改进的χ^2统计文本特征选择方法

肖 婷,唐 雁
XIAO Ting,TANG Yan

西南大学 计算机与信息科学学院,重庆 400715
School of Computer & Information Science,Southwest University,Chongqing 400715,China
E-mail:xiaott2006@163.com

XIAO Ting,TANG Yan.Improved χ^2 statistics method for text feature selection.Computer Engineering and Applications, 2009,45(14):136-137.

Abstract: Feature selection is a hot topic in current search field,especially in the field of text categorization.In this paper,χ^2 statistical method has two defects.One is reducing the weight of the low-frequency words.The other is increasing the weight of the characteristics in the designated class.The characteristics little appear in designated class but other classes.Through simulation and comparison experiment,the result is better than before.

Key words: text categorization;feature selection;χ^2 statistics

摘 要:特征选择是当今研究领域的一个热点,尤其是文本分类领域中的热点。针对χ^2统计方法的两个缺陷:降低了低频词的权重和提高了很少在指定类中出现但普遍存在于其他类的特征在该类中的权重,对χ^2统计方法进行改进,并通过做模拟和对比实验,对比改进前后的方法对文本分类的影响。在模拟和对比实验中,改进后方法的分类效果要好于传统的方法。

关键词:文本分类;特征选择;χ^2统计
DOI:10.3778/j.issn.1002-8331.2009.14.041 **文章编号**:1002-8331(2009)14-0136-02 **文献标识码**:A **中图分类号**:TP39

1 引言

文本挖掘是近几年来数据挖掘领域的一个研究热点[1],文本分类作为文本挖掘的重要手段和工具之一,其目标是对未知类别的文档进行自动处理,判断它们所属的类别。特征选择作为文本分类的前提,其重要性和意义就可想而知。一个合理、有效的特征选择方法可以在数据预处理阶段去掉数据中的冗余,降低特征空间的维数,提高分类的效率。

文本分类中,特征选择的基本思想通常是构造一个评价函数,对特征集的每个特征进行分别评估,后对所有的特征按照其评估分的大小进行排序,选取预定数目的最佳特征作为结果的特征子集[2]。

常用的特征选择方法有:文档频率(DF)、信息增益(IG)、互信息(MI)、χ^2统计方法(CHI)、期望交叉熵(Expected Cross Entropy)、文本证据权(Weight of Evidence text)等。但是在众多的特征选择方法中,Yang指出IG(信息增益)和CHI的效果最好[3],而IG计算量相对其它几种方法较大,因此本文主要针对目前特征选择方法中效果最好χ^2统计方法进行研究和改进。

2 χ^2统计方法相关研究

2.1 χ^2统计方法

χ^2统计方法度量词条t和文档类别c之间的相关程度,假设词条t和类别c之间符合具有一阶自由度的χ^2分布。词条对于某类的χ^2统计值越高,它与该类之间的相关性就越大,携带的类别信息也就越多。令N表示训练语料库中的文档总数,c为某一特定类别,t表示特定的词条,A表示属于c类且包含t的文档频数,B表示不属于c类但包含t的文档频数,C表示属于类别c但不包含t的文档频数,D表示不属于c也不包含t的文档频数。则t对于c的CHI值由下式计算:

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

当特征t与类别c的相关性越强,$\chi^2(t,c)$的值就越大,代表特征t包含与特定类别c之间的独立性就越小,对分类的贡献就越大。有关的信息就越多。当特征t与类别c相互独立时,$\chi^2(t,c)=0$,代表特征t不包含任何与特定类别c有关的信息。

对于多个类别的问题,分别计算t对于每个类别的χ^2值,再选择出整个语料库中χ^2值最大的作为整个语料库的χ^2值。可以事先假定一个阈值,从原始的特征空间中删除那些低于特定阈值的词条,保留大于该阈值的特征词条作为文档表示的特征词。

2.2 存在的问题

Yang的研究表明,χ^2统计方法虽然是目前对于中文文本分类效果最好的一个特征选择方法之一[3]。大多数中文分类系

作者简介:肖婷(1982-),女,硕士研究生,主要研究方向:数据挖掘;唐雁(1965-),女,教授,硕士生导师,主要研究方向:多媒体技术、数据挖掘。
收稿日期:2008-11-20 **修回日期**:2009-02-13

统都采用这种方法[4-5]。但是,根据分析,它依然存在着缺陷,主要表现在两个方面。

(1)降低了低频词的权重。

χ^2统计方法不能准确地保留往往是某类特有的,具有很强代表性的低频词。这里的低频词是指文档频数,但某些低频词往往是某类特有的特征词,这些特征词具有很强的代表性,虽然只出现在指定类的少量文章中,但在这少量文章中频繁的出现,因此对分类的贡献很大。但是通过公式$\chi^2(t,c)$计算出来的χ^2统计值相对来说比较小,而根据TFIDF理论[6],权重低的特征项将被过滤掉。于是,由于低频词的权重比较低,χ^2统计就删除了这本该保留的具有很强代表性的特征项。

(2)提高了很少在指定类中出现但普遍存在于其他类的特征在该类中的权重。

在整个训练集语料库中一些出现频率较高的词,而这些词语在指定类中出现得很少甚至几乎是不出现的(A值很低),显然这些词是不能很好地代表该指定类,应该被过滤掉。但是$BC>>AD$导致用公式$\chi^2(t,c)$计算出来的χ^2统计值相对来说比较大,而根据TFIDF理论[6],权重高的特征项将被保留,作为该指定类的特征项。

3　χ^2统计方法的改进

特征选择方法的好坏直接影响文本分类的效果。前面分析的χ^2统计方法有两个方面的缺陷:第一,降低了低频词的权重;第二,提高了很少在指定类中出现但普遍存在于其他类的特征在该类中的权重。这里认为降低了低频词的权重这一点,对分类效果的影响相对较为明显。所以主要针对降低了低频词的权重这一问题,引入了文档内频度的概念,并且兼顾第二个问题引入类内正确度的概念,对χ^2统计方法进行改进。

(1)文档内频度

针对问题1,把特征项在具体的文档中出现的频度考虑进χ^2统计的计算公式。设训练集中类别为C_i的文本有$d_{i1},d_{i2},\cdots,d_{ik},\cdots,d_{ij}$,特征$t$在文本$d_{ik}(1\leqslant k\leqslant j)$中出现的频度为$tf_{ik}$,则特征$t$在类别$C_i$中出现的频度记为$\alpha$,即文档内频度表示为:$\alpha=\sum_{k=1}^{j}tf_{ik}$。

(2)类内正确度

针对问题2,由于是A值太小,导致$BC>>AD$造成的,所以引入一个调节函数$\beta=\frac{A}{A+B}(0\leqslant\beta\leqslant1)$即词条出现在指定类中的比例。当$\beta$靠近0时,即在整个语料库出现频繁,而在指定类出现较少。当β靠近1时,即在指定类出现较多。由于β的范围是[0,1],这样的β值对计算出的χ^2值影响有限,因此,对β函数变换到β',把[0,1]映射到$[M^{-1},M]$中去,M是正整数。即当$A>B$时(即词条能很好地代表指定类),$\beta'>1$,参数β放大计算出的χ^2值;当$A<B$时(词条指定类出现很少),$\beta'<1$,参数β减小计算出的χ^2值。因此β'记作类内正确度,表示为:

$$\beta'=M^{2\beta-1} \tag{2}$$

鉴于以上分析,考虑将α和β'加入原来的χ^2统计方法的计算公式,使得通过χ^2公式计算出来的词条t和类别c之间的相关程度更为准确,这样提取出来的特征项更能代表指定类。将改进后的χ^2统计方法的计算公式表示如下:

$$\chi^2(t,c)=\frac{N(AD-BC)}{(A+C)(B+D)(A+B)(C+D)}\times\alpha\times\beta' \tag{3}$$

4　实验与结论

特征选择是文本分类的关键技术,特征提取的好与坏直接关系到文本分类的效果,将特征选择方法作为重点研究对象,对χ^2统计方法提出了改进措施,用文本分类的结果来验证改进的有效性。

4.1　朴素贝叶斯分类算法

在文本分类算法的选择中,这里选择最为常见的分类贝叶斯算法。在朴素贝叶斯分类(简称NB)基于贝叶斯定理,可以用来预测类成员关系的可能性,给出文本属于某特定类别的概率。分类时根据预测结果该样本分到概率最高的类别中去即可。在这个方法中,有一个“独立性假设”[7]:在给定的文本类语境下,文本中每个特征词是相互独立的。

设d为一任意文本,它属于文档类$C=\{c_1,c_2,\cdots,c_j\}$中的某一类c_k。根据NB分类方法有:

$$p(c_k|d)=\frac{p(c_k)p(d|c_k)}{p(d)} \tag{4}$$

对文本d进行分类,就是按照上述公式计算所有文本类在给定d情况下的概率,概率值最大的那个类就是文本d所属的类。其中的$p(c_k)=\frac{N_k}{N}$,$p(d_i|c_k)=\frac{1+N_{ik}}{M+\sum_{k=1}^{M}N_{kj}}$,$N_{ik}$表示特征$i$在训练类别$c_k$在文本出现的次数,$N_k$表示类别$k$包含的训练文本数,$i$表示特征项数,$d_i$表示特征$i$。

4.2　实验及结果

从来源于搜狐新闻网站的大量经过编辑手工整理与分类的新闻语料中抽取部分短文共计1 500多篇,采用交叉验证的方法对5个类的文档进行实验。本实验公式(2)中的M值取4。

一般用准确率和召回率来衡量信息挖掘系统的结果。因此,最终的实验结果采用通用的召回率(recall)和准确率(precision)两个指标作为评测标准:

$$\text{准确率(precision)}=\frac{\text{分类正确的文本数}}{\text{实际分类的文本数}}$$

$$\text{召回率(recall)}=\frac{\text{分类正确的文本数}}{\text{应有的文本数}}$$

下面3个表是特征向量空间维数为500、1 000、1 200维的实验数据。

通过对比实验,从表1~表3可以看出,改进后的召回率和准确率普遍高于改进前的。这就证明,本文的对于χ^2统计方法的改进是有较好的效果的。

表1　特征向量空间500维

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.767/0.737	0.190/0.136	0.335/0.403	0.767/0.737	0.466/0.521	0.874/0.894	computer
0.733/0.76	0.028/0.023	0.764/0.806	0.733/0.76	0.748/0.782	0.940/0.959	health
0.830/0.880	0.009/0.009	0.922/0.926	0.830/0.880	0.874/0.903	0.977/0.989	sports
0.847/0.900	0.018/0.008	0.858/0.931	0.847/0.900	0.852/0.915	0.975/0.992	train
0.563/0.687	0.014/0.025	0.833/0.777	0.563/0.687	0.672/0.729	0.899/0.939	travel

表2　特征向量空间1 000维

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.707/0.750	0.100/0.081	0.468/0.537	0.707/0.750	0.563/0.626	0.902/0.925	computer
0.747/0.830	0.026/0.020	0.783/0.836	0.747/0.830	0.765/0.833	0.950/0.962	health
0.863/0.920	0.012/0.012	0.902/0.098	0.863/0.920	0.882/0.914	0.985/0.994	sports
0.880/0.907	0.016/0.006	0.874/0.948	0.880/0.907	0.877/0.927	0.988/0.992	train
0.723/0.757	0.024/0.015	0.792/0.860	0.723/0.757	0.756/0.805	0.935/0.955	travel

(下转140页)