文本分类中互信息特征选择方法的研究

范小丽,刘晓霞
FAN Xiao-li,LIU Xiao-xia

西北大学 信息科学与技术学院,西安 710127
College of Information Science & Technology,Northwest University,Xi'an 710127,China
E-mail:fxl_209@163.com

FAN Xiao-li, LIU Xiao-xia.Study on mutual information-based feature selection in text categorization.Computer Engineering and Applications,2010,46(34):123-125.

Abstract: To solve the problem of the poor effect of mutual information-based feature selection on the unbalanced corpus which arise from not well combining positive feature and negative feature.The ratio of positive feature and negative feature is adjusted with balance factor to strengthen the effect of negative feature.And category strong related feature is distincted with feature distributed factor.The experimental results verify the efficiency and probability of the improved mutual information-based feature selection.

Key words: text categorization;feature selection;mutual information;balance factor;feature distribute difference

摘 要:针对互信息特征选择方法由于没有很好结合正相关特征和负相关特征,影响在不平衡语料集上分类效果的问题,用平衡因子调整正相关和负相关特征比例,加强特征选择时负相关特征的作用。同时引入特征分布差异因子,区分类强相关特征,提高分类效果。最后通过实验证明,改进的互信息特征选择方法具有可行性和有效性。

关键词:文本分类;特征选择;互信息;平衡因子;特征分布差异
DOI:10.3778/j.issn.1002-8331.2010.34.037 **文章编号:**1002-8331(2010)34-0123-03 **文献标识码:A** **中图分类号:**TP391

1 引言

文本分类中高维特征空间不仅增加了分类的时间复杂度和空间复杂度,而且影响分类精度。特征选择是一种常用的特征降维方法,主要作用是删除特征项空间中信息量小,不重要的特征,选择具有类别区分能力的特征,降低原始特征空间维度,从而改善分类效果。

传统特征选择方法在平衡语料集上分类效果较好,但在非平衡语料集上分类效果不明显,大量特征选择研究文献提出通过最优组合正相关特征和负相关特征,提高在不平衡语料集上的分类效果[1-2]。很多研究者强调高频词的作用,即在特征选择方法中加入$p(w)$因子,提高分类效果[3]。互信息特征选择方法是一种常用的特征选择方法,它有效地表达了特征与类别之间的依赖程度,但由于忽略了负相关特征的潜在影响,并且倾向于选择低频词,影响了分类效果。针对互信息特征选择方法存在的以上问题,提出一种改进的互信息特征选择方法,提高其在不平衡语料集上的分类效果。最后通过实验证明,改进的方法是可行的。

2 互信息特征选择方法

文本分类中常用的特征选择方法有文档频率(DF),互信息(MI),信息增益(IG),卡方统计(CHI),期望交叉熵(CE)和几率比(OR)。

互信息根据特征和类别共同出现的概率,度量特征和类别的相关性。特征t和类别c_i互信息值计算公式如公式(1)所示:

$$MI(t,c_i)=\log\frac{p(t,c_i)}{p(t)\times p(c_i)}=\log\frac{p(t|c_i)}{p(t)} \quad (1)$$

其中,$p(t,c_i)$表示训练集中既包含特征t又属于类别c_i的文本出现的概率,$p(t)$表示包含特征t的文本在训练集中出现的概率,$p(c_i)$表示训练集中属于类别c_i的文本的概率。特征t在类别c_i中出现概率高,而在其他类别中出现概率低,即特征t和类别c_i相关性大,将获得较高的互信息值$MI(t,c_i)$。反之,将获得较低的互信息值$MI(t,c_i)$。在m个类别的集合上特征项t的互信息值定义为公式(2):

$$MI(t)=\sum_{i=1}^{m}MI(t,c_i) \quad (2)$$

从公式(1)可以看出,当$p(t|c_i)>p(t)$时,$MI(t,c_i)>0$,说明特征t和类别c_i是正相关的,即特征的出现说明文档可能属于某个类别;反之,$p(t|c_i)<p(t)$,$MI(t,c_i)<0$,说明特征t和类别c_i

基金项目:航空科学基金项目(Aviation Science Fund under Grant No.2006ZC31001)。
作者简介:范小丽(1986-),女,硕士研究生,主要研究方向为网络信息处理;刘晓霞(1965-),女,教授,主要研究方向为信息处理、图形图像处理。
收稿日期:2010-05-18 **修回日期:**2010-07-06

是负相关的,即特征的出现说明文档可能不属于某个类别。用公式(2)计算特征在所有类别集合上的互信息值时,特征与类别负相关的那部分为负的互信息值,会削弱此特征的权值,因此会影响互信息特征选择方法的分类精度,尤其在类分布不均匀的情况下。

从公式(1)还可以看出,特征的 $p(t|c_i)$ 相等时,稀有词具有比普通特征词较高的互信息值,互信息特征选择方法的缺点是没有考虑$p(w)$因子,倾向于选择罕见特征[4]。和类别强相关的特征对类别的区分性较好,而由$p(w)$因子选择的高频词与类别不一定是强相关的。

3 互信息特征选择方法改进

通过以上对互信息特征选择方法不足的分析,提出调整正相关特征与负相关特征比例和区分类强相关高频词的研究方法,对互信息特征选择方法改进。

3.1 平衡因子α

在文本分类中,正相关特征起主要作用;负相关特征起次要作用[5]。很多文献改进特征选择方法时,去除了负相关特征的作用[6]。文献[5]通过实验证明,文本分类时正相关特征有利于提高查准率,而负相关特征有助于提高查全率。负相关特征说明特征项与类别不相关,在特征选择时有助于删除不相关的文档,它在分类中的作用是不可忽视的。

提出以平衡因子α调整互信息特征选择方法中正相关特征和负相关特征比例的方法,提高互信息特征选择方法在不平衡语料集上的分类效果。在公式(1)中,当 $p(t|c_i)>p(t)$,即特征和类别正相关时,互信息计算公式如公式(3)所示:

$$MI(t,c_i)^+=\alpha\times\log\frac{p(t|c_i)}{p(t)},0<\alpha<1 \tag{3}$$

当 $p(t|c_i)<p(t)$,即特征和类别负相关时,互信息计算公式如公式(4)所示:

$$MI(t,c_i)^-=(1-\alpha)\times\log\frac{p(t|c_i)}{p(t)},0<\alpha<1 \tag{4}$$

改进的互信息特征选择方法的公式如公式(5)所示:

$$MI(t,c_i)=MI(t,c_i)^+-MI(t,c_i)^- \tag{5}$$

3.2 特征分布差异因子

某特征集中分布在几个类之间,同时在这几个类内均匀分布,则此特征和类别是强相关的,有较好的类别区分性。特征在类间的分布情况可以用的类间离散度 D_{ac} [7]表示,计算公式如公式(6)所示:

$$D_{ac}=\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(tf_i(T_k)-\overline{tf(T_k)})^2}}{\overline{tf(T_k)}} \tag{6}$$

其中,$tf_i(T_k)$ 表示词条 T_k 在 C_i 类的出现概率,$\overline{tf(T_k)}$ 代表词条 T_k 在各个类的平均词频。类间离散度高的特征词,具有强的类别区分性。特征在类内的分布情况用类内离散度 D_{ic} [7]表示。计算公式如公式(7)所示:

$$D_{ic}=\frac{\sqrt{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(T_k)-\overline{tf_i(T_k)})^2}}{\frac{m}{\sqrt{m-1}}\overline{tf_i(T_k)}} \tag{7}$$

其中,$\overline{tf_i(T_k)}$ 表示在特征词 T_k 在类 C_i 文档中的平均词频,$tf_{ij}(T_k)$ 表示 T_k 在各个类别的文档中的词频。类内离散度小的特征词,具有强的类别区分性。

最后,在互信息特征选择方法中加入词频因素$p(w)$基础上,引入特征分布差异因子 p_r ,提高类强相关高频特征的权重。特征分布差异因子 p_r 如公式(8)所示:

$$p_r=\frac{D_{ac}}{D_{ic}} \tag{8}$$

从公式(8)可以看出,特征具有较大的类间离散度,同时具有较小的类内离散度,则特征对类别的区分性强;反之,特征对类别区分性较弱。将特征分布差异因子 p_r 和词频因素$p(w)$加入公式(5)方法中,如公式(9)所示:

$$MI(t,c_i)=p_r\times p(w)\times(MI(t,c_i)^+-MI(t,c_i)^-) \tag{9}$$

4 实验结果及分析

为了观察比较传统互信息特征选择方法与文中改进的互信息特征选择方法的分类效果,用VC++6.0实现文本分类系统以及文中改进的特征选择方法,并进行实验。

4.1 语料集

实验采用Sogou网站提供的文本分类语料库。选取一个类分布不平衡语料集进行实验,其中各类文档选取情况如表1所示。

表1 语料集上训练集和测试集的选取情况

	教育	文学	医药	经济	科技	体育
训练集	724	1 333	217	890	1 452	323
测试集	657	657	300	657	833	350

4.2 分类器

实验采用*K*NN分类器测试数据。*K*NN(*k*-Nearest Neighbor)分类方法被广泛应用于文本分类,与其他分类方法相比,它具有方法简单,错误率较低等优点。

4.3 评价标准

实验采用查准率P(Precision),查全率R(Recall),平均*F*1值作为评价指标。

$$\text{查全率}=\frac{\text{正确分类的文档数}}{\text{被测试文档的总数}}$$

$$\text{查准率}=\frac{\text{正确分类的文档数}}{\text{被分类器识别为该类的文档数}}$$

平均*F*1值是查全率和查准率的综合,公式如公式(10)所示:

$$F1=\frac{2\times p\times r}{p+r} \tag{10}$$

4.4 平衡因子α选取

通过实验提取平衡因子α的值。因为文本分类时,正相关特征起主要作用,负相关特征起次要作用[5],所以公式(5)中α因子从0.5~1之间,以0.1为平均步长取值,并且对其实验,实验结果如图1所示。

从实验结果看到,平衡因子α取值从0.5增到0.9时,正相关特征相对负相关特征比重增加,分类效果提高,平均*F*1值缓慢增加,当α取值为0.9时,平均*F*1值达到最高值0.48,之后随着负相关特征所占的比重减少,分类精度降低,平均*F*1值降低到0.3。

对实验结果分析发现,负相关特征有助于提高分类效果,公式(5)的方法中平衡因子α取值为0.9时,可以达到较好的分

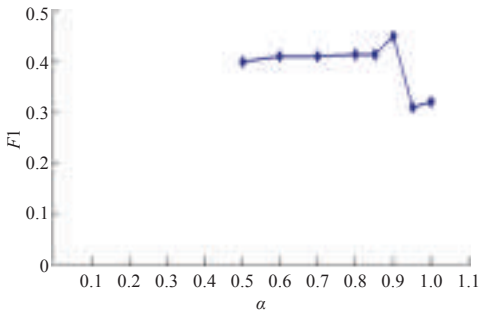

图1 平衡因子α取值所对应$F1$值

类效果。

4.5 实验数据

首先在选取的语料集上分别对互信息特征选择方法和公式(5)方法实验。然后在公式(5)的方法中加入$p(w)$因子和特征分布差异因子区分类强相关的高频特征,并进行实验。

实验1:为了比较互信息特征选择方法和公式(5)中方法的分类效果,在选取的语料集上分别对互信息特征选择方法和公式(5)中的方法实验,实验结果如表2、表3所示。

表2 传统互信息特征选择方法的分类效果

类别	查全率/(%)	查准率/(%)
教育	48.546	59.290
文学	68.402	23.119
医药	2.158	66.667
经济	20.543	28.673
科技	35.266	30.080
体育	20.583	84.328
宏平均查全率/(%)	32.583	
宏平均查准率/(%)		48.692
平均$F1$/(%)	39.041	

表3 公式(5)方法的分类效果

类别	查全率/(%)	查准率/(%)
教育	62.922	71.743
文学	44.346	22.805
医药	4.642	64.865
经济	47.395	44.904
科技	55.769	38.284
体育	46.333	84.091
宏平均查全率/(%)	43.567	
宏平均查准率/(%)		54.448
平均$F1$/(%)	48.403	

从实验结果看到,互信息特征选择方法因为没有考虑互信息量为负的特征的潜在影响,导致特征在依据阈值选取时被淘汰,部分噪音信息被保留,分类效果低。公式(5)的方法加强了负相关特征的作用,在整体上,查全率和查准率较传统互信息特征选择方法好。

对实验结果分析发现,尽管公式(5)方法分类效果较原方法有所提高,但在整体上分类效果并不是很理想。由于互信息特征选择方法的缺点是没有考虑$p(w)$因子[4],提出在加入词频因素的基础上,用特征分布差异因子提高类别区分度高的高频特征的权值的研究方法。

实验2:在公式(5)的方法中加入特征分布差异因子和词频因素$p(w)$,即公式(9)的方法在语料集上实验,实验结果如表4所示。

表4 公式(9)方法的分类效果

类别	查全率/(%)	查准率/(%)
教育	81.735	89.799
文学	71.309	84.615
医药	43.379	88.235
经济	69.559	78.120
科技	97.412	44.077
体育	67.428	96.725
宏平均查全率/(%)	71.803	
宏平均查准率/(%)		80.261
平均$F1$/(%)	75.796	

从实验结果看到,公式(9)方法的分类效果,整体上有了明显的提高。说明文中改进的互信息特征选择方法具有有效性和可行性。

5 结束语

特征选择方法是文本分类的重要环节,互信息特征选择方法是一种常用的特征选择方法,它的缺点是没有考虑负相关特征的潜在影响和词频因素$p(w)$。提出通过调整互信息特征选择方法中正相关特征与负相关特征比例,加强负相关特征项作用,并且以特征分布差异因子的方式区分类强相关特征的改进方法,明显提高了其在不均匀语料集上的分类效果。

参考文献:

[1] 徐燕,李锦涛,王斌,等.不均衡数据集上文本分类的特征选择研究[J].计算机研究与发展,2007,44(增刊):58-62.

[2] Zheng Zhao-hui, Wu Xiao-yun, Rohini S.Feature selection for text categorization on imbalanced documents[J].SIGKDD Explorations Newsletters,2004,6(1):80-89.

[3] 陆玉昌,鲁明羽,李凡,等.向量空间中单词权重函数的分析和构造[J].计算机研究与发展.2002,39(10):1205-1210.

[4] 尚文倩,黄厚宽,刘玉玲,等.文本分类中基于基尼指数的特征选择算法研究[J].计算机研究与发展,2006,43(10):1688-1694.

[5] Forman G.An extensive empirical study of feature selection metrics for text classification[J].Journal of Machine Learning Research,2003,3(1):1289-1305.

[6] Calvo B, Larrariaga P, Lozano J A.Feature subset selection from positive and unlabelled examples[J].Pattern Recognition Letters,2009,30:1027-1036.

[7] 熊忠阳,黎刚,陈小莉,等.文本分类中词语权重计算方法的改进与应用[J].计算机工程与应用,2008,44(5):187-189.