基于上下文重构的短文本情感极性判别研究

杨 震 [1] 赖英旭 [1] 段立娟 [1] 李玉鑑 [1]

摘 要 文本对象所固有的多义性, 面对短文本特征稀疏和上下文缺失的情况, 现有处理方法无法明辨语义, 形成了底层特征和高层表达之间巨大的语义鸿沟. 本文尝试借由时间、空间、联系等要素挖掘文本间隐含的关联关系, 重构文本上下文范畴, 提升情感极性分类性能. 具体做法对应一个两阶段处理过程: 1) 基于短文本的内在联系将其初步重组成上下文 (领域); 2) 将待处理短文本归入适合的上下文 (领域) 进行深入处理. 首先给出了基于 Naive Bayes 分类器的短文本情感极性分类基本框架, 揭示出上下文 (领域) 范畴差异对分类性能的影响. 接下来讨论了基于领域归属划分的文本情感极性分类增强方法, 并将领域的概念扩展为上下文关系, 提出了基于特殊上下文关系的文本情感极性判别方法. 同时为了解决由于信息缺失所造成的上下文重组困难, 给出基于遗传算法的任意上下文重组方案. 理论分析表明, 满足限制条件的前提下, 基于上下文重构的情感极性判别方法能够同时降低抽样误差 (Sample error) 和近似误差 (Approximation error). 真实数据集上的实验结果也验证了理论分析的结论.

关键词 舆情分析, 短文本处理, 情感计算, 误差分析, 遗传算法

DOI 10.3724/SP.J.1004.2012.00055

Short Text Sentiment Classification Based on Context Reconstruction

YANG Zhen[1] LAI Ying-Xu[1] DUAN Li-Juan[1] LI Yu-Jian[1]

Abstract Synonymy and polysemy present a challenge to effective natural language processing, especially in the situations of context absence and sparse feature in short texts, widened semantic gap between low-level text features representation and high-level interpretation. In this work, short texts were reorganized into special context, i.e., the implied internal relationship such as time and space, and a novel two-step scheme for semantic orientation detection based on the special context was proposed. In the first step, the short texts were reorganized into special contexts by the implied internal relationship. In the second step, the unknown short text was categorized into a special context and labeled a polarity tag using the inner semantic orientation classifier. We firstly discussed the effect of special context after a sentiment classification framework based on naive Bayes classifier was presented. Then an enhancement classification method was given using field concept, which was expanded to special context. Finally, a special context reorganizing method was proposed based on genetic algorithm. Theoretical analysis shows the proposed methods can reduce the sample error and approximation error under some constraints. The experimental results in real corpora show the effectiveness of the proposed method.

Key words Public opinion analysis, short text processing, sentiment classification, error analysis, genetic algorithm

近代对情感计算的系统研究始于 Picard 教授的专著 *Affective Computing*[1], 书中指出情感计算是关于情感、情感产生以及影响情感方面的计算. 情感计算相关研究从出现伊始就受到广大研究者的关注, 在诸如自动情绪检测与分析、人机接口与脑机接口方面取得了丰硕的成果. 同时由 Picard 教授主编的情感计算 IEEE 汇刊也于 2010 年正式创刊, 标志着情感计算相关研究已经迈入了一个新的阶段. 具体到文本内容处理领域, 主要面对的情感计算问题是情感识别问题, 即着眼于分析确定一个说话人或者作者对于某些特定主题的态度. 情感识别问题可以看成话题检测与追踪 (Topic detection and tracking, TDT) 问题的一个自然延伸[2], 在实现话题或事件的检测与追踪后, 人们自然希望能够更进一步, 对话题和事件所表达的情感进行分析和识别. 通常文本情感识别包含三个子问题: 1) 极性分类 (Polarity classification)[3−4]; 2) 主客观分类

收稿日期 2011-03-28 录用日期 2011-07-07
Manuscript received March 28, 2011; accepted July 7, 2011
国家自然科学基金 (61001178, 60905017, 60702031, 61002029), 北京市自然科学基金 (4102012, 4112009, 4102013, 4123093), 北京市教育委员会科技发展计划面上项目 (KM201210005024), 国家软科学研究计划项目 (2010GXQ5D317), 北京市高等学校人才强教深化计划 "中青年骨干人才培养计划" 项目 (PHR201108016), 北京工业大学高层人才培养项目, 北京工业大学校青基金资助
Supported by National Natural Science Foundation of China (61001178, 60905017, 60702031, 61002029), Beijing Natural Science Foundation (4102012, 4112009, 4102013, 4123093), Scientific Research Common Program of Beijing Municipal Commission of Education (KM201210005024), National Soft Science Research Program (2010GXQ5D317), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR201108016), Beijing University of Technology High-Level Personnel Development Project, and Beijing University of Technology Research Fund For Young Teachers
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 北京工业大学计算机学院 北京 100124
1. College of Computer Sciences, Beijing University of Technology, Beijing 100124

(Subjective/Objective classification)[3]; 3) 情感强度分类 (Rating inference)[5].

在文本情感识别问题中, 文本的极性分类问题是其中较为重要的一个部分. 即通过分析作者的情绪状态, 或者有意向受众传递的情感讯息, 从而判别文本内容是正面的肯定赞赏还是负面的否定批判[6]. 特别是对于舆情分析问题来说, 话题情感极性的准确判断将有助于弄清大众对话题的倾向性 (赞成或是反对), 从而更深层次的把握舆论走向, 因而具有特别重要的意义, 成为本文重要的研究目标.

文本的情感极性判断也可以看成是一类特殊的文本分类问题, 通过将不同情感倾向的文本视为不同的类别, 利用类别间的差异性建立分类器以区分不同的文本情感极性. 目前, 对于普通网络文本的情感极性识别, 研究者基于概率模型和知识本体 (如 WordNet, HowNet) 等方法[7−9], 已经取得了较为理想的效果. 但是对于短文本情感极性识别, 却因为短文本自身的特性 (稀疏性、实时性、不规范性等)[10], 为其分析处理带来了巨大难题. 随着短信、在线社区、微博客的兴起, 网络短文本已初具规模, 其发展趋势已不可逆转. 特别是针对以即时消息、在线聊天记录、BBS 标题、手机短消息、微博交互、博客评论、新闻评论等为代表的短文本信息, 由于手持输入设备和移动交流方式的限定, 不可能进行长篇累牍的交流. 相较于普通网络文本, 其长度通常要小得多 (1 ∼ 100 词量级). 目前短文本情感计算相关研究相对滞后, 得到了研究者的广泛关注.

基于文本是 “Bag of words” 的假设, 文本分类通常用其所包含的词作为特征. 将语料库中包含的所有词进行选择构成属性特征, 每一篇文本根据相应属性特征是否出现, 进行特征向量表示. 一般来说特征属性规模庞大 (10 000 ∼ 100 000 词量级), 而单一文本长度有限, 不管使用什么特征表示模型, 包括向量空间模型 (Vector space model, VSM)、统计模型和基于句法分析的模型, 单一文本特征相对于特征属性总体来说数量都较少, 因此不可避免出现特征稀疏问题. 对于长度小得多的短文本, 其特征更为稀疏, 使短文本处理在文本表达这一环节就遇到了严重的困难. 以本文处理的情感极性分类短文本材料为例, 分析可知其特征属性多达 29 550 词, 而平均文本长度为 70.4114 词, 文本稀疏度达 0.16573 %, 即每个文本向量中, 只有极少数的维数上是有取值的. 极度稀疏的表达, 给短文本的处理带来了极大的困难. 针对这样的问题, 研究者进行了多方面的尝试. 由于短文本的处理困难一方面可以说是由于特征属性规模庞大造成的, 另外一方面也可认为是由于单一文本自身特征太少造成的. 因此, 研究者从这两方面尝试着手解决问题. 一方面, 研究者通过特征降维避免稀疏性难题. 对高维数据进行有效降维得到了研究者长期以来的关注, 主成分分析 (Principal component analysis, PCA) 方法、潜在语义索引 (Latent semantic indexing, LSI) 及其改进 — 基于聚类的潜在语义索引 (Clustered LSI, CLSI) 方法[11]、基于聚类重心数据降维 (Centroid method, CM) 的方法[12] 在文本分类领域得到了成功的应用. Xu 等[13] 提出了利用改进的潜在语义分析方法 dual-PLSA 解决手机短信分类问题. Mørch 等[14] 等在香港英语外语教学中使用 LSI 方法分析学生提交的英文小作文和范文之间的差异. 但这些降维的方法或者需要计算大矩阵的特征值和特征向量, 或者需要对数据进行频繁的聚类迭代分析, 其计算复杂度和计算时间都比较大. 另外一方面, 研究者希望通过属性联想或组合扩充短文本特征. Wang 等[15] 利用 WR-kmeans 聚类方法综合相关手机短消息解决相似短文本发现问题. Adams 等[16] 利用 WordNet 解决即时聊天信息话题检测与抽取问题. Fan 等[17] 提出一种特征扩展和控制模型, 能有效提高短文本的分类精度. 闫瑞等[18] 通过构造出一种树状组合分类器对短文本进行动态组合分类, 在一定程度上缓解了短文本特征稀疏和样本高度不均衡对分类性能的影响. 但目前属性扩充的方法需要大量人工参与和依赖于计算复杂性非常高的语义计算 (如基于知识本体 WordNet、HowNet 等的语义相似度计算), 限制了方法的使用范围. O′Shea 等[19] 就是依靠人工打分的方式对大约 65 个句子的相似度进行评分. 当然也可同时进行属性约简和特征扩充. 杨锋等[20] 基于 SCP-X 模型, 利用复杂网络理论对特征属性进行约简和对文本特征进行扩充, 解决在线评论情绪倾向性分类问题.

短文本之所以难于处理, 追本溯源是因为文本对象特征 (即语言) 本身所固有的多义性, 即存在一词多义 (Polysemy) 和一义多词 (Synonymy) 的特点, 面对短文本特征的稀疏性和上下文缺失的情况, 造成语义难以明辨, 理解偏差无法消解, 最终形成短文本底层特征和高层语义之间巨大的语义鸿沟. 为了破解这样的难题, 本文试图利用短文本内在的关联关系, 重构文本上下文范畴, 从而克服短文本过于短小的先天缺憾, 提升情感极性分类的性能. 由于这样的上下文关系并非自然上下文关系, 我们将其称为特殊上下文关系. 通过引入特殊上下文的方式进行短文本情感极性判别的研究, 目前未见相关研究报道. 具体做法对应一个两阶段处理过程: 1) 基于短文本的内在联系将其初步重组成上下文 (领域); 2) 将待处理短文本归入适合的上下文 (领域) 进行深入处理. 对于训练过程: 首先, 将短文本按上下文关系进行重组, 并训练上下文分类器; 其次, 对同一

上下文内短文本, 按照其情感极性的不同, 训练相应的域内情感分类判别器. 对于分类过程: 首先, 将待测短文本归类于适合的上下文范畴; 其次, 用相应的域内情感分类器进行分类判别, 输出判决结果. 由于待处理的短文本缺乏明显的自然组织, 因此如何发现文本间隐含的关联关系将其重新组合, 即基于时间、空间等因素的特殊上下文觉察技术成为提升情感极性分类性能的关键. 同时文中还给出了用遗传算法组织任意特殊上下文的方法. 理论分析表明, 满足一定条件的前提下, 基于特殊上下文的分类方法能够同时降低抽样误差 (Sample error) 和近似误差 (Approximation error). 真实数据集上的实验结果也印证了理论分析的结果.

本文的其余部分安排如下: 第 1 节给出了基于 Naive Bayes 分类器的短文本情感极性分类的基本框架; 第 2 节讨论了基于领域归属划分的文本情感极性分类增强方法; 第 3 节进一步探讨了基于上下文关系自动发现的短文本情感极性分类增强方法; 第 4 节给出了论文的总结和未来工作的展望.

1　文本情感极性分类基本框架

毫无疑问, 文本情感极性分类可以看成是一种特殊的文本分类问题, 以判别文本自然语言文字中表达的观点是正面的肯定赞赏, 还是负面的否定批判. 基于这样的思路, 本文采用朴素贝叶斯分类器作为短文本情感极性分类的基本框架.

假设短文本由属性值 $\{w_i\}_{i=1}^m$ 描述, 目标属性集合 $C=\{c_1,c_2\}$, 其中 c_1 表示赞成 (Positive), c_2 表示反对 (Negative). 对于待分类短文本的分类决策为

$$H_{\mathrm{NB}}=\arg\max_{c\in C}\ P(c)\prod_{i=1}^{n}P(w_i|c) \qquad (1)$$

其中, $P(w_i|c)$ 的计算采用 Multinomial 模型. 通过正确估计 $P(c)$ 和 $P(w_i|c)$ 值, 就对待分类短文本的情感极性进行判别.

为了评估基本框架对于短文本情感极性分类的性能, 首先对其进行基线测试. 测试所用情感分析语料集为未去重的平衡语料集 ChnSentiCorp (可在以下网址获得: http://www.searchforum.org.cn/tansongbo/corpus-senti.htm), 包含酒店、电脑与书籍三个领域的相关评价, 其具体的组成如表 1 所示. 每种类型的语料来源单一, 但不同类型间的来源差异性较大. 通过对数据集的详细分析可知, 在文本清洗、分词和预处理之后, 得到的文本有效属性为 29 550 个, 单一文本平均由 76.3525 个词组成. 在去除无效字符后, 平均文本长度为 70.4114 个词. 由于文本中所包含的词有重复, 最后用于计算的特征平均稀疏度为 0.16573 %, 属于特征严重稀疏的情况, 是典型的短文本范式.

表 1　实验用情感分析语料

Table 1　Short text corpora

	正面语料 (篇)	负面语料 (篇)	单篇平均词长	数据来源
书籍评价	2 000	2 000	90.6423	当当图书网
酒店评价	2 000	2 000	95.2885	携程旅行网
电脑评价	2 000	2 000	43.1268	京东网上商城

表 2 给出了基于 Naive Bayes 分类器的短文本情感极性分类测试结果. 评测了使用书籍评价、酒店评价、电脑评价以及全部语料为训练语料, 分别分类其他类别语料时取得的性能. 实验所用评价指标包括每类的精度 (P)、召回率 (R)、准确率 (Auc) 和平均准确率. 实验采用 10 次 10 倍交叉验证 (10-fold cross validation), 即将数据集分成十份, 轮流将其中 9 份做训练 1 份做测试, 10 次的结果的均值作为对算法精度的估计. 同时交叉实验随机重复 10 次, 取其平均值和方差作为最后结果 (即每个实验指标为 100 次实验平均结果).

如表 2 所示, 基线系统处理同源短文本语料的性能尚可, 但对于处理不同领域的其他文本, 其分类能力较弱. 即当训练领域和测试领域不同时, 情感分类器的性能劣化. 有些研究者将这样的性能下降归结为训练 (源) 领域 (Old domain) 和测试 (目标) 领域 (New domain) 的样本分布有着较大差异, 使得仅用源领域语料训练的分类器无法对目标领域的样本进行良好区分. 为此采用情感语料移植的方法, 充分利用目标领域的语料参与训练, 综合得到一个能够良好反映源领域和目标领域特性的分类器[21]. 但我们从表 2 中还发现一个有趣的结果, 虽然只用一个领域的语料做训练, 在其他语料上取得的平均分类性能不高, 但对于同源语料的分类性能却很理想, 甚至优于使用所有语料进行训练分类的结果. 这提示我们, 除了费心进行领域融合之外, 还可以先确定测试语料所属领域, 然后再在该领域中进一步精细处理, 判别情感极性. 从这个角度看, 问题的关键可以转化为如何为待测语料选择一个适合的领域.

2　基于领域归属划分的短文本情感极性分类增强方法

基于第 1 节的结论, 我们希望先确定短文本所归属的领域, 再在领域内进行情感分类. 从而避免进行复杂的领域融合和迁移操作, 降低分类难度, 提升判别性能. 本节探讨了一种基于领域归属划分的文本情感极性分类增强方法, 理论分析表明, 满足一定条件的前提下, 基于领域归属划分的文本情感极性分类方法能够同时降低抽样误差和近似误差.

表 2 Naive Bayes 分类器在 ChnSentiCorp 数据集上的分类准确率

Table 2 Performance of naive Bayes classifer in ChnSentiCorp corpora

测试集	书籍评价					酒店评价					电脑评价					
	负面评价		正面评价		AUC	负面评价		正面评价		AUC	负面评价		正面评价		AUC	平均 AUC
训练集	P	R	P	R		P	R	P	R		P	R	P	R		
书籍评价	0.8093 ± 0.0009	0.9421 ± 0.0003	0.9312 ± 0.0004	0.7782 ± 0.0010	0.8600 ± 0.0003	0.5225 ± 0.0007	0.9828 ± 0.0001	0.8564 ± 0.0060	0.1020 ± 0.0007	0.5424 ± 0.0006	0.5142 ± 0.0006	0.9753 ± 0.0001	0.7601 ± 0.0078	0.0785 ± 0.0005	0.5269 ± 0.0005	0.6431 ± 0.0004
酒店评价	0.6211 ± 0.0010	0.6376 ± 0.0018	0.6279 ± 0.0011	0.6104 ± 0.0018	0.6240 ± 0.0005	0.9243 ± 0.0005	0.7820 ± 0.0009	0.8109 ± 0.0008	0.9359 ± 0.0004	0.8589 ± 0.0003	0.7962 ± 0.0017	0.4453 ± 0.0016	0.6152 ± 0.0008	0.8864 ± 0.0005	0.6657 ± 0.0006	0.7162 ± 0.0004
电脑评价	0.5587 ± 0.0011	0.6450 ± 0.0014	0.5803 ± 0.0013	0.4905 ± 0.0015	0.5676 ± 0.0006	0.7194 ± 0.0009	0.8711 ± 0.0006	0.8366 ± 0.0010	0.6600 ± 0.0014	0.7656 ± 0.0005	0.8950 ± 0.0004	0.9150 ± 0.0004	0.9130 ± 0.0004	0.8925 ± 0.0005	0.9038 ± 0.0002	0.7457 ± 0.0004
全部语料	0.9569 ± 0.0003	0.6474 ± 0.0012	0.7336 ± 0.0007	0.9707 ± 0.0001	0.8090 ± 0.0004	0.7601 ± 0.0009	0.9615 ± 0.0002	0.9474 ± 0.0003	0.6967 ± 0.0012	0.8290 ± 0.0004	0.8529 ± 0.0006	0.8795 ± 0.0005	0.8755 ± 0.0006	0.8483 ± 0.0006	0.8638 ± 0.0003	0.8339 ± 0.0004

2.1 基于领域归属划分的性能增强方法

基于领域归属划分的文本情感极性分类方法如图 1 所示. 建模成一个两阶段处理过程:

第一阶段 (域间归属分类过程): 确定待处理短文本所属的领域;

第二阶段 (域内情感判别过程): 用相应领域的分类器对待处理短文本进行情感判别.

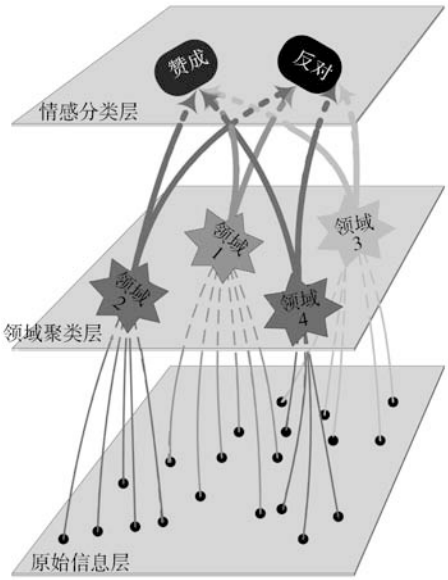

图 1 基于领域归属划分的短文本情感极性分类方法示意图

Fig. 1 Short text sentiment classification based on context reconstruction using field information

具体来说, 对于训练过程: 首先, 将原始文本信息按照领域关系重组, 并训练领域分类器; 其次, 对同一领域内的短文本, 按照其情感极性的不同, 训练相应的域内情感判别器. 对于分类过程: 首先, 将待测短文本归类于适合的领域; 其次, 用相应的域内情感判别器进行分类判别, 输出判决结果. 基于领域归属划分的文本情感极性分类方法体现了 "区域自治" 的原则, 期望将待处理短文本交由最熟悉其特性的领域进行处理. 从而避免进行复杂的领域融合和迁移操作, 降低分类难度, 提升判别性能. 算法细节可见算法 1.

算法 1. 基于领域归属划分的短文本情感极性分类算法

步骤 1. 对于待处理短文本集 N, 按照其领域划分成 m 个簇 $C_1, C_2, \cdots, C_m$;

步骤 2. 为这 m 个簇基于多分类思想, 训练域间分类器 O;

步骤 3. 根据 $C_1, C_2, \cdots, C_m$ 域内样本情感极性分布, 建立相应的域内分类器 θ^i, $i = 1, 2, \cdots, m$;

步骤 4. 对于待处理短文本, 首先由域间分类器 O 判断其领域归属, 然后再用相应的域内情感极性分类器 θ^i 进行判别, 输出判别结果.

2.2 误差分析与性能估计

基于领域归属划分的文本情感极性分类算法简单易操作, 我们只是模糊的期待由于待处理样本是交由最熟悉其特性的领域处理, 因而能够取得良好的性能. 本节将具体分析算法的误差, 并对其性能进行估计.

为了揭示影响分类器性能的因素, 众多研究者对各种可能导致分类错误的原因进行了研究. 发现可以根据误差产生的不同原因, 将其分成不同的部分. 研究者们从不同的角度入手, 得到不同的误差分解式, 包括将分类器错误分解为偏差 (Bias) 和变化

(Variance) 项, 或者偏差和传播 (Spread) 项, 或者抽样误差和近似误差组合. 1995 年, Kong 等[22] 针对 0/1 损失函数给出偏差 –变化 (Bias-variance) 分解式. 1996 年, Breiman[23] 提出了针对 0/1 损失函数的分类器平均错误率的分解式. 同年, Tibshirani[24] 给出了 Breiman 分解式的一个补充, 并探讨了基于 Aggregation effect 的分解. Kohavi 等[25] 提出了 Bias-variance 分解, 但是其分解对于 Bayes 分类器时, Bias 不为 0. 随着研究的深入, 人们发现偏差的产生并不总是坏事, 我们还可以根据误差的不同成因, 采取不同的处理策略. Friedman[26] 对于二分类问题, 其类分布为 Gaussian 分布情况做出了研究, 他指出在某些情况下 Variance 的增加会降低集成错误. Domingos[27] 使用最优估计的概念给出了针对 0/1 损失函数和平方损失函数的统一 Bias-variance 分解. Zhou 等[28] 将误差归因于学习能力、样本抽样、数据噪声, 并提出了一种基于 C4.5 的神经网络集成方法. Cuker 等[29] 在给出学习过程的一般理论的同时, 将期望误差分解为抽样误差和近似误差组合. 基于这些理论成果和分析框架[22−29], 本节对基于领域归属划分的文本情感极性分类算法性能进行分析.

假设数据 X 是欧氏空间中的紧子集或流形, $Y = \mathbf{R}$ 为数据 X 的期望 (真实) 取值, ρ 是定义在 $Z = X \times Y$ 上的 Borel 概率测度, F 为预测函数族空间, $f \in F$ 表示 $X \to Y$ 的映射. 在本文中, X 表示短文本空间, Y 为情感类别空间, F 是分类算子族. 接下来定义损失函数、误差函数.

定义 1. 损失函数 $L(a,b)$ 用于度量由于 a, b 之间的差异而产生的损失, 典型的损失函数有:

1) 二次损失函数 (Squared loss)

$$L(a,b) = (a-b)^2 \tag{2}$$

2) 0-1 损失函数 (Zero-one loss)

$$L(a,b) = 1 - \delta(a-b) = \begin{cases} 0, & a = b \\ 1, & \text{其他} \end{cases} \tag{3}$$

分别对应于所谓回归估计和分类判别问题.

定义 2. 对于任意 $x \in X$, 其期望 (真实) 值为 $y \in Y$, 预测映射 $f : X \to Y$. 那么映射 f 的误差函数定义为

$$\varepsilon_\rho(f) = \int_z L(f(x), y)\mathrm{d}\rho \tag{4}$$

给定 $f_F \in F$ 表示函数族 F 中最小化误差函数 $\varepsilon_\rho(f)$ 的最佳映射.

定义 3. 给定真实数据 Z 的一个抽样 $z \subset Z$, $z = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, 那么映射 f 的经验误差函数定义为

$$\varepsilon_z(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) \tag{5}$$

给定 $f_z \in F$ 表示函数族 F 中最小化误差函数 $\varepsilon_z(f)$ 的最佳映射.

定义 4. 实际的学习过程是一个经验风险最小过程: 基于有限抽样 z, 寻找预测函数族 F 中的最佳映射 $f_z \in F$, 能够最小化经验损失 (误差) 函数 $\varepsilon_z(f)$, 即

$$\min_{f\in F} \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) \tag{6}$$

一个好的学习算法应该能够保证 f_z 在最小化经验损失 $\varepsilon_z(f)$ 的同时, 也能最小化期望损失 $\varepsilon_\rho(f)$. 根据 Cuker 等[29] 的学习理论, 可以将 f_z 的期望误差分解为抽样误差和近似误差:

$$\varepsilon_\rho(f_z) = \overbrace{[\varepsilon_\rho(f_z) - \varepsilon_\rho(f_F)]}^{\text{Sample error}} + \overbrace{\varepsilon_\rho(f_F)}^{\text{Approximation error}} \tag{7}$$

其中, 抽样误差反映了抽样数据 z 的规模和噪声影响, 而近似误差反映了系统的学习能力, 仅仅依赖于预测函数空间 F 的选择, 独立于抽样 z. 简单来说, 固定预测函数族空间 F, 增大抽样样本容量, 能够使得 f_z 收敛于 f_F, 从而减少抽样误差. 固定抽样样本容量, 增大预测函数族空间 F, 能够减小近似误差但增大抽样误差. 因此, 许多研究者将学习过程看作是一个所谓 Bias-variance 折衷, 其中 Bias 指近似误差, Variance 指抽样误差[30], 提出了诸如结构风险最小化原理[31] 及正则化学习理论[32] 等学习框架.

定义 5. 假设 X 表示短文本数据空间, Y 为情感类别空间, ρ 是定义在 $Z = X \times Y$ 上的联合概率分布函数, 给定真实数据 Z 的一个抽样 $z \subset Z$, $z = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. G, F 为预测函数族空间. 对于一般的学习方法, 基于经验风险最小化原则从 F 中选择最佳映射 f_z^0 能够最好地进行短文本情感分类, 其期望误差表示为 $\varepsilon_\rho(f_z^0)$. 对于基于领域归属划分的文本情感极性分类方法, 第 1 阶段域间归属分类过程, 基于数据 z 的领域归属选择预测函数 $g_z^I \in G$ 能够最好地将不同领域间数据分割开来, 其期望误差表示为 $\varepsilon_\rho(g_z^I)$; 第二阶段域内情感判别过程, 基于归属于领域中的数据 $z' \subset z$ 选择域内情感判别函数 $f_{z'}^{II} \in F$ 能够将域内不同情感极性数据分割开来, 其期望误差表示为 $\varepsilon_\rho(f_{z'}^{II})$.

定理 1. 给定数据抽样 $z \subset Z$, $0 < \Delta \ll 1$, 当

$$\varepsilon_\rho(g_z^I) \le \frac{\varepsilon_\rho(f_z^0) - \varepsilon_\rho(f_{z'}^{II}) + \Delta}{1 - \varepsilon_\rho(f_{z'}^{II})} \tag{8}$$

基于领域归属划分的文本情感极性分类方法总能比一般学习方法取得更好的性能.

证明. 对于一般的学习方法, 基于经验风险最小化原则, 从预测函数族 F 中选择最佳映射 f_z^0 能够最好地进行短文本情感分类, 其期望误差表示为 $\varepsilon_\rho(f_z^0)$, 那么算法的正确率可以表示为

$$P0 = 1 - \varepsilon_\rho(f_z^0) \tag{9}$$

对于采用基于领域归属划分的文本情感极性分类, 基于经验风险最小化原则, 从预测函数族 G 中选择最佳映射 g_z^I 能够最好地将领域分割开来, 其期望误差表示为 $\varepsilon_\rho(g_z^I)$; 从预测函数族 F 中选择最佳映射 $f_{z'}^{II}$ 能够最好地进行短文本情感分类, 其期望误差表示为 $\varepsilon_\rho(f_{z'}^{II})$. 那么算法的正确率表示为

$$P1 = [1 - \varepsilon_\rho(g_z^I)][1 - \varepsilon_\rho(f_{z'}^{II})] + \Delta \tag{10}$$

$[1 - \varepsilon_\rho(g_z^I)][1 - \varepsilon_\rho(f_{z'}^{II})]$ 表示将待处理短文本归入正确领域, 并进行正确判别的概率. Δ $(0 < \Delta \ll 1)$ 表示虽然没有将待处理短文本归入正确的领域, 却仍然正确判别的概率. 这样的可能性虽然较小, 但还存在, 所有用一个大于 0 远小于 1 的数 Δ 表示. 当 $P1 > P0$ 时, 基于领域归属划分的文本情感极性分类方法总能比一般学习方法取得更好的性能, 即:

$$1 - \varepsilon_\rho(f_z^0) \le [1 - \varepsilon_\rho(g_z^I)][1 - \varepsilon_\rho(f_{z'}^{II})] + \Delta \tag{11}$$

等价于满足条件:

$$\varepsilon_\rho(g_z^I) \le \frac{\varepsilon_\rho(f_z^0) - \varepsilon_\rho(f_{z'}^{II}) + \Delta}{1 - \varepsilon_\rho(f_{z'}^{II})} \tag{12}$$

式 (12) 说明, 第一阶段领域分类错误率满足一定条件的情况下, 改进算法性能总是优于一般算法. □

接下来分析式 (12) 条件是否容易达到. 根据定义 4 中的 Bias-variance 分解, $\varepsilon_\rho(f_z^0)$ 和 $\varepsilon_\rho(f_{z'}^{II})$ 可以分解为

$$\varepsilon_\rho(f_z^0) = \overbrace{\left[\varepsilon_\rho(f_z^0) - \varepsilon_\rho(f_F)\right]}^{\text{Sample error}} + \overbrace{\varepsilon_\rho(f_F)}^{\text{Approximation error}} \tag{13}$$

$$\varepsilon_\rho(f_{z'}^{II}) = \overbrace{\left[\varepsilon_\rho(f_{z'}^{II}) - \varepsilon_\rho(f_F^{II})\right]}^{\text{Sample error}} + \overbrace{\varepsilon_\rho(f_F^{II})}^{\text{Approximation error}} \tag{14}$$

首先, 近似误差反映了系统的学习能力, 依赖所处理的问题与预测函数空间的选择, 独立于抽样 z. 比较一般学习算法和基于领域的改进算法, 在进行情感极性判别的时候, 虽然选择相同的预测函数空间 F, 但处理问题的复杂度却不相同. 一般学习算法需要处理所有抽样数据, 其分布可能是病态或扭曲的. 然而改进算法只需要处理同一领域内的数据, 因为同一领域中的数据具有一定的相似性, 分布具有一致性, 因而处理难度远低于全部数据. 即 $\varepsilon_\rho(f_F^{II}) \ll \varepsilon_\rho(f_F)$.

其次, 抽样误差反映学习算法受抽样数据的大小和噪声的影响. 对比一般算法和基于领域的改进算法, 一般算法使用全部数据 z 进行分类, 改进算法使用分配到领域内数据 $z' \subset z$ 进行分类. 虽然改进算法可用样本数减少, 从全部数据变成领域内数据, 但这部分数据性质单一, 相当于去除了大量噪声, 同时由于域内分类难度下降, 可能仅需要少量样本也能取得与原先相同的性能. 因此可得 $[\varepsilon_\rho(f_{z'}^{II}) - \varepsilon_\rho(f_F^{II})] \le [\varepsilon_\rho(f_z^0) - \varepsilon_\rho(f_F)]$.

综上可得, $\varepsilon_\rho(f_{z'}^{II}) \le \varepsilon_\rho(f_z^0) \le 1$, 加之 $0 < \Delta \ll 1$, 那么选择适合的领域划分方法, 式 (12) 总是很容易满足. 当然随着 $\varepsilon_\rho(f_{z'}^{II})$ 的下降, $\varepsilon_\rho(g_z^I)$ 就越需要趋近于 0, 式 (12) 的成立条件就越苛刻.

基于领域归属划分的文本情感极性分类方法虽然简单直观, 却能够同时降低抽样误差和近似误差, 突破所谓的 Bias-variance 折衷, 同时给以后的学习算法设计探索了一条可能的新途径.

3 基于上下文关系自动发现的短文本情感极性分类增强方法

前面提出了基于领域归属划分的文本情感极性分类方法, 但领域是一个强概念, 比如本文所处理的短文本对象, 根据其来源不同, 分别归属于书籍评价、酒店评价、电脑评价三个领域. 但领域的概念并不是天然存在的, 和数据特性密切相关, 这就极大的限制了其应用范围. 本节中将领域的概念推广至上下文范畴, 利用短文本内在的关联关系, 重构文本上下文范畴, 从而克服短文本过于短小的先天缺憾, 提升情感极性分类的性能. 当然, 领域可以看成是一种特殊的上下文关系.

3.1 基于特殊上下文关系的话题簇情感极性判别

文本对象特征本身所固有的多义性, 面对短文本特征的稀疏性和上下文缺失的情况, 造成语义难以明辨, 理解偏差无法消解, 最终形成短文本底层特征和高层语义之间巨大的语义鸿沟. 而且由于短文本篇幅有限, 且意义零散, 无法形成统一的观点. 如果能够利用短文本内在的关联关系, 将其连同上下文重组为较长的文本, 形成一定的话题簇, 可以在一定程度上克服短文本过于短小的先天缺憾, 提升情感极性分类的性能. 由于这样的上下文关系并非自然上下文关系, 我们将其称为特殊上下文关系. 但是由于待处理的短文本缺乏明显的自然组织, 因此

如何发现文本间隐含的特殊关联关系将其重新组合, 成为一个现实的难题.

可以考虑利用短文本时间、空间等因素将分散的短文本聚合在一起. 时间关联性反映了话题的出现、流行、凋亡过程的内在规律, 例如可以利用短文本时间关联性对商品在线评价进行挖掘, 以判断消费者对商品评价随时间的变化规律. 而空间关联性反映了话题的局部性、地域性的内在规律. 例如通过空间关联性可以挖掘区域性热点和突发事件. 除此之外, 还可以考虑利用发起者关联性、会话关联性、主题关联性、热点关联性[10] 对短文本进行重组.

基于特殊上下文关系的文本情感极性分类方法如图 2 所示, 也是一个两阶段处理过程:

1) 上下文归属分类过程: 确定待处理短文本所处的上下文;

2) 上下文本情感判别过程: 用相应上下文的分类器对待处理短文进行分类.

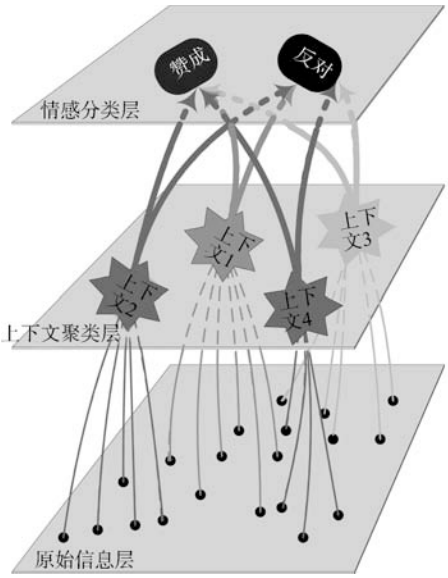

图 2　基于特殊上下文的短文本情感极性分类方法示意图

Fig. 2　Short text sentiment classification based on context reconstruction using special context

具体来说, 对于训练过程: 首先, 将原始文本信息按照文本间隐含的特殊关联关系将其重新组合, 并训练上下文分类器; 其次, 对同一上下文的短文本, 按照其情感极性的不同, 训练相应的情感分类判别器. 对于分类过程: 首先, 将待测短文本归类于适合的上下文范畴; 其次, 用相应的上下文情感分类器进行分类判别, 输出判决结果.

3.2　误差分析与性能估计

同基于领域归属划分的文本情感极性分类方法一样, 在满足一定条件的前提下, 基于特殊上下文关系的分类方法能够同时降低抽样误差和近似误差.

定义 6. 假设 X 表示短文本数据空间, Y 为情感类别空间, ρ 是定义在 $Z = X \times Y$ 上的联合概率分布函数, 给定真实数据 Z 的一个抽样 $z \subset Z$, $z = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. H, F 为预测函数族空间. 对于一般的学习方法, 基于经验风险最小化原则从 F 中选择最佳映射 f_z^0 能够最好地进行短文本情感分类, 其期望误差表示为 $\varepsilon_\rho(f_z^0)$. 对于基于特殊上下文关系的文本情感极性分类方法, 第一阶段, 即上下文归属分类过程, 基于数据 z 的上下文归属选择预测函数 $h_z^I \in H$ 能够最好地将不同上下文数据分割开来, 其期望误差表示为 $\varepsilon_\rho(h_z^I)$; 第二阶段, 即上下文范畴内情感判别过程, 基于归属于上下文的数据 $z' \subset z$ 选择域内情感判别函数 $f_{z'}^{II} \in F$ 能够将域内不同情感极性数据分割开来, 其期望误差表示为 $\varepsilon_\rho(f_{z'}^{II})$.

定理 2. 给定数据抽样 $z \subset Z$, $0 < \Delta \ll 1$, 当

$$\varepsilon_\rho(h_z^I) \leq \frac{\varepsilon_\rho(f_z^0) - \varepsilon_\rho(f_{z'}^{II}) + \Delta}{1 - \varepsilon_\rho(f_{z'}^{II})} \tag{15}$$

基于特殊上下文关系的文本情感极性分类方法总能比一般学习方法取得更好的性能.

证明. 证明方法与定理 1 的证明类似, 此处从略. □

3.3　基于遗传算法的上下文关系发现方法

相较于领域概念, 上下文范畴更为隐式和灵活. 但令人遗憾的是, 目前缺乏相关的标准短文本语料集, 现有的语料库普遍缺少重组上下文的必要信息, 诸如信息的采集时间、地点、来源等关键信息, 导致特殊上下文的重组异常困难. 从式 (8) 和 (15) 可知, 只有第一阶段的分类足够准确, 基于特殊上下文关系的两阶段分类方法才能确定良好的性能.

在上一节中, 领域的内在关联性保证了不同领域间良好可区分性, 从而保证第一阶段的域间分类足够准确. 这同时给我们一个重组上下文的重要思路提示, 即上下文范畴应该是具有较强的可分性.

因此, 本节中提出基于遗传算法的上下文关系重组方法: 通过随机分配的方式将文本组织成上下文关系, 评估不同上下文范畴间的可区分性计算适应度函数, 优胜劣汰迭代寻优, 最终得到理想的上下文关系.

基于遗传算法的上下文关系发现方法的细节如算法 2 所示.

算法 2. 基于遗传算法的上下文关系发现方法

1　初始化算法参数, 包括种群规模 ($popsize$), 选择概率 (P_s), 交叉概率 (P_c), 变异概率 (P_m), 上下文重组数目 (k);

2 $P^{\text{old}} \leftarrow$ 根据事先确定的种群规模 ($popsize$) 和上下文重组数目 (k), 根据字符串编码方式, 初始化种群 P^{old};
3 For $i = 1 : popsize$
4 根据评估不同上下文范畴间的可区分性计算适应度函数, 为 P^{old} 种群中每个个体计算适应度函数值;
5 End for;
6 Do //遗传算法进化迭代
7 $P^{\text{new}} \leftarrow \Phi$ //初始化下一代种群;
8 For $i = 1 : popsize * P_s$ //根据选择概率按比例从 P^{old} 中产生成新种群
9 $I^c \leftarrow$ 根据交叉概率 (P_c) 采用基因座单点交叉的方式进行交叉产生新个体;
10 $I^m \leftarrow$ 根据变异概率 (P_m) 采用随机变异的方式变异 I^c 产生新个体;
11 $P^{\text{new}} \leftarrow P^{\text{new}} \bigcup I^m$;
12 End for;
13 $P^{\text{new}} \leftarrow P^{\text{new}} \bigcup P^{\text{old}}$;
14 $P^{\text{old}} \leftarrow P^{\text{new}}$ 中适应度最大的 $popsize$ 个体组成下一个种群;
15 Until 算法终止条件满足.

接下来还需选择确定编码方式, 遗传算子和适应度函数.

1) 编码方式. 个体采用字符串编码方式. 假设共有 10 个短文本, $T_1, \cdots, T_{10}$, 将其重组为 3 个不同的上下文簇, 那么若由字符串编码的染色体为 (2, 2, 3, 3, 3, 3, 1, 1, 1, 2), 则它表示第 1, 2, 10 篇短文本形成一个特殊上下文, 第 3, 4, 5, 6 篇短文本同属一个特殊上下文关系, 第 7, 8, 9 篇短文本同属一个特殊上下文关系.

2) 初始种群. 采用完全随机生成的方式生成初始种群.

3) 选择算子. 采用轮盘赌的方式根据个体适应度函数值进行选择, 同时执行精英保留策略.

4) 交叉算子. 采用基因座单点交叉方式进行交叉.

5) 变异算子. 采用随机变异的方式. 以给定的变异概率, 对染色体的各个基因座实施随机变异.

6) 适应度函数. 从式 (15) 可以看出, 只有第一阶段的分类足够准确, 基于特殊上下文关系的两阶段分类方法才能确定良好的性能. 当然随着 $\varepsilon_\rho(f_{z'}^{II})$ 性能的下降, $\varepsilon_\rho(h_z^I)$ 就越需要趋近于 0, 式 (15) 的成立条件就越苛刻. 那就将上下文关系的可分性作为适应度函数. 假设共有 N 篇短文本, 形成 M 个上下文簇. 那么为这 M 个上下文簇基于 OVO (One-versus-one) 策略利用 Naive Bayes 分类器训练多类分类器. 将封闭测试的平均准确率作为适应度函数值, 这样一来, 适应度函数值大小就反映了不同上下文簇间的可分性.

3.4 样本不均衡分布下类条件概率加权调整

对于上下文范畴内的情感判别使用的是如式 (1) 所示的朴素贝叶斯决策, 其对数形式可重写为

$$H_{\text{NB}} = [\log P(c_1) - \log P(c_2)] + \left[\sum_k \log P(w_k|c_1) - \sum_k \log P(w_k|c_2)\right] \tag{16}$$

通常由于样本特征较多且分布均衡, 我们通常忽略式 (16) 中的类条件概率部分 $[\log P\{c_1\} - \log P\{c_2\}]$. 但是针对上下文内情感极性判别问题, 由于上下文关系可以任意重组, 造成正负样本分布不均衡. 加之短文本篇幅较小, 特征较少, 因此概率分布扭曲, 必须考虑类条件概率的影响.

我们将式 (16) 进一步重写为

$$H_{\text{NB}} = \lambda(a,b) + \Big[\sum_k \log P(w_k|c_1) - \sum_k \log P(w_k|c_2)\Big] \tag{17}$$

$$\lambda(a,b) = \max\left(\frac{a}{b}, \frac{b}{a}\right) \times \left[\log\left(\frac{a}{a+b}\right) - \log\left(\frac{b}{a+b}\right)\right] \times \mathrm{e}^{\left|\log\left(\frac{a}{a+b}\right) - \log\left(\frac{b}{a+b}\right)\right|} \tag{18}$$

其中, a, b 分别是不同情感极性样本的数量, $\lambda(a,b)$ 体现了对样本分布不均衡的补偿.

如式 (18) 所示, 当样本均衡时, λ 趋近为 0, 当样本分别越不均衡, 取值越大, 补偿越明显. 这样一来, 在确定上下文关系以后, 就可以根据图 2 所示的两阶段处理过程, 进行短文本情感极性判别, 这部分工作和第 2 节工作类似, 就不再赘述.

4 实验分析

文中所有实验均采用表 1 所示语料集. 实验所用评价指标包括每类的精度 (P)、召回率 (R)、准确率 (Auc) 和平均准确率. 采用 10 次 10 倍交叉验证 (10-fold cross validation), 使用了 TMG 工具箱[33] 进行了基本数据采集和清洗, SVM 算法使用的是 libsvm 工具包[34].

首先评测基于领域归属划分的文本情感极性分类方法性能. 将第 1 阶段域间归属分类过程看成是多类分类问题, 实验比较了三种不同多分类策略; 第二阶段域内情感判别过程, 为了方便比较采用同基本框架一致的 Naive Bayes 分类器.

实验结果如表 3 所示, VSM + NB, NB + NB,

表 3　基于领域归属划分的短文本情感极性分类方法性能

Table 3　Short text sentiment classification performance based on context reconstruction using field information

性能	书籍评价					酒店评价					电脑评价					
	负面评价		正面评价		AUC	负面评价		正面评价		AUC	负面评价		正面面评价		AUC	平均 AUC
方法	P	R	P	R		P	R	P	R		P	R	P	R		
VSM + NB	0.7902 ± 0.0006	0.8947 ± 0.0004	0.8784 ± 0.0005	0.7624 ± 0.0007	0.8287 ± 0.0003	0.8447 ± 0.0006	0.8084 ± 0.0007	0.8162 ± 0.0007	0.8511 ± 0.0007	0.8298 ± 0.0004	0.8062 ± 0.0007	0.9194 ± 0.0004	0.9061 ± 0.0006	0.7792 ± 0.0008	0.8493 ± 0.0003	0.8359 ± 0.0003
NB + NB	0.8129 ± 0.0005	0.9412 ± 0.0003	0.9302 ± 0.0003	0.7830 ± 0.0007	0.8623 ± 0.0003	0.9227 ± 0.0006	0.7832 ± 0.0007	0.8115 ± 0.0005	0.9340 ± 0.0005	0.8586 ± 0.0003	0.8945 ± 0.0005	0.9168 ± 0.0005	0.9148 ± 0.0005	0.8926 ± 0.0004	0.9047 ± 0.0003	0.8752 ± 0.0003
SVM + NB	0.8129 ± 0.0005	0.9412 ± 0.0003	0.9302 ± 0.0003	0.7830 ± 0.0007	0.8623 ± 0.0003	0.9227 ± 0.0006	0.7832 ± 0.0007	0.8115 ± 0.0005	0.9340 ± 0.0005	0.8586 ± 0.0003	0.8945 ± 0.0005	0.9168 ± 0.0005	0.9148 ± 0.0005	0.8926 ± 0.0004	0.9047 ± 0.0003	0.8752 ± 0.0003
PCA + KNN	0.8878 ± 0.0004	0.9241 ± 0.0005	0.9208 ± 0.0006	0.8834 ± 0.0003	0.9036 ± 0.0002	0.8210 ± 0.0007	0.8034 ± 0.0007	0.8077 ± 0.0006	0.8245 ± 0.0008	0.8141 ± 0.0004	0.7798 ± 0.0013	0.6262 ± 0.0012	0.6877 ± 0.0010	0.8235 ± 0.0008	0.7246 ± 0.0005	0.8141 ± 0.0057
CLSI + KNN	0.8764 ± 0.0005	0.9233 ± 0.0005	0.9191 ± 0.0006	0.8698 ± 0.0005	0.8965 ± 0.0002	0.8178 ± 0.0008	0.8173 ± 0.0006	0.8174 ± 0.0005	0.8180 ± 0.0007	0.8177 ± 0.0004	0.7667 ± 0.0010	0.6332 ± 0.0009	0.6875 ± 0.0009	0.8074 ± 0.0007	0.7203 ± 0.0005	0.8115 ± 0.0056
CM + KNN	0.8828 ± 0.0006	0.9283 ± 0.0006	0.9244 ± 0.0006	0.8771 ± 0.0005	0.9025 ± 0.0003	0.8176 ± 0.0007	0.8166 ± 0.0007	0.8166 ± 0.0008	0.8172 ± 0.0008	0.8169 ± 0.0003	0.7587 ± 0.0012	0.6371 ± 0.0012	0.6875 ± 0.0008	0.7978 ± 0.0007	0.7174 ± 0.0005	0.8123 ± 0.0061

SVM + NB 分别表示: 第一阶段采用基于 VSM 中心向量夹角余弦距离分类, 基于 Naive Bayes 分类器 OVO 集成分类, 基于 SVM 分类器 OVO 集成分类方法; 第二阶段采用 Naive Bayes 分类器进行域内情感判别所得到的性能指标. VSM 中心向量夹角余弦距离分类方法是将短文本利用向量空间模型 (VSM) 表示成短文本空间中的向量, 然后计算每一领域的中心向量; 那么对于待分类短文本, 通过比较其和各领域中心向量的夹角余弦距离, 将其归为最相似的一类. 基于 Naive Bayes 分类器 OVO 集成分类是采用经典的 OVO 策略[35] 进行多分类处理, 具体做法即为任意两类样本之间训练一个 Naive Bayes 分类器, 通过对未知样本进行投票的方式, 确定该未知样本的类别. Naive Bayes 分类中, 首先将中文文本分词, 去除无意义字符 (过长和过短), 将得到的词均作为特征, 不特别区分情感词. 这些词的概率依据其属于不同领域和不同极性用相应的语料进行计算. 例如有 S_1, S_2 两个领域, "+" 和 " −" 两种极性, 对 "中国" 一词来说, 依据训练语料分别计算 P (中国 $|S_1,+$), P (中国 $|S_1,-$), P (中国 $|S_2,+$) 和 P (中国 $|S_2,-$) 的概率, 然后利用多项式模型进行平滑, 最后利用 Naive Bayes 规则进行决策. 基于 SVM 分类器 OVO 集成分类方法也是为任意两类样本之间训练一个 SVM 分类器, 通过对未知样本进行投票的方式, 确定该未知样本的类别. 为了降低运算复杂度, 实验中采用的是线性核 SVM. 由于文本分类对象通常是高维属性、稀疏特征, 因此对高维特征进行降维处理也是常用的处理手段. 同时文本分类规模庞大, 对处理速度要求高, 一些没有复杂训练过程的分类方法受到广泛的青睐, 例如 Naive Bayes 和 KNN 等. 因此实验还比较了基于主成分分析 (PCA) 方法、基于聚类的潜在语义索引 (CLSI) 方法、基于聚类重心数据降维 (CM) 的方法降维, 并利用 K 最近邻 (K-nearest neighbor, KNN) 分类器进行直接一层分类的方法性能.

从表 3 的实验结果可以得到以下结论:

1) 基于领域归属划分的三种方法的平均准确性均优于使用全部语料进行分类的情况, 性能分别提升了 0.2398 %, 4.9526 %, 4.9526 %. 基于 VSM 的方法性能提升较为不明显, 可能和处理简单有关, 实验中使用的是基于简单词频 (Term frequency, TF) 的 VSM 模型, 没有进行特征选择和降维去噪.

2) NB + NB 和 SVM + NB 取得一样好的性能, 可能是因为为短文本的领域来源差异性较大, 具有较好的可分性. 因此两种方法都能将不同领域良好区隔开来, 而域内分类算法又是同一种算法, 所以两种算法能够取得一样好的性能. 但是需要注意的是, 使用 OVO 策略进行多分类, 对于 k 个类别需要训练 $k*(k-1)/2$ 个分类器. 显然训练 Naive Bayes 分类器的代价远低于建立 SVM 分类器的代价. 这也是后续在上下文实验中选用 Naive Bayes 分类器进行 OVO 集成的重要原因. 最引人注目的是, NB + NB 和 SVM + NB 方法在单个领域的分类精度逼近甚至超过使用同源语料进行分类的结果.

这充分体现了两阶段处理方法的优越性.

3) 目前流行的基于降维数据利用 KNN 分类的三种方法性能相似, 均劣于基于领域归属划分的方法. 且分类结果显示出较大的分类倾向性: 在书籍评价类中的性能较高, 而在电脑评价类中的性能较低. 表 3 中的 PCA 的降维数为 400, CLSI 方法维数为 300, CM 方法维数为 200, KNN 分类中近邻数选择为 2. 选择依据是通过比较不同降维维度的分类准确率得到的.

4) 我们比较了 CLSI + KNN, CM + KNN, PCA + KNN 方法在降维到不同维度 (10 ∼ 1 000) 时不同分类方法的分类准确率. KNN 分类方法近邻数选择为 2. 实验取 10 折平均, 结果如图 3 所示, 图中曲线中心点代表均值, 上下位横线代表取值波动范围. 三种方法得到相似的分类性能, 当降维数目大于 100 后, 分类性能稳定, 但随着降维数目的增加, 呈现逐渐下降的趋势. PCA, CLSI, CM 方法在分别降维到 400, 300, 200 维时, 取得了最好的分类性能. 这也是我们在表 3 中参数选择的依据.

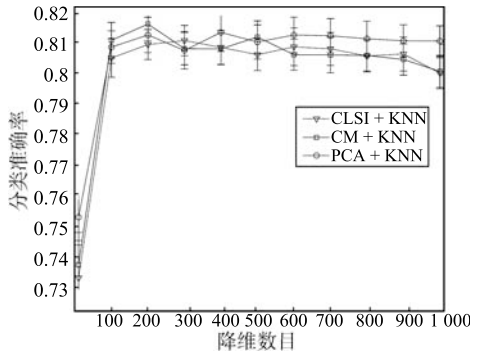

图 3 降维到不同维度 (10 ∼ 1 000) 时不同分类方法的分类准确率比较示意图

Fig. 3 Performance comparison of different classification methods in 10 ∼ 1 000 dimensions

接下来评测基于上下文关系的文本情感极性分类方法性能. 第一阶段即不同上下文间分类问题看成是多类分类问题, 首先基于遗传算法重组上下文关系, 然后基于 Naive Bayes 分类器 OVO 集成分类, 实验比较了不同上下文聚类数目的性能; 第二阶段即上下文内情感判别过程, 为了方便比较采用同基本框架一致的 Naive Bayes 分类器. 但是由于第一阶段所处理对象的规模庞大, 共有 12 000 篇文本, 如果重组为 k 个上下文簇, 算法需要在 $k^{12\,000}$ 规模的解空间中迭代寻优. 考虑到现有的计算能力, 兼顾性能和效率, 实验中不得不做出一些折衷. 我们先将 12 000 篇文本用 K-means 方法聚为 30 个类, 在对这 30 个类进行重组. 这样解空间的规模简化为 k^{30}, 这大大简化了我们的工作. 遗传算法其他参数设置如下: 种群数量为 100; 进化代数为 10; 编码长度为 30; 选择算子为轮盘赌算子, 概率为 0.9; 交叉算子为单点交叉, 概率为 0.9; 变异概率为 0.1; 将上下文关系的可分性作为适应度函数: 即用划分得到的不同上下文当作不同的类别训练相应的多类分类器 (基于 OVO 策略的 Naive Bayes 分类器), 将封闭测试的平均准确率作为适应度函数值. 这样一来, 适应度函数值的大小就反映了不同上下文簇间的可分性.

实验结果如表 4 所示, $k = 3, 4, 5$ 表示利用遗传算法将上下文划分 3, 4, 5 个不同的簇. 由于文本分类问题规模庞大, 随着聚类数目的增加, 算法复杂度急剧上升, 因此只比较了三个不同聚类数目. 从实验结果可以得到以下结论:

1) 三种 k 取值下的平均准确性均优于使用全部语料进行分类的情况, 性能分别提升了 4.8687 %, 3.3817 %, 2.3504 %. 由于利用遗传算法迭代寻优是以平均准确率为适应度函数, 因此只考虑了平均准确率的提高, 如果对某些类别有特殊要求, 可对其加权, 简单改进适应度函数即可实现.

2) 随着 k 取值的增加, 算法性能逐渐下降, 且性

表 4 基于特殊上下文关系的文本情感极性分类方法性能

Table 4 Short text sentiment classification based on context reconstruction using special context

性能	书籍评价					酒店评价					电脑评价					
	负面评价		正面评价		AUC	负面评价		正面评价		AUC	负面评价		正面评价		AUC	平均 AUC
方法	P	R	P	R		P	R	P	R		P	R	P	R		
特殊	0.8604	0.8911	0.8900	0.8502	0.8709	0.9082	0.8269	0.8422	0.9150	0.8711	0.8973	0.8676	0.8755	0.8955	0.8816	0.8745
上下文	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±
$k=3$	0.0020	0.0025	0.0016	0.0038	0.0004	0.0009	0.0017	0.0012	0.0012	0.0004	0.0025	0.0035	0.0021	0.0056	0.0011	0.0006
特殊	0.8383	0.8711	0.8706	0.8236	0.8467	0.9024	0.8108	0.8311	0.9107	0.8612	0.9024	0.8526	0.8641	0.9046	0.8784	0.8621
上下文	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±
$k=4$	0.0044	0.0056	0.0030	0.0077	0.0012	0.0012	0.0039	0.0019	0.0017	0.0008	0.0015	0.0051	0.0032	0.0028	0.0009	0.0012
特殊	0.8363	0.8629	0.8664	0.8207	0.8419	0.8944	0.7746	0.8040	0.9075	0.8412	0.8946	0.8620	0.8704	0.8934	0.8776	0.8535
上下文	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±	±
$k=5$	0.0039	0.0106	0.0040	0.0089	0.0018	0.0017	0.0045	0.0022	0.0018	0.0012	0.0025	0.0039	0.0024	0.0053	0.0010	0.0016

能波动加剧. 一方面是因为数据来源于三个明显不同的类别, 因此将其划分为三个不同的上下文范畴是符合其自然属性的; 另外一方面, 随着 k 取值的增加, 其优化难度也在增加. 但实验中对于不同的 k 取值采用了相同的遗传参数, 包括相同的种群数目、进化代数等等. 因此在一定程度上制约了其收敛程度, 然而出于性能和效率的综合考虑, 这样的损失是可接受的.

3) 基于上下文关系的情感判别方法性能稍逊于基于领域归属划分的情感判别方法. 这一方面可能是因为领域的概念较强且结构清晰, 而上下文的概念更为隐式和内在; 另外一方面也可能是因为基于遗传算法进行上下文重组的时候, 虽然希望是任意上下文重组, 但出于性能和效率的综合考虑, 进行了一定程度的折中所致. 但是考虑到领域的概念并不是天然存在的, 而上下文关系可以根据我们的需求任意重组, 因此有更广泛的应用前景.

4) 我们分析了基于遗传算法所发现的上下文之间的关系和领域之间的关系. 假设对短文本语料, 按照其领域划分成 m 个簇 $C_1, C_2, \cdots, C_m$. 通过遗传算法重组为 n 个簇 $S_1, S_2, \cdots, S_n$. 那么上下文和领域间的相似度定义为

$$S^{\text{smility}} = \frac{1}{n}\sum_{i=1}^{n}\max_{j=1,\cdots,m}\left(\frac{S_i \bigcap C_j}{|C_j|}\right) \tag{19}$$

其中, $S_i \bigcap C_j$ 表示上下文 S_i 和领域 C_j 共同的样本数, $|C_j|$ 表示领域 C_j 的样本数.

图 4 分别给出了使用遗传算法重组上下文, 聚类数目为 3～11 时, 每代进化种群最优个体的相似度值变化规律. 为了细化比较, 我们的领域划分采用两种方式: 单一领域方式是将短文本领域按电脑评价、酒店评价、书籍评价进行领域划分: 单一情感方式是将短文本按电脑评价 (正面)、电脑评价 (负面)、酒店评价 (正面)、酒店评价 (负面)、书籍评价 (正面)、书籍评价 (负面) 进行领域划分.

图中曲线中心点代表均值, 上下位横线代表取值波动范围. 从中可知, 当上下文重组数目等于 3, 上下文关系和领域相似度最大, 接近于文本的自然划分. 随着上下文重组数目的增加, 相似性逐渐降低. 从图中还可以看出, 比较单一领域和单一情感相似度, 单一情感相似度重要略高于单一领域相似度. 这说明现重组的上下文中以单一情感文本的为主. 这也反映了上下文范畴内文本极性分布严重不均衡, 因此基于类条件概率的加权调整是有必要进行的.

5　结论与展望

面对短文本处理难题, 本文试图利用短文本内在的关联关系, 重构文本上下文范畴, 从而克服短文本过于短小的先天缺憾, 提升情感极性分类的性能. 理论分析表明, 满足一定条件的前提下, 基于特殊上下文的分类方法能够同时降低抽样误差和近似误差. 真实数据集上的实验结果也印证了理论分析的结果.

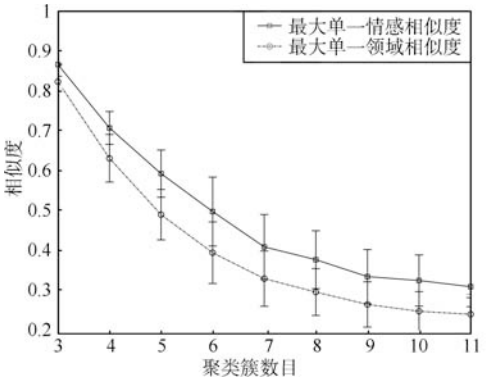

图 4　不同上下文聚类簇数目和领域相似度关系图

Fig. 4　Similarity between fields and clustering generated by GAs

研究初步理清了特殊上下文作用的机制, 但随着研究的深入进行, 一些新的问题摆到了我们的面前, 其潜在的影响仍需评估, 接下来的工作将围绕以下几点展开:

1) 本文基于遗传算法进行上下文发现, 期望依据给定的适应度函数进行任意上下文关系组合. 但在实验中发现, 由于所处理对象的规模庞大, 对于 k 个上下文簇, 算法需要在 $k^{12\,000}$ 规模的解空间中迭代寻优. 考虑到现有的计算能力, 兼顾性能和效率, 实验中不得不做出一些折中. 但我们仍然希望能找到更好的方法处理这样的问题, 这也是下一步需要解决的问题之一.

2) 本文以提升分类精度的为目的重组上下文, 但是短文本信息处理并不只是分类一个目的. 上下文范畴是按照文本的内在关联关系重组, 组织隐含高层语义, 是一种有意义的团簇, 能够帮助我们更深入的理解文本意义. 但如何发现和利用这样的隐含意义, 是一个值得研究的方向.

3) 实验中也发现, 短文本预处理部分, 包括分词算法的性能以及特征选择的方法, 对算法的性能有一定的影响, 会造成算法性能的波动. 但这些因素并不影响问题的实质, 只会造成细微的差别.

4) 值得注意的是, 本文所提出的两阶段分类策略, 虽然简单却探索了一条设计学习算法的思路. 对于任意学习算法, 都可以利用两阶段学习策略, 首先确定领域, 再进行域内处理, 从而提升算法性能. 算法同时降低抽样误差和近似误差. 当然前提是第一阶段的分类要足够准确.

除此之外, 由于待处理的短文本缺乏明显的自

然组织, 因此如何发现文本间隐含的关联关系将其重新组合, 即基于时间、空间等因素的特殊上下文觉察技术成为提升情感极性分类性能的关键. 下一步我们将研究如何依据短文本间真正的内蕴关系, 重建上下文关系, 但这部分工作受制于语料库的建设. 因此着手建立标准短文本内容信息数据资源库, 并进行精细整理和标注, 是今后工作开展的重要前提.

References

1 Picard R W. *Affective Computing*. Cambridge: MIT Press, 1997

2 Wayne C. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Greece: ELRA, 2000. 1487−1494

3 Finn A, Kushmerick N. Learning to classify documents according to genre: special topic section on computational analysis of style. *Journal of the American Society for Information Science and Technology*, 2006, **57**(11): 1506−1518

4 Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 2006, **22**(2): 110−125

5 Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, USA: ACL, 2005. 115−124

6 Osman D J, Yearwood J L. Opinion search in web logs. In: Proceedings of the 18th Conference on Australasian Database. Ballarat, Australia: ACS, 2007. 133−139

7 Liao Xiang-Wen, Cao Dong-Lin, Fang Bin-Xing, Xu Hong-Bo, Cheng Xue-Qi. Research on blog opinion retrieval based on probabilistic inference model. *Journal of Computer Research and Development*, 2009, **46**(9): 1530−1536
(廖祥文, 曹冬林, 方滨兴, 许洪波, 程学旗. 基于概率推理模型的博客倾向性检索研究. 计算机研究与发展, 2009, **46**(9): 1530−1536)

8 Zhou Li-Zhu, He Yu-Kai, Wang Jian-Yong. Survey on research of sentiment analysis. *Journal of Computer Applications*, 2008, **28**(1): 2725−2728
(周立柱, 贺宇凯, 王建勇. 情感分析研究综述. 计算机应用, 2008, **28**(1): 2725−2728)

9 Zhu Yan-Lan, Min Jin, Zhou Ya-Qian, Huang Xuan-Jing, Wu Li-De. Semantic orientation computing based on HowNet. *Journal of Chinese Information Processing*, 2006, **20**(1): 14−20
(朱嫣岚, 闵锦, 周雅倩, 黄萱菁, 吴立德. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006, **20**(1): 14−20)

10 Gong Cai-Chun. Research on Short Text Language Computing [Ph. D. dissertation], Institute of Computing Technology, Chinese Academy of Sciences, China, 2008
(龚才春. 短文本语言计算的关键技术研究 [博士学位论文], 中国科学院研究生院 (计算技术研究所), 中国, 2008)

11 Zeimpekis D, Gallopoulos E. Linear and non-linear dimensional reduction via class representatives for text classification. In: Proceedings of the 6th IEEE International Conference on Data Mining. Hong Kong, China: IEEE, 2006. 1172−1177

12 Park H, Jeon M, Rosen J B. Lower dimensional representation of text data based on centroids and least squares. *Bit Numerical Mathematics*, 2003, **43**(2): 427−448

13 Xu W R, Liu D X, Guo J, Cai Y C, Hu R L. Supervised dual-PLSA for personalized SMS filtering. In: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology. Sapporo, Japan: Springer-Verlag, 2009. 254−264

14 Mørch A I, Cheung W, Wong K, Liu J, Lee C, Lam M, et al. Grounding collaborative knowledge building in semantics-based critiquing. In: Proceedings of the 4th International Conference on Advances in Web-based Learning. Hong Kong, China: Springer-Verlag, 2005. 244−255

15 Wang L, Jia Y, Han W. Instant message clustering based on extended vector space model. In: Proceedings of the 2nd International Symposium on Advances in Computation and Intelligence. Wuhan, China: Springer-Verlag, 2007. 435−443

16 Adams P H, Martell C H. Topic detection and extraction in chat. In: Proceedings of the IEEE International Conference on Semantic Computing. Santa Clara, USA: IEEE, 2008. 581−588

17 Fan X, Hu H. A new model for Chinese short-text classification considering feature extension. In: Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence. Sanya, China: IEEE, 2010. 7−11

18 Yan Rui, Cao Xian-Bin, Li Kai. Dynamic assembly classification algorithm for short text. *Acta Electronica Sinica*, 2009, **37**(5): 1019−1024
(闫瑞, 曹先彬, 李凯. 面向短文本的动态组合分类算法. 电子学报, 2009, **37**(5): 1019−1024)

19 O′Shea J, Bandar Z, Crockett K, McLean D. A comparative study of two short text semantic similarity measures. In: Proceedings of the 2nd KES International Symposium on Agent and Multi-agent Systems: Technologies and Applications. Incheon, Korea: Springer-Verlag, 2008. 172−181

20 Yang Feng, Peng Qin-Ke, Xu Tao. Sentiment classification for online comments based on random network theory. *Acta Automatica Sinica*, 2010, **36**(6): 837−844
(杨锋, 彭勤科, 徐涛. 基于随机网络的在线评论情绪倾向性分类. 自动化学报, 2010, **36**(6): 837−844)

21 Tan S B, Cheng X Q, Wang Y F, Xu H B. Adapting naive Bayes to domain adaptation for sentiment analysis. In: Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval. Toulouse, France: Springer-Verlag, 2009. 337−349

22 Kong E B, Dietterich T G. Error-correcting output coding corrects bias and variance. In: Proceedings of the 12th International Conference on Machine Learning. Tahoe City, USA: Morgan Kaufmann, 1995. 313−321

23 Breiman L. Bias, Variance and Arcing Classifiers, Technical Report 460, Department of Statistics, University of California at Berkeley, USA, 1996

24 Tibshirani R. Bias, Variance and Prediction Error for Classification rules, Technical Report No. 9602, Department of Statistics, University of Toronto, Canada, 1996

25 Kohavi R, Wolpert D H. Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the 13th International Conference on Machine Learning. Bari, Italy: Morgan Kaufmann, 1996. 275−283

26 Friedman J H. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1997, **1**(1): 55−77

27 Domingos P. A unified bias-variance decomposition for zero-one and squared loss. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. Austin, USA: AAAI, 2000. 564−569

28 Zhou Z H, Jiang Y. NeC4.5: neural ensemble based C4.5. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **16**(6): 770−773

29 Cucker F, Smale S. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 2002, **39**(1): 1−49

30 Niyogi P. *The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar*. Norwell: Kluwer Academic Publishers, 1997

31 Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995

32 Poggio T, Rifkin R, Mukherjee S and Niyogi P. General conditions for predictivity in learning theory. *Nature*, 2004, **428**: 419−422

33 Zeimpekis D, Gallopoulos E. TMG: A Matlab toolbox for generating term-document matrices from text collections. *Grouping Multidimensional Data: Recent Advances in Clustering*. Berlin: Springer-Verlag, 2006. 187−210

34 Chang C C, Lin C J. LIBSVM: A library for support vector machines [Online], available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, Nov 20, 2011

35 Knerr S, Personnaz L, Dreyfus G. Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Neurocomputing: Algorithms, Architectures and Applications*. Berlin: Springer-Verlag, 1990. 41−50

杨 震 北京工业大学计算机学院副教授. 主要研究方向为信息内容安全, 网络舆情分析, 可信计算. 本文通信作者.
E-mail: yangzhen@bjut.edu.cn
(**YANG Zhen** Associate professor at the College of Computer Science, Beijing University of Technology. His research interest covers information content security, public opinion analysis, and trusted computing. Corresponding author of this paper.)

赖英旭 北京工业大学计算机学院副教授. 主要研究方向为网络安全, 可信计算.
E-mail: laiyingxu@bjut.edu.cn
(**LAI Ying-Xu** Associate professor at the College of Computer Science, Beijing University of Technology. Her research interest covers information network security and trusted computing.)

段立娟 北京工业大学计算机学院副教授. 主要研究方向为网络与信息安全, 内容检索与内容安全.
E-mail: ljduan@bjut.edu.cn
(**DUAN Li-Juan** Associate professor at the College of Computer Science, Beijing University of Technology. Her research interest covers network security, information retrieval, and content security.)

李玉鑑 北京工业大学计算机学院教授. 主要研究方向为模式分析与机器智能.
E-mail: liyujian@bjut.edu.cn
(**LI Yu-Jian** Professor at the College of Computer Science, Beijing University of Technology. His research interest covers pattern analysis and machine intelligence.)