doi:10.3969/j.issn.1673-4785.2009.05.011

基于粗糙集的文本分类特征选择算法

张志飞[1,2],苗夺谦[1,2]
(1.同济大学 计算机科学与技术系,上海 201804;2.同济大学 嵌入式系统与服务计算教育部重点实验室,上海 201804)

摘 要:文本分类是根据未知文本的内容将其划分到一个或多个预先定义的类别的过程,是许多基于内容的信息管理任务的重要组成部分.文本分类问题的难点是特征空间的高维性,通常采用特征选择作为降维的重要方法.将属性约简和文本分类的特点相结合,提出了一种基于粗糙集的特征选择算法即改进的快速约简算法.实验表明该算法是有效的,不仅可以降低特征空间的维度,而且能够维持高精度.

关键词:文本分类;粗糙集;特征选择;快速约简
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-4785(2009)05-0453-05

Feature selection for text categorization based on rough set

ZHANG Zhi-fei[1,2], MIAO Duo-qian[1,2]
(1. Department of Computer Science and Technology, Tongji University, Shanghai 201804, China; 2. The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804, China)

Abstract: Text categorization assigns text documents to one or more predefined categories based on their contents. This assists content-based information management. A difficult problem in this task is the high dimensionality of the feature space. To resolve this, a feature selection method was employed to reduce the dimensions. A new approach based on rough sets, that we call it the improved quick reduction (IQR) algorithm, was proposed. It involved both attribute reduction and text categorization. The experimental results demonstrated the effectiveness of the proposed algorithm. It reduced the dimensionality of feature space, while maintaining high accuracy.

Keywords: text categorization; rough set; feature selection; quick reduction

自20世纪80年代以来,信息化的浪潮席卷全球."信息爆炸"虽然提供了丰富多彩的信息资源,但是限制了人们有效地获取信息的能力.文本分类是根据给定文本的内容将其判为事先确定的若干个文本类别中的某一类或某几类的过程[1-2],具有广泛的应用.但其面临的一大难题就是,文本特征空间的高维性,因此需要在保证一定分类精度的同时对文本特征进行降维.降维的2种常用方法是特征选择和特征抽取.特征选择是指从原始特征项集中选取一个子集构造新的特征空间.常用算法是基于阈值的过滤,如卡方(CHI)统计、信息增益(information gain,IG)、互信息(mutual information,MI). Yang[3]通过实验分析说明了CHI的分类效果比IG和MI好.但是CHI也存在一个不足之处,即过多地考虑了特征项和类别的负相关程度,可能选择在某类中出现较少而在其他类中普遍存在的特征,这会对分类结果产生干扰.

粗糙集理论是20世纪80年代由波兰数学家Pawlak首先提出的一个分析数据的数学理论.它不需要任何预备的或额外的有关数据信息,能够有效地分析和处理不完备、不一致、不精确的数据[4].粗糙集的核心是属性约简、删除冗余属性、获取对于决策分类最有用的属性,与特征选择有相似之处.于是,将粗糙集的属性约简和文本分类的特点相结合,提出改进的快速约简(improved quick reduction,IQR)算法,选择有用的特征表示文本,并通过实验验证了该算法的有效性.

1 粗糙集的基本理论

1.1 基本概念

在粗糙集理论中,一个信息表定义为二元组

收稿日期:2008-11-16.
基金项目:国家自然科学基金资助项目(60775036,60475019);高等学校博士学科点专项科研基金资助项目(20060247039).
通信作者:张志飞.E-mail:zzf_tj01@126.com.

$I=(U,A)$,其中非空有限集合 U 称为论域,非空有限集合 A 称为属性集. 对于任一属性 $a\in A$,存在一个信息函数 $f_a:U\to V_a$,V_a 称为属性 a 的值域. 通常属性集 A 可以划分为2个子集:条件属性集和决策属性集,分别用符号 C 和 D 表示,此时的信息表称为决策表.

设属性子集 $B\subseteq A$,称二元关系 $\mathrm{IND}(B)$ 为论域 U 上的 B-不可分辨关系,定义[4]

$$\mathrm{IND}(B)=\{(x,y)\in U^2\mid \forall a\in B,\ a(x)=a(y)\}. \tag{1}$$

如果 $(x,y)\in\mathrm{IND}(B)$,则称对象 x 和 y 在属性集 B 上是不可分辨的. 显然,$\mathrm{IND}(B)$ 是一个等价关系,$[x]_B$ 表示包含对象 x 的 $\mathrm{IND}(B)$ 的等价类.

设论域子集 $X\subseteq U$ 和属性子集 $B\subseteq A$,X 的 B-下近似和 B-上近似分别用 $\underline{B}X$ 和 $\overline{B}X$ 表示,定义[4]

$$\underline{B}X=\{x\mid [x]_B\subseteq X\}, \tag{2}$$

$$\overline{B}X=\{x\mid [x]_B\cap X\neq\varnothing\}. \tag{3}$$

前者表示根据 B 可以确定归入到 X 中的对象集合,后者表示根据 B 可能归入到 X 中的对象集合;而集合

$$\mathrm{POS}_B(X)=\underline{B}X \tag{4}$$

称为 X 的 B-正域.

1.2 属性约简

约简是保证正域不变的最小属性集合. 一个信息表可能存在多个约简. 计算最小约简是 NP-hard 问题,因此采用启发式方法寻找最优或次优约简,如 Maudal 提出的基于正域的快速约简算法[5],苗夺谦等人提出的基于互信息的属性约简算法[6]等.

快速约简(quick reduction, QR)算法[5]的基本思想是不断选择使正域大小变化最大的属性,直到正域大小和条件属性正域的大小相等为止.

算法:快速约简(QR)算法.

输入:决策表 $I=(U,C\cup D)$.

输出:约简属性 R.

1) 初始化 $R=\varnothing$.

2)令辅助集合 $T=R$.

3)对每个属性 $x\in C-R$,如果满足

$$\mathrm{card}(\mathrm{POS}_{R\cup\{x\}}(D))>\mathrm{card}(\mathrm{POS}_T(D)),$$

则令 $T=R\cup\{x\}$.

4)令 $R=T$,如果满足

$$\mathrm{card}(\mathrm{POS}_R(D))=\mathrm{card}(\mathrm{POS}_C(D)),$$

则转到5);否则转到2).

5)输出约简属性 R,算法终止.

2 基于粗糙集的特征选择

2.1 模型构建

根据训练文本集 $D=\{d_1,d_2,\cdots,d_n\}$ 得到候选特征项集 $T=\{t_1,t_2,\cdots,t_m\}$,每篇文本在特征项上的取值为1和0,分别表示该特征项在文本中出现和不出现. 将 D 作为论域,T 作为条件属性集,文本类别 c 作为决策属性,c 的值域 $V_c=\{c_1,c_2,\cdots,c_p\}$,构成文本分类决策表,如表1所示.

表1 文本分类决策表

Table 1 A decision table for text categorization

D	T				V_c
	t_1	t_2	$\cdots$	t_m	
d_1	1	0	$\cdots$	0	c_1
d_2	0	1	$\cdots$	1	c_1
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
d_n	1	1	$\cdots$	0	c_p

在文本分类中,此决策表具有如下特点:

1)条件属性集规模庞大,即 m 值很大,原因是文本向量空间的高维性.

2)属性取值极不均匀,即0、1分布差异大,原因是每篇文本的项相对很少.

2.2 改进的快速约简

文本分类决策表的约简要求不是很严格,既不需要是最小约简,也不需要具有完备性[7]. 因此可以牺牲完备性,以减少时间上的开销.

2.2.1 改进的不可分辨关系

粗糙集的基础是不可分辨关系,传统意义上认为2个对象在属性集 B 上不可分辨当且仅当 B 中所有属性的取值都相等. 由于文本分类决策表中0、1分布很不均匀,传统概念过于严格,使得2个文本在属性集上总是不可分辨的,无法进行后续的分析. 于是引入差异度的概念,表示在属性集上取值不相等的属性所占的比例.

设对象 x 和 y,属性子集 $B\subseteq T$,x 和 y 在 B 上的差异度定义为

$$D_B(x,y)=\frac{\mathrm{card}(\{a\in B\mid a(x)\neq a(y)\})}{\mathrm{card}(B)}. \tag{5}$$

设定阈值 β,只要2个对象的差异度低于该阈值,就说明2个对象在给定的属性集上不可分辨. 于是不可分辨关系可修改为

$$\mathrm{IND}'(B)=\{(x,y)\in U^2\mid D_B(x,y)<\beta\}. \tag{6}$$

2.2.2 属性局部排序和筛选

利用统计学的方差来衡量属性在类别间的波动情况. 波动越大,属性越具有类别区分度. 对于波动特别不明显的,将其删除. 因此该步骤在一定程度上

删除了不具有类别区分度的属性. 为了适合本文的模型,对文献[7]中的方法做一定的修改.

算法:属性局部排序和筛选算法.

输入:决策表 $I=(D,T\cup\{c\})$.

输出:条件属性子集 T'.

1)将决策表 I 按文本类别分割成类别矩阵.

2)对每个类别矩阵,计算每个列向量的和(即属性在该类别中的文档频率),得到向量 $\boldsymbol{X}$.

3)将不同类别的向量 $\boldsymbol{X}$ 组成一个矩阵 $\boldsymbol{M}$,计算每个列向量的方差,得到向量 $\boldsymbol{Z}$,同时记录每个列向量中最大值对应的文本类别,得到向量 $\boldsymbol{B}$.

4)按照 $\boldsymbol{Z}$ 中的值将属性从高到低排序,然后将大于指定阈值(一般为接近 0 的较小正数)的属性依序添加进 T'.

5)输出条件属性子集 T',算法终止.

2.2.3　阈值的计算

根据训练文本计算改进的不可分辨关系中的阈值. 原则是保证各类训练文本在 T' 上是可分辨的,此时有 $\mathrm{card}(\mathrm{POS}_{T'}(D))=\mathrm{card}(D)$.

算法:改进不可分辨关系的阈值计算.

输入:决策表 $I=(D,T\cup\{c\})$和属性子集 T'.

输出:阈值 β.

1)初始化 $\beta=1$.

2)对 D 中的每一篇文本和之后的所有文本进行如下操作:

a)如果 2 篇文本的类别相同,则转 d);

b)计算 2 篇文本在 T' 上的差异度 t;

c)如果满足 $t<\beta$,则重置 $\beta=t$;

d)转到 2)进行下一轮循环.

3)对 β 做略微调整,如 $\beta=0.9\cdot\beta$.

4)输出最后的阈值 β,算法终止.

2.2.4　改进的快速约简算法

在 1.2 中介绍的快速约简算法基础上提出了改进的快速约简(improved quick reduction, IQR)算法,将它作为本文进行属性约简的启发式算法.

算法:改进的快速约简(IQR)算法.

输入:决策表 $I=(D,T\cup\{c\})$,特征数目 N.

输出:特征项集 S.

1)对 T 进行局部排序和筛选,得到 T'.

2)初始化 $R=\varnothing$.

3)令辅助集合 $T=R$.

4)按序选择 $T'-R$ 中的属性 ,如果满足

$$\mathrm{card}(\mathrm{POS}_{R\cup\{x\}}(\{c\}))>\mathrm{card}(\mathrm{POS}_{T'}(\{c\})),$$

则令 $T=R\cup\{x\}$.

5)令 $R=T$,如果满足

$$\mathrm{card}(\mathrm{POS}_{R}(\{c\}))=\mathrm{card}(D),$$

则转到 6);否则转到 3).

6)初始化特征项集 $S=R$.

7)根据属性局部排序和筛选算法中得到的 $\boldsymbol{Z}$ 和 $\boldsymbol{B}$,将 $T'-R$ 中的属性按照 $\boldsymbol{B}$ 的值分组,组内按照 $\boldsymbol{Z}$ 的值从大到小排序.

8)平均从每组中按序选择属性构成 $N-|R|$ 个属性添加到 S.

9)输出特征项集 S,算法终止.

与快速约简 QR 算法的区别在于增加了第 1)步,对属性集进行局部排序和筛选,时间复杂度为 $O(|V_c\|D|)$,相比约简复杂度 $O(|T'|^2|D|^2)$来说影响不大. 此外,属性集规模减小,能够避免盲目搜索,有助于快速找到约简,提高整个算法的效率. 算法还增加了第 7 和 8 步,得到约简之后,按照剩余属性的类间波动及出现最多的类别补充一定数目的属性,构成文本分类的特征项集,时间复杂度相比约简可以忽略. 之所以这样做,是因为若约简出的特征词较少,很多文本因不含这些特征词而表示为空,导致分类器无法识别,误分几率增大,分类效果降低.

3　实验及其分析

3.1　实验设置

实验采用中科院计算所谭松波提供的语料库,共有财经、电脑、房产、教育、科技、汽车、人才、体育、卫生和娱乐 10 个类别,1 500 篇文本. 预处理采用中科院 ICTCLAS 进行分词,文本采用 LTC 权重评价方法[8]和向量空间模型[1]表示,分类器采用支持向量机[1].

文本分类一般采用查全率和查准率等指标来衡量分类系统的性能. 查全率 R 是指分类正确的文本数与应有的分类文本数之比,反映分类的全面性;查准率 P 是指分类正确的文本数与实际分类的文本数之比,反映分类的准确性. 通常使用宏平均和微平均作为衡量性能参数,前者是先对每个类计算,然后对所有类求平均值,而后者是根据所有类计算,比前者更为常用[1]. F_1 是查全率 R 和查准率 P 的综合评价:

$$F_1=2PR/(P+R). \tag{7}$$

因此,实验采用微平均 F_1 来评价算法的分类效果.

3.2　实验结果及分析

考虑到属性约简的时间复杂度,该实验仅从 10 个类别中抽取 4 个类别,每个类别 30 篇训练文本.

50篇测试文本.得到在不同组合下属性约简结果如表2所示,以及IQR和CHI在不同的特征维度下的微平均F_1对比情况,如图1所示.

表2 4种组合下属性约简结果

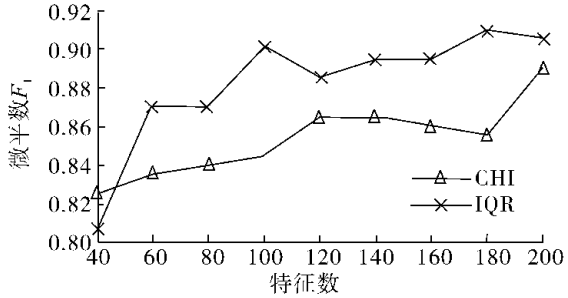
Table 2 Results of four combinations' attribute reduction

组合	属性维度		微平均F_1
	原始	约简	
财经+教育+人才+娱乐	7 696	11	0.665
电脑+房产+汽车+科技	6 329	12	0.745
教育+汽车+人才+卫生	6 531	11	0.740
科技+汽车+体育+卫生	6 536	12	0.740

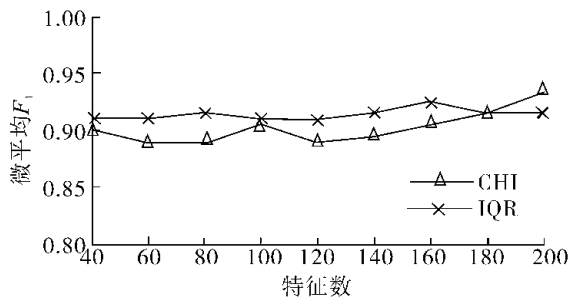
从表2可以看出,经过快速约简之后,特征维度从六七千维减小到十几维,降低幅度很大,此时的微平均F_1值为70%左右,说明用粗糙集属性约简方法得到的特征词具有很强的分类能力.

结合表2和图1可以看出,补充适量的特征词增大特征维度之后,例如40维时,微平均F_1值为80%以上,提高的幅度很大.原因在于,当特征维度太小时,一部分文本因不含有这些特征词而表示为空,导致分类器无法识别,误分几率增大.

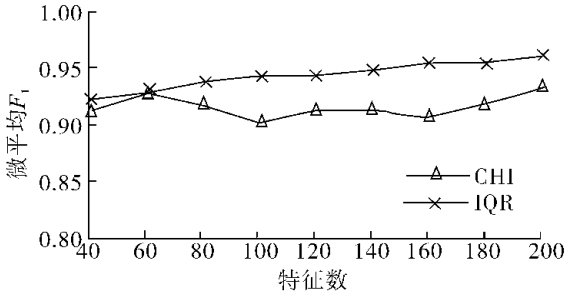
从图1可以看出,特征维度增大到一定量时,微平均F_1值将达到90%以上,分类效果较佳;同时发现当特征维度继续增大时,微平均F_1值反而减小,说明过多的特征词会对分类产生干扰.



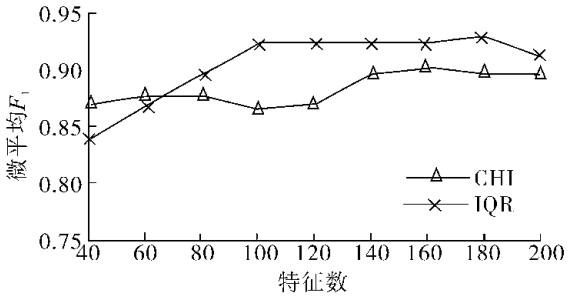
(a) 财经/教育/人才/娱乐类组合



(b) 电脑/房产/汽车/科技类组合



(c) 教育/汽车/人才/卫生类组合



(d) 科技/汽车/体育/卫生类组合

图1 4种组合下IQR、CHI微平均F_1比较

Fig. 1 Micro_F_1 comparison of four combinations' IQR and CHI

综合考虑以上4种组合的情况,如图2所示.对于原始维度在7 000左右的小文本集,用IQR和CHI将维度降到200左右,均能保持相当好的分类效果,而且IQR比CHI的分类效果有一定程度的提高.主要原因是:IQR用快速约简得到的特征词具有很强的分类能力,再根据波动性和类别情况补充特征词,不仅尽可能解决了文本表示为空的问题,而且能够区分有些低频词,所以分类效果比CHI稍好.

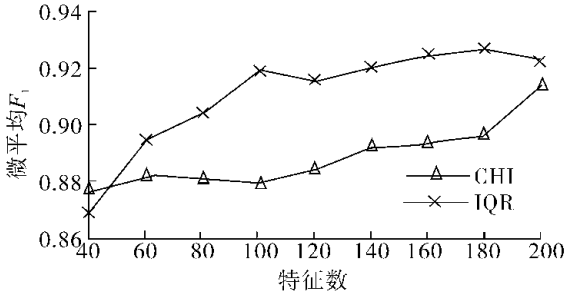


图2 IQR、CHI微平均F_1比较

Fig. 2 Micro_F_1 comparison of IQR and CHI

4 结 论

文本分类的分类效果很大程度上取决于文本特征项的选取.在小文本集上,IQR的效果比CHI稍好,验证了IQR算法的有效性.但实验仍存在不足之处:

1)决策表的属性值为布尔型,丢失了部分信息,可尝试用词频表示;

2)IQR 和 CHI 的对比验证是在小文本集上完成的,没有推广到大文本集上,原因在于本实验采用的属性约简算法的时间复杂度高.

因此,本文下一步的工作主要在于设计时间复杂度较小的属性约简算法,以提高约简速度,并将其应用到大文本集上,分析其实际应用效果.

参考文献:

[1]苗夺谦, 卫志华. 中文文本信息处理的原理与应用[M]. 北京: 清华大学出版社, 2007: 214-230.

[2]周　屹. 基于 Naive Bayes 的文本分类器的设计与实现[J]. 黑龙江工程学院学报, 2007, 21(2): 28-30.
ZHOU Yi. A text classifier's design and realization based on Naive Bayes method[J]. Journal of Heilongjiang Institute of Technology, 2007, 21(2): 28-30.

[3]YANG Yiming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning. Nashville, USA, 1997: 412-420.

[4]王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001: 1-100.

[5]MAUDAL O. Preprocessing data for neural network based classifiers: rough sets vs principal component analysis[R]. Edinburgh: University of Edinburgh, 1996.

[6]苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681-684.
MIAO Duoqian, HU Guirong. A heuristic algorithm for reduction of knowledge[J]. Computer Research and Development, 1999, 36(6): 681-684.

[7]盛晓炜, 江铭虎. 基于 Rough 集约简算法的中文文本自动分类研究[J]. 电子与信息学报, 2005, 27(7): 1047-1052.
SHENG Xiaowei, JIANG Minghu. Automatic classification of Chinese documents based on rough set and improved quick-reduce algorithm[J]. Electronics and Information Technology, 2005, 27(7): 1047-1052.

[8]AAS K, EIKVIL L. Text categorisation: a survey[R]. Oslo: Norwegian Computing Center, 1999.

作者简介:

张志飞,男,1986 年生,硕士研究生,主要研究方向为文本挖掘、智能信息处理.

苗夺谦,男,1964 年生,教授、博士生导师. 中国计算机学会人工智能与模式识别专业委员会委员,中国人工智能学会理事,上海市计算机学会理论与人工智能专业委员会委员. 主要研究方向为粗糙集理论、粒计算、主曲线、网络智能、数据挖掘等. 已主持完成多项国家、省部级自然科学基金与科技攻关项目,并参与完成"973"计划子项目 1 项,"863"计划项目 2 项. 曾获国家教委科技进步三等奖、山西省科技进步二等奖、教育部科技进步一等奖等. 发表学术论文 120 余篇,其中被 SCI 和 EI 等收录 50 余篇,出版学术专著 3 部.