

基于粗糙集与信息增益的情感特征选择方法

蒲国林

(四川文理学院 计算机学院, 四川 达州 635000)

摘 要: 为了提高情感特征提取的准确率, 为高性能情感分析打下坚实的基础, 提出了一种融合粗糙集与信息增益的情感特征选择方法. 该方法借助信息增益判据选出高相关性的特征子集, 再通过粗糙集剔除高冗余性的特征, 从而得到最优的特征子集. 在多个数据集上的测试表明, 该方法可将若干经典方法的准确率提高 4~9 个百分点, 是一种优秀的特征选择方法, 对提升情感分析的整体性能有明显意义.

关键词: 情感分析; 特征选择; 粗糙集; 信息增益

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)01-0096-04

A Sentiment Feature Selection Method Based on Rough Set and Information Gain

PU Guo-lin

(School of Computer Science, Sichuan University of Arts and Science, Dazhou 635000, China)

Abstract: A Rough Set and Information Gain based on sentiment feature selection method is proposed for building a solid foundation in sentiment analysis. The novel method firstly uses Information Gain to select a feature subset which has high relativity with the class attribute. Secondly, the features which have high redundancy will be eliminated by Rough Set. Experimental results on several datasets reveal the method makes accuracy increase 4-9 percentages than other methods. It is an outstanding feature selection method and has significance in sentiment analysis.

Key words: sentiment analysis; feature selection; Rough Set; Information Gain

1 引言

互联网沉淀的海量用户发言信息, 蕴含着丰富的说话人主观态度和情感. 这些情感对商业部门的产品调查及市场预测、公共管理部门的舆情分析、安全部门的反恐维稳等方面都有巨大价值. 因此, 文本的情感倾向性分析逐渐成为机器学习领域的热门课题.

情感特征选择是指从文档中选择最能表达说话人情感的字、词或词组等特征的过程. 与文档的主题挖掘或者话题挖掘不同, 情感分析的目标不是分析文本的客观内容, 而是发现文本所体现的说话人的

主观感情. 在客观内容分析领域, 名词和名词性短语在特征项中贡献巨大^[1], 而在主观感情分析领域, 特征的选择更加复杂. 文本的情感特征选择是情感分析性能的基石^[2], 成为情感挖掘领域最活跃的方向之一.

文献[3]提出了一种基于词法模型的情感特征选择算法, 虽然实验表明该算法有较高的性能, 然而通用性有限. LDASA 是一种基于隐含狄利克雷分配(Latent Dirichlet Allocation, LDA)模型的文本情感特征提取方法. 实现了完全自适应的文本情感特征提取^[4], 但该方法的时间复杂度较高. 文献[5]提出了一种基于特征划分的情感分析方法, 该方法

收稿日期: 2015-03-17; **修回日期:** 2015-05-22

基金项目: 国家自然科学基金(61152003); 国家档案局科技计划项目(2014-X-65); 四川省教育厅科研项目(15ZB0323); 四川文理学院智能计算与物联网工程技术中心资助项目

的重点是对特征的划分,而非降维.文献[6]通过构建基于期望最大化的特征选择函数过滤大量无关特征^[6],取得了一定的成果.

2 基于粗糙集的文本分析

2.1 粗糙集定义

粗糙集理论是一种通过删除冗余条件属性来实现属性约简的数学工具.粗糙集通过一对称作上近似和下近似的精确概念对来近似模糊概念集^[7].

设问题空间信息系统 U 可以被表示为式(1)的形式:

$$U = (W, C, X, h) \tag{1}$$

式中, W 是目标对象集合,也叫做论域, C 是属性集. X 的定义满足式(2).

$$X = \bigcup_{a \in C} X_a \tag{2}$$

式中, X_a 是属性 a 的范围, $h: W \times C \rightarrow X$ 是为论域 W 中的每个目标 z 分配属性值的信息方程,即通过函数 h 可得到式(3).

$$z \in W, c \in C \Rightarrow h(z, c) \in X_c \tag{3}$$

设 ZW 是一组目标, DC 是一组属性,则关于 Z 的属性集 D 的下近似 $\underline{D}(Z)$ 和上近似 $\bar{D}(Z)$ 可通过式(4)、式(5)定义.

$$\underline{D}(Z) = \{z \mid z \in W, [z]_D \subseteq Z\} \tag{4}$$

$$\bar{D}(Z) = \{z \mid z \in W, [z]_D \cap Z \neq \Phi\} \tag{5}$$

目标集合 ZW 称为概念,高于上近似的区域称为负域,上下近似之间称为边界,低于下近似的区域称为正域.

2.2 决策表约简

若属性 C 可以被划分为条件属性 E 和决策属性 F ,并满足式(6),则信息系统可以被称为决策系统或决策表.

$$\begin{aligned} E \cup F &= C \\ E \cap F &= \Phi \end{aligned} \tag{6}$$

通常 F 中只含有一个决策属性 $\{f\}$.关于 D 的决策属性 f 的正域可通过式(7)计算.

$$POS_D(f) = \bigcup_{z \in W / IND(\{f\})} DZ \tag{7}$$

式中, $IND(\{f\})$ 是一个不可区分等价关系,其定义满足式(8).

$$IND(f) = \{(x, y) : x, y \in W, h(x, f) = h(y, f)\} \tag{8}$$

将粗糙集理论应用于知识获取时通常需要学习样例数据.因此决策表的定义就很重要.与知识表类似,决策表包含大量的样例数据记录,此外它还包含

知识表所不具有的决策属性.决策属性 f 的正域代表了关于 f 的近似质量.当决策时,并非 C 中的所有属性都是必须的,通过学习可以发现哪些属性是重要的,哪些属性是冗余的.若决策属性集为 F ,则正域的划分可通过式(9)实现.

$$POS_C(F) = \bigcup_{Z \in W/F} CZ \tag{9}$$

属性集 C 和决策集 F 的依赖度可定义如式(10).

$$\delta_C(F) = \frac{|POS_C(F)|}{|W|} \tag{10}$$

对于任意的 $x \in F$,其对于决策属性集 F 的重要性可通过式(11)定义.

$$POS_C(F) = \bigcup_{Z \in W/F} CZ \tag{11}$$

通过以上步骤,可以通过属性对决策的重要性来约简决策表.

2.3 文本特征建模

设给定文档集合为 $D = \{d_1, d_2, \dots, d_n\}$,其特征集合为 $S = \{s_1, s_2, \dots, s_m\}$.在特征空间中文档被用特征表示为向量,各维度的值为 0 或者 1,1 代表该特征在文档中出现,为 0 代表不出现.若以文档集 D 作为论域,候选特征集 T 作为候选属性集 E ,以文本的情感类别集合 $C = \{c_1, c_2, \dots, c_k\}$ 作为决策属性集 F ,则文本情感分类问题的决策表可以构建如表 1 所示.

表 1 文本情感分类问题的决策表

D	T				C
	t_1	t_2	$\dots$	t_m	
d_1	1	1	$\dots$	0	c_1
d_2	0	1	$\dots$	0	c_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
d_n	1	0	$\dots$	1	c_k

在文档集的向量空间中,决策表具有以下三个重要性质:

(1)候选属性集规模庞大,也就是说文档集的特征维度是非常高的;

(2)稀疏性,也就是说每个特征在具体某篇文档(决策表的某一行)中出现的概率很低;

(3)特征不均衡,某个特征在不同文档中出现的次数差异巨大.

以上三点特性,使直接基于决策表的情感分类或其他应用很困难,也使得传统的决策表约简方法效率低下,需要新的降维方式.

3 基于信息增益的决策表约简

信息增益 (Information Gain, IG) 是一种经典的特征选择方法, 它通过衡量包含与不包含某特征式特征集的信息量变化来选择前 k 个特征作为文本的特征集合。

某个属性是否有用, 同时取决于它的相关性和冗余性。如果能够通过某个特征预测到类别, 则认为它是相关的, 否则就是不相关的。如果某个特征和其他特征相关, 则认为它是冗余的。

单纯基于信息增益的特征选择的不足是需要预先设置阈值, 并且当文档集规模庞大时, 信息增益选择出的特征集合也比较庞大。基于特征集的方法可以排除大量无关属性和噪声特征, 从而大大降低特征集的维度。然而单纯基于粗糙集的方法也有不足。首先, 得到最优消减是一个 NP 难题, 只能通过一些启发式的算法近似求解; 其次, 单纯基于粗糙集的方法, 计算开销往往很大^[7]。因此, 可以通过结合信息增益与粗糙集实现一种高降维率、低时间开销的情感特征选择算法。

混合算法的执行大概分为两步。首先计算每个特征的信息增益值, 并且选择出所有信息增益大于 0 的特征项。经过这一步, 就可以初步排除粗糙集属性表中无关或噪声特征, 从而大大降低了后续约简的计算开销。此外, 初步排除之后的特征集将交给粗糙集属性约简算法, 通过启发式运算, 得到最优的特征解。

本文提出的基于信息增益和粗糙集的情感特征选择方法在文中称为 IG-RSR (Information Gain-Rough Set Reduction) 算法。

4 实验

4.1 实验数据

为了测试本文提出情感特征选择方法的性能, 我们爬取了 1 万篇 IMDB 的影评文档、1 万篇 Amazon 的书评文档, 并下载了 Datatang 提供的 2 万篇去除隐私信息的带标注微博短文本文档。数据经过正则表达式过滤排除了 HTML 标签等噪声数据作为实验的训练和测试语料。图 1 所示为实验语料示例。

4.2 直接降维效果测试

文档集降维之后的维度, 是特征选择算法最直观的衡量标准。在此测试了本文提出算法和现有的基于粗糙集的特征选择算法 RSFD^[8]、基于信息增

评论者  于 2011 年 6 月 2 日
版本: 精装
距离第一次看这本书已经十二年了, 仍然记得初读时的震撼。书上的每一个字都认识, 但是这样的小说前所未见, 马尔克斯描述的那个世界一会儿真实得触手可及, 一会儿又像史前神话一样不可思议。后来我知道了一个专有名词: 魔幻现实主义。这版换了译者, 已读了几十页, 文字功底扎实, 不输黄锦炎 的译本, 可以放心了。

图 1 情感分析语料示例

益的算法 IG-UTS^[9] 以及热门的 LDASA 算法对各数据集进行情感特征选择之后的特征维度。实验结果如表 2 所示。

表 2 直接降维效果对比

数据 集	IMDB 影评数据集	Amazon 书评数据集	微博数据集
RSFD	397	254	485
IG-UTS	738	501	880
LDASA	601	576	712
IG-RSR	314	209	444

如表 2 所示, 基于粗糙集的特征选择方法的降维效率明显高于基于信息增益的特征选择方法, 其降维效率优于基于 LDA 模型的方法。而将信息增益引入粗糙集理论, 可以进一步提高原粗糙集情感特征选择方法的降维能力, 从而使本文提出的 IG-RSR 算法具有实验对照组中最高降维效率。

4.3 间接降维效果测试

对于情感分析而言, 更大的降维率意味着更低的计算复杂度。然而这些被选择的特征, 是否最能代表说话人的情感倾向性? 这一问题无法通过降维的结果直接衡量, 或者说直接衡量所需的人工开销是巨大的。在此, 可以通过相同情感分类算法基于不同特征选择方法的分类准确率来判断特征选择是否保留了对情感分析最有价值的特征。

在此我们选择基于朴素贝叶斯的算法^[10]、基于支持向量机的算法^[11] 以及基于 AdaBoost 的算法^[12], 在经过表 1 所述各种降维方法进行特征选择之后的情感分类准确率。实验结果如表 3 所示。

表 3 基于不同降维方法的分类结果对比

分类 算法	NBTC	B-SVM	IMC- AdaBoost
RSFD	0.769	0.843	0.853
IG-UTS	0.772	0.839	0.829
LDASA	0.794	0.825	0.837
IG-RSR	0.832	0.896	0.904

如表 3 所示,无论是朴素贝叶斯类方法、支持向量机类算法,还是 Boosting 算法,以本文提出的 IG-RSR 作为特征选择工具之后的分类准确率都高于基于其他降维算法之后的分类准确率.证明了本文提出的基于粗糙集与信息增益的情感特征选择方法能够有效选择对情感分类最有价值的特征、排除无关特征.

5 结束语

情感分析是当前互联网数据挖掘和机器学习等相关领域的热门问题. 在各类情感分析任务中,选择适当的情感特征、排除无关的非情感特征是关乎应用成败的基本技术.

现有情感特征选择方法存在降维效率有限或选择结果不理想等问题. 本文将信息增益判据引入粗糙集理论,提出了一种基于信息增益筛选的粗糙集属性约简情感特征选择方法. 该方法通过粗糙集决策表对文档集进行建模,并通过信息增益判据进一步优化决策表属性约简的过程,实现了对情感特征更精准的选择. 实验结果表明,本文具有更高的降维效率,并且基于本文降维结果的深入分析,具有更高的精度.

随着移动互联网与智能终端的发展,移动社交网络应用已渗透到社会各个群体,并且其份额逐渐压倒传统基于 PC 的社交网络. 对于移动应用,流量和时延是两大重要因素. 在移动互联网上发表的用户主观倾向性,为了快速性,包含了大量语音、图像等元素,为了节约流量和进行快速交流,包含了大量缩略语和短文本. 对于包含多种数据格式的混杂的短文本的情感分析,特别是这一类文本的快速分析,对未来这一领域的应用具有重要价值.

参考文献:

- [1] 殷春霞, 彭勤科. 利用复杂网络为自由评论鉴定词汇情感倾向性[J]. 自动化学报, 2012, 38(3): 389-398.
- [2] Bollegala D, Weir D, Carroll J. Cross-domain sentiment classification using a sentiment sensitive thesaurus[J]. IEEE Transactions on Knowledge and Data

Engineering, 2013, 25(8): 1719-1731.

- [3] Isa Maks, Piek Vossen. A lexicon model for deep sentiment analysis and opinion mining application[J]. 2012, 53(4): 680-688.
- [4] Shams M, Shakey A, Faili H. A non-parametric LDA-based induction method for sentiment analysis[C]// 16 th CSI International Symposium on Artificial Intelligence and Signal Processing, Shiraz, Fars; IEEE, 2012: 216-221.
- [5] 李素科, 蒋严冰. 基于情感特征聚类的半监督情感分类[J]. 计算机研究与发展, 2013, 50(3): 2570-2577.
- [6] 文翰. 面向信息检索的 Web 文本挖掘方法研究[D]. 广州: 华南理工大学, 2012.
- [7] Haoyang Wu, Yuyuan Wu, Jinping Luo. An interval type-2 fuzzy rough set model for attribute reduction [J]. IEEE Transactions on Fuzzy Systems, 2012, 17(2): 301 - 315.
- [8] You-Shyang Chen, Ching-Hsue Cheng. A soft-computing based rough sets classifier for classifying IPO returns in the financial markets[J]. Applied Soft Computing, 2012, 12(1): 462-475.
- [9] Mirzakhani M, Moghadam A E. Two layer algorithm for data classification based on rough set and bayesian network classifiers[C]// 13th Iranian Conference on Fuzzy Systems (IFSC). Qazvin, Iran; IEEE, 2013: 1-6.
- [10] Troussas C, Virvou M, Junshean Espinosa K, et al. Twitter news classification: theoretical and practical comparison of SVM against naive bayes algorithms [C]// Fourth International Conference on Information, Intelligence, Systems and Applications. Piraeus, Greece; IEEE, 2013: 13-18.
- [11] Mordelet F, Vert J P. A bagging SVM to learn from positive and unlabeled examples[J]. Pattern Recognition Letters, 2014, 37(1): 201-209.
- [12] 伍杰华, 倪振声. 改进多分类器集成 AdaBoost 算法的 Web 主题分类[J]. 计算机应用与软件, 2013, 30(11): 273-279.

作者简介:

蒲国林 男, (1971-), 博士, 副教授. 研究方向为人工智能.
E-mail: puguols@126.com