第 51 卷 第 4 期 厦门大学学报(自然科学版) Vol. 51 No. 4
2012 年 7 月 Journal of Xiamen University (Natural Science) Jul. 2012

基于特征相关的改进加权朴素贝叶斯分类算法

饶丽丽[1],刘雄辉[2],张东站[1]*
(1. 厦门大学信息科学与技术学院,福建 厦门 361005;2. 龙岩烟草工业有限责任公司信息技术部,福建 龙岩 364021)

摘要:朴素贝叶斯分类算法的特征项间强独立性的假设在现实中是很难满足的.为了在一定程度上放松这一假设,提出了基于特征相关的改进加权朴素贝叶斯分类算法,该算法采用一种新的权重计算方法,这种权重计算方法是在传统词频-反文档频率(TF-IDF)权重计算基础上,考虑到特征项在类内和类间的分布情况,另外还结合特征项间的相关度,调整权重计算值,加大最能代表所属类的特征项的权重,将它称之为 TF-IDF-FC 权重计算.与基于传统 TF-IDF 权重的加权朴素贝叶斯分类算法和其他常用加权朴素贝叶斯分类算法比较,如基于属性加权的朴素贝叶斯分类算法,这种算法的分类效果均有一定的提高.

关键词:朴素贝叶斯文本分类器;加权朴素贝叶斯文本分类算法;TF-IDF 权重;特征项间的相关度
中图分类号:TP 391.1 **文献标志码**:A **文章编号**:0438-0479(2012)04-0682-04

朴素贝叶斯方法[1]是目前公认的一种简单有效的分类方法,它是一种基于概率的分类方法,被广泛地应用于模式识别、自然语言处理、机器人导航、规划、机器学习以及利用贝叶斯网络技术构建和分析软件系统.朴素贝叶斯文本分类方法是基于特征项间独立的假设,但是这与现实是不相符的,为此很多人研究出一种加权朴素贝叶斯算法,对后验概率计算中的每个条件概率项进行加权,并且对不同的特征项提供不同的加权值,从而使得特征项之间是不独立的,它们对类别的重要程度是不一样的[2].本文提出一种基于特征相关的改进加权朴素贝叶斯算法,在传统词频-反文档频率(TF-IDF)权重的基础上,考虑到类内和类间分布,同时根据特征项之间的相关程度,对它们的权重进行调整,突出相关性比较大的特征项权重,从而提高了加权朴素贝叶斯的分类能力.

1 朴素贝叶斯文本分类

朴素贝叶斯算法[1]的工作过程如下:

1) 设 D 是训练元组和相关联的类标号的集合.照例,每个元组用一个 n 维属性向量 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}$ 表示,描述由 n 个属性 $A_1,A_2,\cdots,A_n$ 对元组的 n 个测量.

2) 假定有 m 个类 $C_1,C_2,\cdots,C_m$.给定元组 $\boldsymbol{X}$,分类法将预测 $\boldsymbol{X}$ 属于具有最高后验概率(在 $\boldsymbol{X}$ 条件下)的类.也就是说,朴素贝叶斯分类法预测 $\boldsymbol{X}$ 属于类 C_i,当且仅当

$$P(C_i/\boldsymbol{X})>P(C_j/\boldsymbol{X}),1\leqslant j\leqslant m,j\neq i,$$

这样,最大化 $P(C_i/\boldsymbol{X})$.其中 $P(C_i/\boldsymbol{X})$ 最大的类 C_i 称为最大后验假设.

根据贝叶斯定理,得到

$$P(C_i/\boldsymbol{X})=\frac{P(\boldsymbol{X}/C_i)P(C_i)}{P(\boldsymbol{X})}. \tag{1}$$

3) 由于 $P(\boldsymbol{X})$ 对于所有类为常数,只需要 $P(\boldsymbol{X}/C_i)P(C_i)$ 最大即可.如果类的先验概率未知,则通常假定这些类是等概率的,即 $P(C_1)=P(C_2)=\cdots=P(C_m)$,并据此对 $P(\boldsymbol{X}/C_i)$ 最大化.

4) 由于计算 $P(\boldsymbol{X}/C_i)$ 的开销可能非常大,为降低计算 $P(\boldsymbol{X}/C_i)$ 的开销,可以做类条件独立的朴素假定.给定元组的类标号,假定属性值有条件地相互独立,即在属性之间,不存在依赖关系.这样,

$$P(\boldsymbol{X}/C_i)=\prod_{k=1}^{k=n}P(x_k/C_i)=P(x_1/C_i)\times P(x_2/C_i)\times\cdots\times P(x_n/C_i), \tag{2}$$

可以容易地由训练元组估计概率 $P(x_1/C_i),P(x_2/C_i),\cdots,P(x_n/C_i)$.但是,独立性假设在许多实际问题中并不成立,如果忽视这一点,会引起分类的误差.

2 加权朴素贝叶斯文本分类

朴素贝叶斯分类方法认为所有条件属性对决策属

收稿日期:2011-10-20
* 通信作者:zdz@xmu.edu.cn

性的分类重要性是一致的(权重均为 1),这种方式使得冗余的、与分类无关的、相互影响的以及被噪声污染的特征和其他特征具有相同的地位,并使得分类的正确性降低,事实并非如此,有些因素对分类影响大一些,而另外的要小一些[3]. 基于上述观察,人们提出将各种特征加权算法与朴素贝叶斯分类器相结合,对不同的特征根据其分类重要性赋予不同的权值,使朴素贝叶斯扩展为加权朴素贝叶斯,以提高分类器的性能[3]. 加权朴素贝叶斯模型[3]大多为

$$C(x)=\underset{C_i\in C}{\operatorname{argmax}}P(C_i)\prod_{k=1}^{k=n}P(x_k/C_i)W_{j,k}\ , \tag{3}$$

其中 $W_{j,k}$ 是特征项 t_j 在类别 c_k 中的权重,权重越大,该特征项对分类的影响越大.

特征权重的计算方式[1]有很多种,比如布尔权重、词频权重、TF-IDF 权重等. 而 TF-IDF 权重应用最广泛,因为它将词频和反文档频率结合使用,克服了前面种种权重计算方法的缺点,TF-IDF 计算的归一化公式[4]如

$$W_i=\frac{\mathrm{TF}(t_i)\times\mathrm{IDF}(t_i)}{\sqrt{\sum_{i=1}^{n}(\mathrm{TF}(t_i)\times\mathrm{IDF}(t_i))^2}},$$

$$\mathrm{IDF}(t_i)=\log(\frac{N}{n_i}+L), \tag{4}$$

其中 $\mathrm{TF}(t_i)$ 是特征项 t_i 的词频,$\mathrm{IDF}(t_i)$ 是反文档频率[5],在它的计算公式中,L 的取值通过实验来确定. N 为文档集的总文档数,n_i 为出现特征项 t_i 的文档数. IDF 算法的核心思想是,在大多数文档中都出现的特征项不如只在一小部分文档中出现的特征项重要. IDF 算法能够弱化一些在大多数文档中都出现的高频特征项的重要程度,同时增强一些在小部分文档中出现的低频特征项的重要程度.

3　基于特征相关的改进加权朴素贝叶斯文本分类

3.1　特征值选取

在将文本表示成特征向量时,原始特征空间由出现在文本中的所有词条组成,文本分类问题所对应的原始特征空间通常都高达几万维,甚至更高. 如果直接在这样一个高维特征空间上进行分类器的训练和分类,不仅使文本自动分类的计算量过大,而且会使样本统计特征的估计变得非常困难. 因此,在分类器对训练文本进行训练之前,在不影响分类准确率的前提下,减少原始空间的维数(也称降维),将特征维数压缩到与训练文本个数相适应的情况[6]. 降维就是从原始特征空间中提取出部分特征的过程. 下面是本文采取的降维方法:

根据互信息的定义公式[7]

$$I(t,c)=\log\frac{p(t\wedge c)}{p(t)\times p(c)}, \tag{5}$$

其中 $p(t\wedge c)$ 为单词 t 和类别 c 同时出项的概率,$p(t)$ 为单词出现的概率,$p(c)$ 为类别 c 出现的概率. 它可近似为[7]

$$I(t,c)=(A/N)/[(B/N)\times(C/N)]=\frac{A\times N}{B\times C}, \tag{6}$$

其中 A 为 t 和 c 在训练集中的同现频率,N 为训练集中文本的数目,B 为 t 在训练集中出现的文本频数,C 为 c 在训练集中出现的文本频数.

由于互信息没有考虑到词频,所以可能会选中词频比较低的特征项,因此对互信息公式做了如下改进:

$$I_{\mathrm{new}}(t,c)=p(t)\times p(c/t)\times p(c)\times I(t,c), \tag{7}$$

其中 $p(t)$ 为 t 出现的概率(t 的词频与特征项总数之比),$p(c/t)$ 为 t 属于类别 c 的概率,$p(c)$ 为类别 c 的概率,该公式考虑到了频数,同时还考虑到特征在每个类别中的分布情况.

最后对训练集中的每个单词计算其互信息值,从原始特征空间中移除低于特定阈值的单词,保留高于阈值的词条作为特征值.

3.2　基于特征相关的改进权重计算方法(TF-IDF-FC 权重)

1) 计算测试文本的所有特征项的权重. 由于传统的 TF-IDF 权重[8]是将训练文本作为一个整体来进行考虑的,而没有考虑特征项在类间和类内中的分布情况这一重要信息,比如,如果某一特征项在某个类别大量出现,而在其他类别出现很少或者是某一特征项只在某个类别的一两篇文档中大量出现,而在类内的其他文档中出现的很少,这样的特征项的分类能力其实是很强的. 但这在 TF-IDF 算法中是无法体现的. 为了弥补这些缺陷,考虑到特征项对某类的影响程度与它在该类内的文档数成正比,与在其他类的文档数成反比,本文对 TF-IDF 权重的计算进行了改进,提出 TF-IDF-FC 权重,公式如下所示:

$$W_{j,k}=\frac{\mathrm{TF}(t_{j,k})\times\mathrm{IDF}(t_{j,k})}{\sqrt{\sum_{j=1}^{j=n}(\mathrm{TF}(t_{j,k})\times\mathrm{IDF}(t_{j,k}))^2}}, \tag{8}$$

$$\mathrm{IDF}(t_{j,k})=\log(\frac{n_k+1}{n_j-n_k+1}+L), \tag{9}$$

其中 $W_{j,k}$ 是特征项 t_j 在类 C_k 中的权重,值越大,越能代表该类,$\mathrm{TF}(t_j,k)$ 是特征项 t_j 在类 C_k 中的词频,n_k

是在类 C_k 中包含特征项 t_j 的文档数，n_j 是训练文本中所有包含特征项 t_j 的文档数，因此 n_j-n_k 是除类 C_k 外包含特征项 t_j 的文档数，L 的取值通过实验来确定，在本文中 L 取 2，式(9)中的加 1 是为了避免分子和分母为 0 的情况，从式(8)、(9)可以看出，特征项在类内文档数越大(类内分布越均匀)，在其他类的文档数越小(类间分布越不均匀)，则它的权重越大.

2) 计算测试文本特征项之间的互信息值(特征项相关度). 虽然能够赋予每个特征项不同的权重以示它们不再独立，但是不同的权重只是代表对类的贡献程度不同，它们之间的复杂联系没有考虑到. 比如，太极是属于宗教常出现的词语，老虎是动物，属于自然类，但它们同时出现时，很有可能指的是韩国(太极虎)，属于政治类. 因此，根据上面的互信息公式计算特征项之间的相关程度，此时，计算公式如下：

$$I(t,c,d)=\log\frac{p(t\wedge c\wedge d)}{p(t\wedge d)\times p(t\wedge c)}, \tag{10}$$

其中 c,d 代表 2 个特征项，可以用式(7)的近似代替，如下所示：

$$I(t,c,d)=(A/N)/[(B/N)\times(C/N)]=\frac{A\times N}{B\times C}, \tag{11}$$

只是现在 A 为 t,c,d 在训练集中的同现频率，N 为训练集中文本的数目，B 为 t,c 在训练集中出现的文本频数，C 为 t,d 在训练集中出现的文本频数.

根据式(11)计算测试文本之间的相关系数，设定阈值 λ，如果特征项 c 和 d 的互信息值小于 λ，它们之间就没有连接，否则将它们连接，由此最后能够得到一个无向图，每个特征项代表图的节点.

3) 调整相关特征项的权重. 根据 2)步构造出来的图，计算每个节点(特征项)的度数，将 1)步计算出的权重乘以该特征项的度数即为特征项的最终权重，度数越大，权重越大，则能代表该类的能力越强，也就是分类能力越强.

4) 将 3)步计算的权重代入式(3)，计算属于每类的概率，其中概率最大的即是测试文本所属类.

4 实验数据及分析

假设 TP_i 表示测试文档集中本来属于类别 c_i 而且被分类器分类到类别 c_i 的文档数，FP_i 表示测试文档集中本来不属于类别 c_i 但却被分类器错误分类到类别 c_i 的文档数，FN_i 表示本来应该属于类别 c_i 但被分类器分类到别的类别的文档数，而 TN_i 表示本来不属于类别 c_i 也没有被分类器分类到类别 c_i 的文档数. 那么，分类器在类别 c_i 上的召回率(R)定义为

$$R_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}, \tag{12}$$

分类器在类别 c_i 上的精确率(P)[8]定义为

$$P_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i}, \tag{13}$$

对于类别 c_i，其 $F1$(值越大，分类效果越好)定义为

$$F1=\frac{2\times R\times R}{R+P}. \tag{14}$$

本文采用的是复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的文本分类语料库，它分为训练语料和测试语料，均分为 20 类，其中训练语料共有 9 804 篇文档，测试语料共有 9 833 篇文档，基本是按照 1∶1 划分的，而在本实验中只取其中的 5 类，训练数据每类取 200 篇文章，共 1 000 篇文章，测试数据每类取 150 篇，共 750 篇文章. 下面将基于传统 TF-IDF 权重，改进后的 TF-IDF-FC 权重和文献[3]中的基于属性加权的朴素贝叶斯分类算法的实验结果进行对比，如表 1 所示.

从表 1 可以看出，改进后较改进前的算法分类效果明显要好，且较基于属性加权的朴素贝叶斯算法的分类效果也有一定的提高. 原因在于 TF-IDF-FC 权重计算考虑到了特征项在类间和类内的分布，类内分布越均匀，类间分布越不均匀，对类的贡献能力越大，因此它的权重越大，不仅如此，还根据特征项间的相关性，调整特征项权重，使得代表类的能力强的特征项的权重增大，分类效果越好.

5 结 论

特征项间独立性假设在很大程度影响了朴素贝叶斯文本分类的分类能力，为了克服这一问题，本文提出一种改进的特征项加权朴素贝叶斯文本分类方法，该算法的创新部分在于提出了一种改进的 TF-IDF 权重计算方法——TF-IDF-FC 权重计算方法，该方法是在传统 TF-IDF 权重计算方法的基础上，考虑到特征项在类内分布越均匀和类间分布越不均匀，则代表类的能力越大，权重越大，并且根据特征项之间的相关程度，对权重计算值进行了调整，使得对所属类贡献度较大的特征项的权重增加，有利于分类结果质量的提高. 另外，通过实验也表明本文采用的这种文本分类算法的分类效果不但较基于传统 TF-IDF 权重的加权朴素贝叶斯分类算法的分类有了很明显的提高，而且与文献[3]介绍的基于属性加权的朴素贝叶斯分类算法的

表 1　基于传统 TF-IDF 权重,属性加权和改进 TF-IDF 权重的加权朴素贝叶斯文本分类算法的实验效果对比

Tab. 1　Experimental results comparison of naive Bayes classification algorithm based on the traditional TF-IDF weight, attribute weight and improved TF-IDF weight

文档目标	基于属性加权的朴素贝叶斯分类算法			基于传统 TF-IDF 权重的加权朴素贝叶斯文本分类算法			基于特征相关的改进 TF-IDF 权重的加权朴素贝叶斯文本分类算法		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
美术(150)	0.9306	0.8933	0.9116	0.8966	0.8667	0.8814	0.9379	0.9066	0.9220
文学(150)	0.9252	0.9067	0.9159	0.9034	0.8792	0.8911	0.9388	0.9200	0.9293
教育(150)	0.9060	0.9000	0.9030	0.9085	0.8716	0.8897	0.9195	0.9133	0.9164
哲学(150)	0.9396	0.9333	0.9364	0.9128	0.9067	0.9097	0.9530	0.9467	0.9499
历史(150)	0.8875	0.9467	0.9117	0.8383	0.9333	0.8833	0.9057	0.9600	0.9321

分类效果有较大的提高,因此,研究此类算法意义还是比较大的.

参考文献:

[1] Han J W, Kamber M. 数据挖掘概念与技术[M]. 范明,孟小锋,译. 北京:机械工业出版社,2000:173-175.

[2] 程克非,张聪. 基于特征加权的朴素贝叶斯分类器[J]. 计算机仿真,2006,23(10):92-94.

[3] 秦锋,任诗流,程泽凯. 基于属性加权的朴素贝叶斯分类算法[J]. 计算机工程与应用,2008,44(6):107-109.

[4] 刘林. 基于词语权重改进的朴素贝叶斯分类算法的研究与应用[D]. 广州:中山大学,2009.

[5] 鲁明羽,李凡,庞淑英. 基于权值调整的文本分类改进方法[J]. 清华大学学报:自然科学版,2003,43(4):513-515.

[6] 罗海飞,吴刚,杨金生. 基于贝叶斯的文本分类方法[J]. 计算机工程与设计,2006,27(24):4746-4748.

[7] 郑伟. 文本分类特征选取技术研究[D]. 呼和浩特:内蒙古大学,2008.

[8] 徐凤亚,罗振声. 文本自动分类中特征权重算法的改进研究[J]. 计算机工程与应用,2005,41(1):181-184.

An Improved Weighted Naive Bayes Classification Algorithm Using Feature Correlation

RAO Li-li[1], LIU Xiong-hui[2], ZHANG Dong-zhan[1]*

(1. School of Information Science and Technology, Xiamen University, Xiamen 361005, China;
2. Department of Information Technology, Longyan Tobacco Industrial Co. Ltd, Longyan 364021, China)

Abstract: The strong independence condition between the feature required by naive Bayes classification algorithm is very difficult to realize in reality. This paper puts forward an improved weighted naive naive Bayes classification algorithm using feature correlation to loose this condition to some extent, this algorithm adopts a new weighting method called TF-IDF-FC weight calculation, it takes into account the feature distribution within and between class based on the traditional TF-IDF weight calculation method and adjusts feature weight in combination with feature correlation in order to make the weight of the feature which can represent its class mostly. Compared with weighted naive Bayes classification based on the traditional TF-IDF weight and other commonly used weighted naive Bayes classification algorithms, such as attribute weighted naive Bayes classification, this algorithm improve the performance of classification to a certain extent.

Key words: naive Bayes text classification; weighted naive Bayes text classification; TF-IDF weight; feature correlation