$\bar{\chi} = \frac{1}{n}\sum_{i=1}^{n}\chi_i \tag{5}$$

$$\sigma_i = \left[\frac{1}{n-1}\sum_{i=1}^{n}(\chi_i - \bar{\chi_i})^2\right]^{\frac{1}{2}} \tag{6}$$

其中:χ_i 是单篇博文的倾向性评分;$\bar{\chi}$ 是所有博文倾向性评分的平均值,往往反映博主整体倾向性用语的强弱;而 σ_i 是博主所有博文倾向性评分的方差,反映博主书写博文风格的稳定性。

一般而言,激进型博主的倾向性得分往往偏高,μ 值比较大,保守型博主倾向性得分往往偏低,μ 值较小,可利用平均值 μ 对倾向性进行平滑;并且风格稳定的博主其方差 σ 较小,较为可信,而风格不稳定的博主其方差 σ 较大,较为不可信。为了能更合理地度量博文倾向性强度,本文对式(4) 修正如下:

$$D = (D - \mu)/\sigma \tag{7}$$

其中 D 为最终的博文倾向度。当 $D > 0$ 时,文章为褒义;当 $D < 0$ 时,文章为贬义,当倾向值在零值附近时表示博文接近中立。

3　实验结果及分析

实验所用的博客集合是从网站 http://blog.sina.com.cn 中下载的。人工从中找出对同一主题具有兴趣的所有博客的全部博文,由于每个博客都可以对不止一个主题感兴趣,本文对所有博文认真审查,去除语言不规范的文本,最终选出具有相同主题的博文 300 篇进行实验。与手工标注的结果进行比较,采用召回率和准确率两个指标评价实验结果,同时,将本文方法和文献[16] 基于 SVM 的分类方法以及文献[6] 基于 HowNet 的语义相似度的计算方法进行比较。实验结果如表 3 ~4所示。

表 3　本文方法的博文分类结果

极性	本文方法识别篇数			手工标注篇数	准确率/%	召回率/%
	总篇数	正确篇数	错误篇数			
褒义	153	146	7	159	95.42	91.82
贬义	147	125	22	141	85.00	88.65

通过表 3 可以看出,本文方法对褒义的博文取得较好的效果,而对极性为贬义的识别则较差,主要因为没有考虑到词语的动态极性。有些词语通常不表现为极性词,只有在修饰某个特征时才含有褒贬意义,如:"大"不表现为极性词,在修饰"功耗"时表现为贬义,但在修饰"功劳"时表现为褒义,在修饰"数字"时表现为中性,这需要建立动态极性词表解决。

从表 4 可以看出不同分类方法用于篇章的分类效果也不同。本文的分类方法先通过极性词典判断极性,若无法判断再进行语义相似度计算,同时利用博主的语言风格因素对倾向值进行修正,与文献[6]方法相比,提高了判断的精度和效率。文献[16] 是基于 SVM 的分类方法,需要大量的训练数据,而篇章级别的博文有利于特征的提取,所以精度略高于本文方法。

表 4　博文的极性分类准确率比较

分类方法	正确识别数目	准确率/%
文献[16]方法	277	92.33
文献[6]方法	254	84.67
本文方法	271	90.33

4　结语

中文博客文本的情感分析是建立在 Web 信息挖掘、信息抽取基础上的一项非常有意义的研究。本文提出了基于语义的博客文本倾向性分析方法。以 HowNet 情感词语集为基础构建中文基础情感词词典,并计算词语的情感倾向权值,同时分析了程度副词和否定副词在语义上对情感词倾向性的影响,提出计算词语修饰极性的方法,最后分析博主背景因素对博文倾向性的影响并建立博主的倾向性用语风格数学模型,对博文倾向性判定进行修正。实验表明,本文提出的解决方案有较好的性能,具有一定的实用价值。但是,博文倾向性还受文档结构等诸多因素影响,还需要通过中心句中心段的识别来提高文档级的分类精度,并且本文研究的博文都是基于同一个主题,如何进行主题的自动识别、抽取还有待进一步更深入细致的研究。

参考文献:

[1] China Internet Network Information Center. The 23th statistical report of China Internet network development [EB/OL]. [2011 - 01 - 10]. http://www.cnnic.net.cn/uploadfiles/pdf/2009/1/13/92458.pdf.

[2] 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.

[3] YI J, NASUKAWA T, BUNESCU R, *et al*. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques [C]// Proceedings of the 3rd IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2003: 427 - 434.

[4] 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.

[5] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002: 79 - 86.

[6] 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.

[7] TURNEY P D, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315 - 346.

(下转第 2137 页)

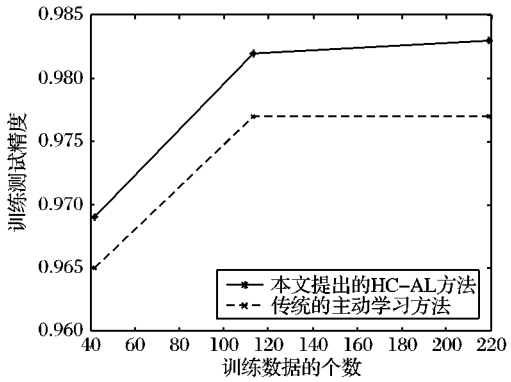

图 5　两种方法在 Thyroid 上的实验结果

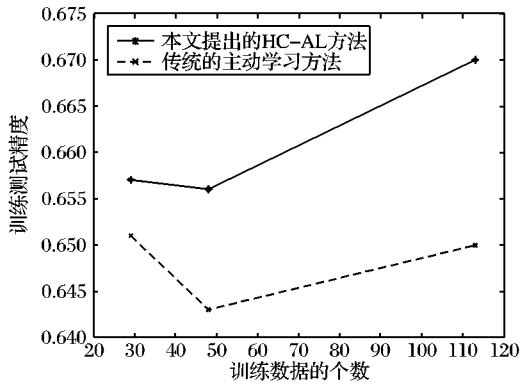

图 6　两种方法在 Flare_solar 上的实验结果

3　结语

主动学习方法由于其训练规模小、速度快的优点,在很多领域得到成功的应用,已经成为机器学习领域的一个研究热点。本文提出的 HC_AL 方法采用分层细化、逐步求精的方法,既提高学习器的学习效率,又获得满意的泛化能力。在未来的工作中,可以考虑将本文提出的 HC_AL 主动学习方法与不同的分类器如神经网络、决策树等相结合,以取得更令人满意的结果,从而使主动学习的性能得到进一步的提高。

参考文献:

[1] DIMA C, HEBERT M, STENTZ A. Enabling learning from large datasets: Applying active learning to mobile robotics [C]// ICRA 2004: Proceedings of the 2004 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2004: 108 - 114.

[2] VLACHOS A. A stopping criterion for active learning [J]. Computer Speech and Language, 2008, 22(3): 295 - 312.

[3] 张健沛,徐华. 支持向量机(SVM)主动学习方法研究与应用[J]. 计算机应用,2004,24(1):1 - 3.

[4] CORD M, COSSELIN P H, PHILIPP-FOLIGUET S. Stochastic exploration and active learning for image retrieval [J]. Image and Vision Computing, 2007, 25(1): 14 - 23.

[5] 田春娜,高新波,李洁. 基于嵌入式 Bootstrap 的主动学习示例选择方法[J]. 计算机研究与发展,2006,43(10):1706 - 1712.

[6] 韩光,赵春霞,胡雪蕾. 一种新的 SVM 主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发展,2009,46(11):1934 - 1941.

[7] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2002, 2(1): 45 - 66.

[8] ABE N, MAMITSUKA H. Query learning strategies using boosting and bagging [C]// Proceedings of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 1998: 1 - 9.

[9] BAEZA-YATES R, HURTADO C, MENDOZA M. Query clustering for boosting Web page ranking [C]// AWIC 2004: Proceedings of the Second International Atlantic Web Intelligence Conference, LNCS 3034. Berlin: Springer-Verlag, 2004, 3034: 164 - 175.

[10] FINE S, GILAD-BACHRACH R, SHAMIR E. Query by committee, linear separation and random walks [J]. Theoretical Computer Science, 2002, 284(1): 25 - 51.

[11] TONG JUN, YIN JIANPING, ZHU EN. An active learning method based on most possible misclassification sampling using committee [C]// MDAI'07: Proceedings of the 4th International Conference on Modeling Decisions for Artificial Intelligence. Berlin: Springer-Verlag, 2007: 104 - 113.

[12] ROVER B, JEM J R, ROSS D K. Active learning for regression based on query by committee [C]// IDEAL'07: Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 4881. Berlin: Springer-Verlag, 2007: 209 - 218.

[13] ZHU JINGBO, WANG HUIZHEN, TSOU B K, *et al*. Active learning with sampling by uncertainty and density for data annotations [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1323 - 1331.

[14] 陈锦禾,沈洁. 基于信息熵的主动学习半监督分类研究[J]. 计算机技术与发展,2010,20(2):110 - 113.

[15] 龙军,殷建平,祝恩,等. 选取最大可能预测错误样例的主动学习算法[J]. 计算机研究与发展,2008,45(3):472 - 478.

[16] YU H, YANG J, HAN J W, *et al*. Making SVMs scalable to large data sets using hierarchical cluster indexing [J]. Data Mining and Knowledge Discovery, 2005, 11(3): 100 - 128.

(上接第 2133 页)

[8] 柳位平,朱艳辉,栗春亮,等. 中文基础情感词词典构建方法研究[J]. 计算机应用,2009,29(10):2875 - 2877.

[9] ESULI A, SEBASTIANI F. Sentiwordnet: A publicly available lexical resource for opinion mining [C]// Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation. Genova, Italy: [s. n.], 2006: 417 - 422.

[10] 姚天昉,娄德成. 汉语情感词语义倾向判别的研究[C]//ICCC2007:第七届中文信息处理国际会议论文集. 北京:电子工业出版社,2007:221 - 225.

[11] HATZIVASSILOGLOU V, MCKEOWN K. Predicting the semantic orientation of adjectives [C]// ACL-97: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid, Spain: [s. n.], 1997:174 - 181.

[12] TURNEY P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// ACL-02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA: [s. n.], 2002: 417 - 424.

[13] 蔺璜,郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报:哲学社会科学版,2003,26(2):71 - 74.

[14] 林炜,林世平. 中文倾向性挖掘中情感词修饰极性的研究[J]. 计算机科学,2008,35(8):208 - 210.

[15] 廖祥文. 基于博主背景的博客倾向性检索归一化策略[J]. 中文信息学报,2010,24(3):75 - 80.

[16] 应伟,王正欧,安金龙. 一种基于改进的支持向量机的多类文本分类方法[J]. 计算机工程,2006,32(16):74 - 76.