一种基于样本空间动态划分的中文情感识别方法*

刘 智, 杨宗凯, 刘三妍, 铁 璐
(华中师范大学 国家数字化学习工程技术研究中心, 武汉 430079)

摘 要: 为了提高中文文本情感识别的精度,从集成学习的角度出发,提出了一种基于样本空间动态划分的机制构建文本情感分类器。该算法充分利用训练样本空间内的鉴别信息,通过引入核平滑方法对样本空间进行自适应划分,形成若干个具有差异性的多粒度样本子集,然后分别在每个子集上构造基分类器,最后将所有基分类器的输出进行融合以产生最终的预测结果。实验结果表明,该算法在查准率和查全率等方面均优于 Bagging、AdaBoost 等算法,并且在大规模样本集的情感识别中具有良好的应用前景。

关键词: 文本情感识别; 样本子空间; 动态划分; 集成分类模型; 核平滑
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)05-1443-05
doi:10.3969/j.issn.1001-3695.2013.05.040

Dynamic partition mechanism in sample space for sentiment recognition of Chinese texts

LIU Zhi, YANG Zong-kai, LIU San-ya, TIE Lu
(*National Engineering Research Center for E-Learning*, *Huazhong Normal University*, *Wuhan* 430079, *China*)

Abstract: To improve the performance of sentiment recognition for Chinese texts, this paper proposed a dynamic partition mechanism based on sample space to build the sentiment classifier in terms of ensemble learning method. Firstly, it utilized a kernel smoothing method to adaptively divide the sample space into several multi-granularity sample subsets. Then, it trained a base classifier on each subset by a learning algorithm. Finally, it combined the outputs from all base classifiers to form a final recognition result. To evaluate the method, the experiment was conducted on a Chinese benchmark dataset. The results indicate that the method is better than Bagging and AdaBoost algorithm in both precision and recall rate. Furthermore, it has a good application prospect in Chinese sentiment recognition for plenty of samples.

Key words: text sentiment recognition; sample subspace; dynamic partitioning; ensemble classification model; kernel smoothing

0 引言

随着 Web 2.0 技术的发展以及人们对信息需求的日益增长,越来越多的互联网用户通过博客、微博、论坛和在线评论网等公共社区发表个人对某事物的观点和意见。这些带有主观情绪的信息主要以文本的形式存储。在网络文本指数式增长的趋势下,舆论分析者和用户每天需要浏览大量的评论内容,并从中获取有价值的信息。蕴涵在这些文字中的观点、态度、情感等信息能为顾客、公司和政府部门提供重要的研究和调查资源。尽管如此,这些日益增长的网络评论也给普通用户造成了很多干扰,大量的噪声数据并不利于用户准确地判断事物的好坏。采取人工方式已经难以快速、稳定地挖掘潜在的重要信息,如何通过自动化的方式对这些带有情感信号的文本进行识别和分析,成为目前网络智能信息处理中的一个研究热点。因此,有效的情感识别方法对于网络舆情调查和分析具有重要的研究价值和实际意义。

1 相关工作

情感识别是模式识别中一个新兴的研究领域,其目的主要是将评论文的整体篇章、句子或单词分类为正面或负面的态度(有时也会出现中立态度)。其中,中文文本的情感识别已经受到了国内外学者的广泛关注。目前的研究工作主要分为基于情感知识的方法和基于机器学习的方法。在基于情感知识的研究中,Zhai 等人[1]提出一种识别并消除情感噪声词的方法,并验证了这种方法在使用不同规模情感词典时的识别效果;段建勇等人[2]提出一种基于句法语义的情感倾向判别算法,该算法基于特定领域的情感语料库,运用基于扩展句法树的语言处理模型,从句子到篇章识别文本情感倾向;杨江等人[3]从篇章结构的角度,计算篇章中存在的主题句的情感极性值,然后采取加权求和法对评论文本整体倾向性进行识别。在基于机器学习的研究中,Li 等人[4]将多分类模型应用于商品评论的情感识别中,该模型为不同领域的评论数据训练特定的分类器以克服领域依赖问题;胡熠等人[5]提出了一种基于

收稿日期: 2012-08-20; **修回日期**: 2012-10-08 **基金项目**: 国家"核高基"重大专项基金资助项目(2010ZX01045-001-005);国家"十二五"科技支撑计划资助项目(2011BAK08B03);国家教育部新世纪优秀人才支持计划资助项目(NCET-11-0654);华中师范大学中央高校基本科研业务费专项基金资助项目(CCNU09A02006)

作者简介: 刘智(1986-),男,湖北武汉人,博士研究生,主要研究方向为智能软件与知识服务、数据挖掘(liuzhi8673@gmail.com);杨宗凯(1963-),男,教授,博导,主要研究方向为现代信息网络、数字信息处理;刘三妍(1973-),男,教授,博士,主要研究方向为人工智能、计算机应用;铁璐(1988-),女,硕士研究生,主要研究方向为中文文本处理.

语言建模的方法识别文本情感,该方法在训练集中分别构建正面和负面两类情感倾向的语言模型,然后通过计算测试文本的语言模型与这两类情感模型之间的 Kullback-Leibler 距离以得到与其最相近的情感类别。Wang 等人[6]基于 Fisher 判别分析提出一种情感特征词的选择算法,并结合支持向量机(SVM)分类器对影视、手机、教育等多种评论语料进行情感分类,取得了较高的识别准确率。文献[7]对各种基于情感知识以及机器学习的方法作了对比和分析,得出的结论为机器学习方法在情感分类中能获得更好的识别效果,本文也将使用机器学习技术构造情感识别模型以识别评论文本的整体情感极性。

上述关于机器学习的研究主要基于情感特征空间构建分类模型,但随着网络中评论样本数的增加,情感特征数与样本数的比例将不断减小,这使得特征空间中可利用的情感鉴别信息变得更为有限。特别在中文文本中,情感表达较含蓄,某些含有情感信息的短文本中甚至不包含任何情感词,这使得特征空间变得更为稀疏。与上述研究相比,基于样本空间划分的集成学习模型仍然研究较少,实际上,采用有效的划分算法从样本空间中提取属于不同情感子区域的样本子集能为识别模型提供更为稳定和准确的情感判别信息。

本文提出一种动态划分算法,重点在于从训练集中划分出不同粒度的子空间以构造集成识别模型。充分利用样本空间中情感鉴别信息的分布,首先选取若干样本子空间作为初始种子集,对不同划分粒度产生的识别结果进行估计,为后续的动态划分提供依据;然后构造子空间划分粒度的概率密度分布,基于该分布自适应地选择多粒度的训练子空间;最后在这些子空间上构造相应的情感基分类器,并采用多数投票法结合这些基分类器的判别结果形成一个强分类模型以提高情感识别的性能。

2 基于样本子空间的集成识别方案

2.1 问题描述

为了构造向量空间模型,用于情感识别的训练和测试文本可表示为不同情感词组成的向量。设 $W=\{w_1,w_2,w_3,\cdots,w_d\}$ 为训练集中出现频率最高的 d 个情感词的降序集合,f_{ij} 为第 j 个情感词(w_j)在第 i 个文本中的频率,则每个样本 s_i 可表示为 $(f_{i1},f_{i2},f_{i3},\cdots,f_{id})$。此外,定义训练样本全空间为 $S=\{(s_i,c_i)\mid 1\leqslant i\leqslant n\}$,空间中任一样本 $s_i\in\Re^d$,其情感类别为 $c_i\in C=\{-1,+1\}$,-1 表示负面情感,$+1$ 表示正面情感。

设 $S^{r_k}_{k:n}$ 为 S 的第 k 个样本子空间,其中 $r_k(r_k<n)$ 为第 k 次迭代时自适应更新的子空间选取粒度。另设 $\Psi_k(S^{r_k}_{k:n})$ 为在子空间 $S^{r_k}_{k:n}$ 上训练生成的基分类器 Ψ_k,则基于 L 个样本子空间的集成分类模型可表示为

$$\mathrm{Ens}=\mathrm{RDS}\{\Psi_k(S^{r_k}_{k:n}),R,\mathrm{fus},1\leqslant k\leqslant L\} \tag{1}$$

其中:fus 为基分类器的融合策略。为了对每个子空间确定合适的粒度值,需要构造一种自适应更新的 R 分布函数,RDS 表示基于 R 分布的动态子空间(R-based dynamic subspace)方法,即本文所采用的关键集成分类算法。

2.2 总体方案

根据以上描述,在基于样本空间动态划分的集成识别方案中,需要进一步确定的问题包括情感词典的制定、情感特征选择方法、样本子空间数量(基分类器数量)的确定、子空间内样本子集选择方法、基分类器的类型以及融合策略的确定。对于这些问题,本文拟采取以下方案:

a)文本中包含的具有情感倾向信息的词汇是最好的表示该文本整体情感的特征,因此本文采用文献[8,9]中使用的情感评价词集作为情感词典。词典融合了清华大学[8]和 HowNet[10]发布的标准情感词集。其中包含褒义词 8 015 个,贬义词 6 733 个。

b)对于情感特征的选择,首先采用中科院分词工具 ICTCLAS2012[11]对文本集进行批量分词,并将分词结果进行存储,然后根据方案 a)中制定的情感词典提取分词结果中包含的情感词作为特征。

c)将样本子空间的数量 L 作为参数进行考量,分析不同的 L 值对情感识别结果的影响。

d)对于样本子集的选择,根据均匀分布 U 对样本设定等概率分布进行随机选择。对于子空间 S_k,选取步骤如下:

(a)在[0, 1]区间内产生一个服从分布 U 的随机数 u;

(b)如果 u 满足 $F_U(i-1)<u<F_U(i)\ (i\leqslant i\leqslant n)$,则选择第 i 个样本放入子空间内,其中 F_U 表示服从 U 分布的密度函数;

(c)返回(a),直到子空间 S_k 内 r_k 个样本全部选择结束。

e)基分类器采用基于线性核函数的支持向量机 LSVM[12](linear support vector machine),因其广泛地应用在文本分类中,并在稀疏特征空间上具有良好的分类性能和处理速度。LSVM 采用适合于大规模样本处理的 LIBLINEAR[13]实现。

f)采用多数投票法作为基分类器的融合策略

$$\mathrm{Senti}(\mathrm{Ens},x)=\arg\max_{c\in\{-1,+1\}}\{\sum_{k=1}^{L}\Psi_k(S^{r_k}_{k:n})\} \tag{2}$$

式中:基分类器 $\Psi_k\in\{0,1\}$,$\mathrm{Senti}(\mathrm{Ens},x)$ 表示对输入测试样本 x 采用集成分类模型 Ens 所得到的情感极性。

3 动态子空间算法

3.1 基于 R 分布的子空间选择

在传统的集成学习方法中,子空间的选取粒度需要预先设定,这使得在每个子空间内含有的样本数量为等粒度分布,基分类器之间差异度则无法保证。为了克服这个缺陷,本文引入动态子空间模型。该模型采用迭代法构造每个子空间,其具体方法是根据不同子区域上所构造的分类器训练性能自适应地为其选取合适的粒度大小。为此,需要构造 R 分布为子空间的粒度选择提供概率依据。

R 分布的构造过程包括两个步骤:a)构建 R 分布的初始分布序列 R_0;b)采用核平滑 kernel smoothing (KS)[14]方法以平滑 R_0 使其连续化。该过程如图 1 所示。

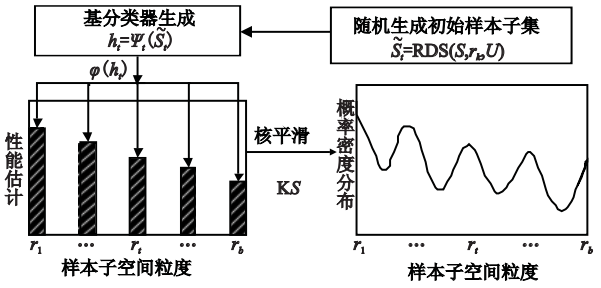


图1 初始分布R_0的概率密度估计示意图

如图 1 中左边的连续分布图所示,为了构建 R_0 分布,首先需要产生 b 个训练子集,其产生机制服从均匀分布 U,即 $\tilde{S}_t=\mathrm{RDS}(S,r_t,U)\ (t=1,2,\cdots,b)$,其中,初始样本子空间的粒度 r_t 根据式(3)确定,其中 floor 表示取整计算:

$$r_t=b+\mathrm{floor}(\frac{1}{2}\times\frac{n-1}{t-1}\sum_{s=1}^{t-1}h_s) \tag{3}$$

然后,在 b 个训练子集上构造初始基分类器 $h_t(t=1,2,$

…,b),并估计每个基分类器 h_t 在初始子空间 $\tilde{S}_t$ 的训练精度 $\varphi(h_t)$。该训练精度代表选择粒度为 r_t 时的训练效果。根据这 b 个训练精度值得到离散的 R_0 分布,最后使用 KS 方法将其连续化处理,KS 平滑过程如式(4)所示:

$$f_R(r)=\frac{1}{\sum_{t=1}^{b}\varphi(h_t)\sigma}\left[\sum_{t=1}^{b}\varphi(h_t)K\left(\frac{r-r_t}{\sigma}\right)\right]\quad r=1,2,\cdots,n \tag{4}$$

其中:K 为核函数,σ 为平滑因子。

在本研究中,将以 R 分布为依据,采用伪随机数产生的逆方法以决定子空间的粒度。为此,设 $F_R(r)$ 为累计概率密度值,定义如下:

$$F_R(r)=\sum_{t=1}^{r}f_R(t)\quad t=1,2,\cdots,r \tag{5}$$

子空间选取步骤如下所示:

a)产生一个在[0,1]区间上服从均匀分布的随机数 v;

b)每次迭代中,累计概率密度值 $F_R(r)$,如果存在 $F_R(r-1)<u<F_R(r)(1\leqslant r\leqslant n)$,选择 r 为该子空间的粒度。

在子空间形成的过程中,R 分布将会不断被更新,这样避免了大量样本被重复选择的情况,并且不同粒度的子空间内情感鉴别信息会分布得更加均衡,为后续基分类器的构造提供了分类互补信息,提高了集成系统的有效性。

3.2　算法设计

RDS 算法从两方面着手:a)通过子空间选取粒度的差异性提高基分类器之间的互补能力;b)通过 R 分布的更新机制获取最合适的子空间选择区域。

该算法采取一种级联的方式对基分类器进行构造,第 k 层子空间的选取参数将根据第 $k-1$ 层基分类器的训练精度作相应的调整,以形成一种自适应的选取过程。生成序列号相邻的基分类器之间具有较强的互补能力,这样训练出的基分类器充分利用了原始样本全空间内不同情感子区域的情感信息,特别是在短文本中情感信息缺损的情况下,动态子空间选择能自适应地选择出含有更多情感特征的样本,并与其结合形成包含完整情感信息的子空间。最后融合所有基分类器的决策结果,对测试样本作出稳定的情感极性判别。RDS 算法流程如下所示:

输入:训练样本集 S,测试样本 x,一种学习算法(分类器)Ψ,基分类器的数量 L,基于 R 分布的动态子空间选择算法 RDS。

输出:基分类器的预测值 $H=\{h'_1,h'_2,\cdots,h'_L\}$ 以及测试样本 x 情感极性 Senti:$x\rightarrow c\in\{-1,+1\}$,$-1$ 为负面情感,$+1$ 为正面情感。

a)训练阶段

(a)随机生成 b 个训练子集 $\tilde{S}_t\in S(1\leqslant t\leqslant b)$,估计初始分布 R_0;

(b)根据 R_{k-1} 分布函数的累计密度值 $F_R(r)(r=1,2,\cdots,n)$ 获取一个新子空间的采样粒度 r'_k;

(c)由 RDS 选择机制以及(b)获取的粒度值 r'_k,迭代抽取子空间 $\tilde{S}'_k=\mathrm{RDS}(S,r'_k,R)(k=1,2,\cdots,l)$;

(d)使用学习算法 Ψ 通过 $h_k=\Psi_k(\tilde{S}'_k)$ 训练第 k 个基分类器;

(e)在样本子集 $\tilde{S}'_k$ 上估计第 k 个基分类器的训练精度 $\varphi(h'_k)$;

(f)将 $\varphi(h'_k)$ 作为反馈参数以更新第 k 次迭代的密度分布 R_k,更新机制如式(6)所示:

$$f_R(r)=\frac{1}{\left(\sum_{i=1}^{b}\varphi(h_i)+\sum_{j=1}^{k}\varphi(h'_j)\right)\sigma}\times\left[\sum_{i=1}^{b}\varphi(h_i)K\left(\frac{r-r_i}{\sigma}\right)+\sum_{j=1}^{k}\varphi(h'_j)K\left(\frac{r-(r'_j)}{\sigma}\right)\right],r=1,2,\cdots,n \tag{6}$$

(g)返回(b),直到 L 个基分类器全部训练结束。

b) 识别阶段

(a)应用构造的 L 个基分类器分别对 x 进行分类;

(b)统计基分类器的识别结果 $h'_k(k=1,2,\cdots,L)$,最后采用多数投票法的融合策略得到 x 情感极性,如式(7)所示:

$$\mathrm{Senti}(x)=\arg\max_{c\in\{-1,+1\}}\mathrm{card}(k\mid h'_k(x)=c) \tag{7}$$

3.3　算法分析

在进行 RDS 算法的多分类器融合时,核函数 K 及其带宽参数 σ 的选取对最终的识别性能有较大的影响。σ 越大,密度分布函数 R_k 的曲线抖动越显著;σ 越小,分布曲线越平缓,σ 过小会使 R_k 逼近于均匀分布,则选取的子空间覆盖区域和粒度会出现重叠的情况,集成分类模型的差异度和置信度会大大降低。为此,在 R 分布的更新中,如式(8)(9)所示,核函数设置为高斯分布函数,而 σ 主要由基分类器的训练精度来决定,以保持概率密度函数 $f_R(r)$ 的平滑性:

$$K\left(\frac{r}{\sigma}\right)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2} \tag{8}$$

$$\sigma=0.9\,\mathrm{std}(H)\times[\mathrm{mean}(H)]^{-\frac{1}{5}} \tag{9}$$

其中:H 表示前 k 个基分类器的训练精度组成的向量,$H=\{h'_1,h'_2,\cdots,h'_k\}$,std 与 mean 分别表示向量 H 中训练精度的标准差与均值。

为了分析该集成系统的复杂度,设算法总体时间复杂度为 $T(p)$,$T_k(p)$ 表示第 k 个基分类器的时间复杂度,$T_s(p)$ 为动态子空间选择所需的时间复杂度,$g(p)$ 为子空间采样规模 p 的某个函数,则算法的总体时间复杂度为

$$T(p)=T_s(p)+\sum_{k=1}^{L}T_k(p) \tag{10}$$

考虑到算法是利用基分类器的训练结果进行的自适应迭代处理,从算法流程上看,随着问题规模 p 的增大,算法运行时间的增长率与问题规模 $g(p)$ 的增长率成正比,因此 $T_s(p)$ 是呈线性阶。如果忽略子空间生成过程的时间复杂度,$T_s(p)$ 可转换为

$$T_s(p)=O(g(p))=O(p) \tag{11}$$

可以看出,总体时间复杂度 $T(p)$ 与采用的基分类器算法密切相关。而对基分类器的训练和识别过程可以采用成熟的并行处理软件快速完成,最后只需将所有并行输出结果进行简单投票统计即可获得识别结果。通过算法的时间复杂度分析不难看出,不考虑子空间选择过程中线性阶的时间复杂度 $T_s(p)$,该集成分类算法具有良好的可行性。

4　实验研究

4.1　数据集

本文使用的实验数据来自于谭松波[15]提供的关于宾馆和笔记本的中文评论语料库,该语料集为平衡数据集,正负类各 2 000 篇。数据集的预处理包括三部分:a)去除与评论主题无关的字符,包括 HTML 代码以及一些引导类文字如“免费注册”“宾馆反馈”“补充点评”等;b)数据规范化处理,包括繁体字转换为简体字、半角全角符号转换等;c)采用 2.2 节的方案 a)中制定的情感词典,从两种语料集中统计所有情感特征词出现的频率。经过预处理后,数据集的相关信息如表 1 所示。

表 1　实验数据集信息

语料	平均文本长度	平均单词数	平均情感词数	情感词总数
宾馆	50.58 字符	71.80	8.56	2 099
笔记本	31.61 字符	33.61	4.92	1 185

从表 1 中可以看出,评论文本的长度普遍较短,且平均每个文本所含有的情感特征词较少,在文本中所占比例均低于 15% 。另外,在两类数据集中情感词与样本规模比例分别为 52.48% 和 29.63% ,该信息粒度表明基于样本空间的动态划分方法具有可行性。

4.2 实验设计

为了验证 RDS 算法的识别性能以及在中文情感识别中的精度和鲁棒性,具体实验设计方案如下:

a)考察子空间的采样数量(即基分类器数量)L 以及初始分布 R_0 中训练子集产生的数量 b 对 RDS 算法识别性能的影响。L 分别取值为 20, 40, 80, 100, 120;b 的取值分别为 5, 10, 20, 30, 40。

b)对比 RDS 与集成算法 Bagging、AdaBoost 以及单分类器 LSVM 的识别性能,考察四种方法在标准数据集上的识别效果。其中,LSVM 是基于原始样本空间进行分类,Bagging 算法[16]是对样本空间进行等粒度随机划分,AdaBoost 算法[17]是在 Bagging 基础上对样本权重作动态调整以迭代方式形成等粒度的子空间。实验中所有集成方法均采用 LSVM 作为底层基分类器。三种集成识别算法分别使用不同的基分类器数量进行实验(L = 20, 40, 120)。另外,Bagging 和 AdaBoost 算法中子空间的选取粒度均设置为原始训练样本规模的 50% ,RDS 中初始训练子集 b 设置为 30。

c)在底层基分类器 LSVM 中,线性核函数的损失参数 C 设置为 0.08,代价参数 E 设为 0.01。

d)在以上两种实验中,训练集与测试集按照 4∶1 的比例分割,分别包含 3 200 和 800 个样本。由于子空间的划分具有随机性,因此三种集成分类算法均采取 10 次独立运行的平均值,所有实验均在 MATLAB 7.1 平台上完成。

e)对于情感识别结果的对比,采用查准率 P、查全率 R 以及 $F1$ 值三种评价指标进行衡量:

$$P=\frac{1}{2}\left(\frac{\sum_{c_i=-1}\mathrm{True}(c_i)}{\sum_{c_i=-1}\mathrm{Samp}(c_i)}+\frac{\sum_{c_j=+1}\mathrm{True}(c_j)}{\sum_{c_j=+1}\mathrm{Samp}(c_j)}\right) \tag{12}$$

$$R=\frac{1}{2}\left(\frac{\sum_{c_i=-1}\mathrm{True}(c_i)}{\sum_{c_i=-1}\mathrm{Res}(c_i)}+\frac{\sum_{c_j=+1}\mathrm{True}(c_j)}{\sum_{c_j=+1}\mathrm{Res}(c_j)}\right) \tag{13}$$

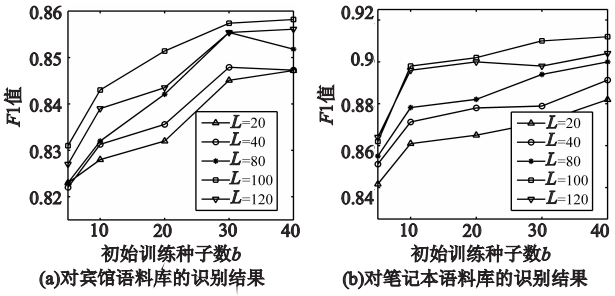
$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{14}$$

式(12)(13)中:$\mathrm{True}(c_i)$ 表示被识别为 c_i 且正确的样本数,式(13)中的 $\mathrm{Res}(c_i)$ 表示识别结果为 c_i 的样本数。

4.3 实验结果与分析

部分实验结果如图 2 与表 2 所示。图 1 表示采用不同的参数(子空间采样数量 L 以及初始训练子集数量 b)时,RDS 算法在两种评论数据集上的情感识别性能的变化。表 2 为四种不同识别方法的识别结果。其中,实验 2 中,三种集成学习方法均使用了三种子空间的生成数量(L = 20,40,100)进行对比。

初始训练子集的选择是 RDS 算法的关键环节。初始训练种子数量 b 越大,为后续子空间的生成提供的反馈数据越多;相反,b 越小,子空间的构造将会接近完全随机选择的状态,且集成识别模型的鲁棒性会降低。但 b 值过大也会引起算法性能的下降,且初始分布 R_0 的计算复杂度将提高。由图 2 可知,b 值的提高一定程度上改善了识别效果,从对两种语料的识别情况看,b 值在[5,30]区间内提高时具有较好的效果,但继续提高 b 值,性能提升并不显著,说明 b = 30 是识别性能趋于稳定的一个拐点。另外,子空间的生成数量 L 对识别模型的性能也具有较大的影响。从集成学习中差异度的角度看,L 越大,基分类器的差异度也会下降,会出现多个基分类器具有同样的分类能力,超过一定数量的子空间并没有对集成模型提供有效的分类互补信息,极端情况下会出现集成分类器与单分类器性能等效的情况。可以观察到,在对宾馆数据集识别时,当基分类器数量 L = 20 时,由于迭代数量过少,$F1$ 值仅为 84% ;L = 40 时,识别性能较好,达到 85.74% 的水平;当 L = 120 时,子空间中选择的样本覆盖率达到了 96% 以上,但由于基分类器之间差异度下降,$F1$ 值不再提高,反而有下降的趋势。在笔记本语料的处理中,RDS 的识别效果较好,当取 b = 30,L = 100 时,取得了最高,$F1$ 值为 91% 。此外,基分类器数量 L 在[20,100]区间内递增时,识别性能的提升较为显著,最高提升幅度约为 5% 。



(a)对宾馆语料库的识别结果 (b)对笔记本语料库的识别结果

图 2 初始训练种子数 b 与子空间数量 L 对 RDS 算法识别性能的影响

在表 2 中,从查准率、查准率以及 $F1$ 值的统计结果可观察到,与基于训练样本全空间的单分类器 LSVM 相比,三种集成分类算法在情感识别上表现出更高的识别性能。其中,关于宾馆语料的识别结果中,在基分类器数量从 40 提高到 100 时,Bagging 算法识别的查全率与查准率均有所下降,并未体现出集成算法的优势,表明完全随机的子空间划分机制会引起子空间内大量噪声数据的出现,一些稳定性较差的基分类器影响集成模型识别的鲁棒性,因此 Bagging 算法在基分类器数量提高的同时,性能提升并不明显。而 AdaBoost 算法考虑到样本空间内样本的权重概率分布,能在一定程度上提高基分类器之间的差异度。对两类语料集的处理中,AdaBoost 的性能总体上略高于 Bagging,但在实验的中间结果中发现,由于样本分布权重的更新过多集中在错分的训练样本上,使子空间的采样仅局限于某些权重较大的样本上,覆盖率仅为 72% 左右,并且多个基分类器的识别精度仅略高于 50% ,表明个体基分类器的性能并不稳定。

表 2 不同算法在两种语料集(宾馆和笔记本)上的识别性能比较

算法	L	宾馆			笔记本		
		P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
RDS	20	84.61	84.47	84.54	87.23	87.01	87.12
	40	85.20	84.38	84.79	88.13	87.65	87.89
	100	85.85	85.81	85.83	91.17	90.83	91.00
Bagging	20	84.52	84.51	84.52	85.16	84.92	85.04
	40	85.15	85.13	85.14	85.27	84.83	85.05
	100	84.94	84.92	84.93	86.34	86.16	86.25
AdaBoost	20	83.48	83.34	83.41	84.89	84.90	84.88
	40	84.12	84.03	84.08	85.23	85.14	85.14
	100	85.14	84.96	85.05	86.74	86.36	86.55
LSVM	N/A	83.92	83.91	83.92	85.04	84.80	84.92

注:L 表示基分类器数量,P 表示查准率,R 表示查全率

与 Bagging 和 AdaBoost 算法相比,RDS 充分考虑到不同子空间内鉴别信息的粒度分布,并且避免了预先设定子空间的选取粒度,有效提高了子空间选择的自由度。对以上结果进行对比可以看到,RDS 在两类语料上的识别结果均得到了提升,特别在对笔记本语料的识别中,当基分类器数量提高到 100 时,

RDS 的查准率达到了 91.17% 的最高精度,比其他方法最高高出了约 6%。这说明多粒度的子空间划分机制能有效改善集成识别系统的性能。但值得注意的是,基分类器的数量太大也会降低集成系统的运行效率,虽然在一定范围内提高基分类器数量能提升分类的性能,但超过某个阈值性能会趋于稳定,甚至会略有下降,其根本原因在于大量的基分类器生成后,差异度会随之减小。

5　结束语

通过一系列实验可以看出,在使用集成学习方法进行中文文本情感识别时,其识别性能取决于子空间的划分粒度以及不同子空间之间的差异度。针对中文情感识别中样本空间规模大、稀疏度高等特点,提出一种基于样本空间划分的集成分类方法,并通过核平滑方法对子空间进行自适应选取。实验表明,这种多粒度的子空间选择方法能有效改善中文文本情感识别的性能,可以实现大规模评论集的情感识别、观点挖掘等工作。下一步将就样本空间内情感特征的分布对动态子空间算法性能的影响展开进一步的研究。另外,将结合基于情感特征空间的集成分类机制设计更为优化的中文情感识别算法。

参考文献:

[1] ZHAI Zhong-wu, XU Hua, JIA Pei-fa. An empirical study of unsupervised sentiment classification of Chinese reviews[J]. Tsinghua Science and Technology, 2010,15(6):702-708.

[2] 段建勇,谢宇超,张梅. 基于句法语义的网络舆论情感倾向性评价技术研究[J]. 情报杂志,2012,31(1):147-150.

[3] 杨江,彭石玉,侯敏. 基于主题情感句的汉语评论文倾向性分析[J]. 计算机应用研究,2011,28(2):569-572.

[4] LI Shou-shan, HUANG Chu-ren, ZONG Cheng-qing. Multi-domain sentiment classification with classifier combination[J]. Journal of Computer Science and Technology, 2011,26(1):25-33.

[5] 胡熠,陆汝占,李学宁,等. 基于语言建模的文本情感分类研究[J]. 计算机研究与发展,2007,44(9):1469-1475.

[6] WANG Su-ge, LI De-yu, SONG Xiao-lei, *et al*. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification[J]. Expert Systems with Applications, 2011, 38(7):8696-8702.

[7] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008,2(1):1-135.

[8] ZHAI Zhong-wu, XU Hua, KANG Ba-da, *et al*. Exploiting effective features for Chinese sentiment classification[J]. Expert Systems with Applications, 2011,38(8):9139-9146.

[9] 钟将,邓时滔. 基于多特征融合的汉语情感分类研究[J]. 计算机应用研究,2012,29(1):98-100.

[10] 知网[EB/OL]. [2011-04-22]. http://www.keenage.com.

[11] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87-93.

[12] 顾亚祥,丁世飞. 支持向量机研究进展[J]. 计算机科学,2011,38(2):14-17.

[13] HALL M, FRANK E, HOLMES G, *et al*. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations Newsletter, 2009,11(1):10-18.

[14] BISHOP C M. Neural networks for pattern recognition[M]. New York: Oxford University Press, 1995.

[15] 谭松波. 中文情感挖掘语料 ChenSentiCorp[EB/OL]. (2010-06-29)[2011-04-22]. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm.

[16] BREIMAN L. Bagging predictor[J]. Machine Learning, 1996, 24(2):123-140.

[17] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997,55(1):119-139.

(上接第 1442 页)队伍调配的实际需求,从行政等级、地震烈度、人口规模三方面确定受灾点的优先等级,为突出应急救援的时效性和救援目标的多属性特征,以应急救援效率最大化、物资损耗最小化为优化目标,构建最短时限应急救援多目标指派模型以优化救援力量。本文给出了相对完整的技术思路和初步应用,但模型仍存在一些不足,如未考虑救援队伍的特征以及受灾点对救援队伍数量的要求,因此,将受灾点对专业队伍的救援需求引入到多目标时限指派问题是进一步研究的方向。

参考文献:

[1] 陈颙,陈棋福,黄静,等. 减轻地震灾害[J]. 地震学报,2003,25(6):621-629.

[2] ERDIK M, AYDINOGLU N, FAHJAN Y, *et al*. Earthquake risk assessment for Istanbul metropolitan area[J]. Earthquake Engineering and Engineering Vibration, 2003,2(1):1-23.

[3] 李柞泳,邓新民. 自然灾害的物元分析灾情评估模型初探[J]. 自然灾害学报,1994,3(2):28-33.

[4] 雷秋霞,陈维锋,黄丁发,等. 地震现场搜救力量部署辅助决策系统研究[J]. 地震研究,2011,34(7):384-388.

[5] OZDAMAR L, EKINCI E, KUCUKYAZICI B. Emergency logistics planning in natural disasters[J]. Annals of Operations Research, 2004,129(3): 217-245.

[6] YI W, OZDAMAR L. A dynamic logistics coordination model for evacuation and support in disaster response activities[J]. European Journal of Operational Research,2007,179(3):1177-1193.

[7] CHANG M, TSENG Y, CHEN J. A scenario planning approach for the flood emergency logistics preparation problem under uncertainty [J]. Transportation Research Part E: Logistics and Transportation Review, 2007,43(6):737-754.

[8] TZENG G, TSUNG H, HUANG D. Multi-objective optimal planning for designing relief delivery systems[J]. Transportation Research Part E: Logistics and Transportation Review,2007,43(6):673-686.

[9] HANDRIGAN M T, BECKER B M, JAGMINAS L, *et al*. Emergency medical services in the reconstruction phase following a major earthquake: a case study of the 1988 Armenia earthquake[J]. Prehospital Disaster Med, 1998,13(1):35-40.

[10] FIEDRICH F, GEHBAUER F, RICKERS U. Optimized resource allocation for emergency response after earthquake disasters[J]. Safety Science, 2000,35(1-3): 41-57.

[11] YATES D, PAQUETTE S. Emergency knowledge management and social media technologies: a case study of the 2010 Haitian earthquake[J]. International Journal of Information Management, 2011,31(1):6-13.

[12] 樊治平,刘洋,袁媛,等. 突发事件应急救援人员的分组方法研究[J]. 运筹与管理,2012,21(2):1-7.

[13] 袁媛,樊治平,刘洋. 生命线网络系统多节点失效的应急抢修队伍派遣模型研究[J]. 运筹与管理,2012,21(1):131-135.

[14] 李亦纲,张媛,李志伟. 地震灾区救援力量优化调配模型[J]. 自然灾害学报,2012,21(3):150-154.

[15] 陈守煜. 系统模糊决策理论与应用[M]. 大连:大连理工大学出版社,1994.

[16] 夏少刚,刘佳. 利用最小调整法求解特殊的二维 0-1 规划[J]. 运筹与管理,2008,17(1):24-28.