词间相关性在贝叶斯文本分类中的应用研究

章舜仲[1,2],王树梅[1],黄河燕[3],陈肇雄[3]
ZHANG Shun-zhong[1,2],WANG Shu-mei[1],HUANG He-yan[3],CHEN Zhao-xiong[3]

1.南京理工大学 计算机科学系,南京 210094
2.南京财经大学 电子商务系,南京 210046
3.中国科学院 计算机语言信息工程研究中心,北京 100083
1.Department of Computer Science,Nanjing University of Science and Techology,Nanjing 210094,China
2.Department of Electronic Business,Nanjing University of Finance and Economics,Nanjing 210046,China
3.Computer Language Information Engineering Research Center,Chinese Academy of Sciences,Beijing 100083,China
E-mail:zszws@jlonline.com

ZHANG Shun-zhong,WANG Shu-mei,HUANG He-yan,et al.Research on application of word correlation in Naive Bayes text classification.Computer Engineering and Applications,2009,45(16):159-161.

Abstract: Aiming at the deficiency of Naive Bayes' attribute independence assumption,the concept of correlation and that between multi-variations were discussed,and the definition of correlation degree between terms was presented.Based on the analysis of the correlation between terms of TAN classifier,authors proposed a fomula to evaluate the correlation degree between document feature words and the algorithm of its application to ameliorating Naive Bayesian classifier.The experiments on Reuters-21578 collection show the improvement of algorithm to be simple,effective and easy to implement.

Key words: text classification;Naive Bayes;event correlation;correlation degree;Tree Augmented Naive Bayes(TAN) classifier

摘 要:针对朴素贝叶斯分类的属性独立性假设的不足,讨论了相关性及多变量相关的概念,给出词间相关度的定义。在TAN分类器的词间相关性分析基础上,提出一种文档特征词相关度估计公式及其在改进朴素贝叶斯分类模型中应用的算法,在Reuters-21578文本数据集上的实验表明,改进算法简单易行,能有效改进贝叶斯分类性能。

关键词:文本分类;朴素贝叶斯;事件相关;相关度;树扩展型朴素贝叶斯分类器
DOI:10.3778/j.issn.1002-8331.2009.16.046 **文章编号**:1002-8331(2009)16-0159-03 **文献标识码**:A **中图分类号**:TP311

1 引言

自动文本分类指计算机将一篇文章自动地分派到一个或多个预定义的类别中去,其过程通常包括两步:第一步,将一组预先分好类的文档作为训练集,并在训练文档的基础上构造分类器;第二步是利用获得的分类器对未知类别的文档进行归类。目前较为常见的文本分类器构造算法有Rocchio、贝叶斯、KNN、决策树、SVM等。其中贝叶斯分类方法由于具有坚实的数学理论基础并能综合先验信息和数据样本信息,一直是文本分类的研究热点之一。

朴素贝叶斯(Naïve Bayes)分类器是贝叶斯分类器的一种,是一种简单而有效的概率分类方法,其性能可与决策树、神经网络等算法相媲美,在某些领域甚至表现更为优越[1]。然而它的属性独立性假设无法表示属性之间的依赖关系,影响了分类性能,于是关于朴素贝叶斯分类器的改进算法研究引起人们持续的关注。TAN(Tree Augmented Naive Bayes)分类器[2-3]放松了独立性假设条件,它采用树形结构近似分解以类变量为条件的条件概率,有效改进了朴素贝叶斯分类器的分类性能。关于TAN结构的进一步改进与扩展的已有研究如BAN(Bayesian Network Augmented Naive Bayes)[4]允许属性之间可以形成任意的有向图,使其表示依赖关系的能力增强。

本文提出了一种新的基于词间相关性表示的贝叶斯文本分类方法,通过词间两两相关性对文档特征向量词间多相关性进行评估,实验表明,该方法能够稳定有效地改善贝叶斯文本分类器的性能。

2 朴素贝叶斯文本分类模型

设有 m 个类 $C_1,\cdots,C_m$,词全集 T,文档表示为其特征词向量 $\boldsymbol{X}=\{t_1,t_2,t_3,\cdots,t_n\}$,$\boldsymbol{X}\subset T$。按贝努利模型,$\boldsymbol{X}$ 中特征词 t_k 以布尔值表示该词是否出现在文档中,而词的出现次数则不被考虑。根据贝叶斯定理,有 $P(C_j|X)=\frac{P(X|C_j)*P(C_j)}{P(X)}$,$(j=1,\cdots,m)$。$P(C_j|X)$ 表示待分类文本 X 属于 C_j 的概率,$P(X)$ 是 X 的先验概

作者简介:章舜仲(1972-),男,博士研究生,讲师,主要研究领域为模式识别、机器学习、数据挖掘;王树梅(1957-),女,博士,教授,主要研究领域为数据库、信息处理;黄河燕(1963-),女,博士,研究员,主要研究领域为自然语言处理与机器翻译、大型智能应用系统;陈肇雄(1961-),男,博士,研究员,主要研究领域为自然语言处理、大型智能应用系统。

收稿日期:2008-04-01 **修回日期**:2008-06-11

率,其含义可解释为:任意文本恰为 X 的概率。$P(C_j)$表示任意文本属于类 C_j 的概率。$P(X|C_j)$是在条件 C_j 下,X 的后验概率,其含义为:已知任意文本属于类 C_j,该文本恰为 X 的概率。贝叶斯分类预测文档 X 所属类别 $C^*=\arg\max(C_j)\{P(C_j|X)\}$,即 $P(C_j|X)$最大的类,由于 $P(X)$对于所有类为常数,$P(C_j|X)\propto P(X|C_j)*P(C_j)$,分类模型如下:

$$C^*=\arg\max(C_j)\{P(t_1,t_2,\cdots,t_n|C_j)*P(C_j)\} \quad (1)$$

式(1)中 $P(t_1,t_2,\cdots,t_n|C_j)$和 $P(C_j)$可由训练样本进行估计,设训练文本总数 S,其中属于类 C_j 的为 S_j,则有:

$$P(C_j)=S_j/S$$

$$P(t_1,t_2,\cdots,t_n|C_j)=S_j(t_1,t_2,\cdots,t_n)/S_j \quad (2)$$

式(2)为特征词集联合概率公式在训练集上的估计,$S_j(t_1,t_2,\cdots,t_n)$为类 C_j 训练文本中同时包含词 $t_1,t_2,\cdots,t_n$ 的文本数,在文本分类的高维特性,特征词组合的数量近于无穷,因此 $S_j(t_1,t_2,\cdots,t_n)$无法预先计算,只能分类时在训练集上直接计算,如果训练集不是足够大,$S_j(t_1,t_2,\cdots,t_n)$一般为 0,而如果训练集太大,计算 $S_j(t_1,t_2,\cdots,t_n)$的开销非常大,因此为简化计算,朴素贝叶斯假定属性相互间类条件独立,则词集联合概率为词概率乘积:

$$P(t_1,t_2,\cdots,t_n|C_j)=\prod_{k=1}^{n}P(t_k|C_j)=\prod_{k=1}^{n}S_j(t_k)/S_j \quad (3)$$

$S_j(t_k)$为类 C_j 的训练文本中包含词 t_k 的文本数。

3 基于词间相关性分析的改进贝叶斯分类模型

3.1 相关性分析

相关性一般狭义指随机变量间的线性相关程度,常局限为两个随机变量间的线性相关,如变量 x 和 y,则其相关系数定义为:$\rho(x,y)=\dfrac{Cov(x,y)}{\sqrt{D(x)D(y)}}$。对于多个变量间的相关系数的定义及其应用已有一些研究[5-6],然而由于文本分类问题的高维复杂性,在应用过程中,未能取得较好效果,因此本文关于词间相关性定义采用广义的基于事件概率相关。

若事件 A 与 B 满足 $P(AB)=P(A)P(B)$,则称 A 与 B 相互独立,若 $P(AB)>P(A)P(B)$,则称 A 与 B 正相关,反之为负相关,由于在分类问题中,类别特征往往由正相关属性决定,从而以下“相关性”均指正相关。A 和 B 之间的相关性通过计算 $Corr_{A,B}=\dfrac{P(AB)}{P(A)P(B)}$来度量,称 $Corr_{A,B}$ 为相关度,推广至 n 个事件有:$Corr(A_1,A_2,\cdots,A_n)=\dfrac{P(A_1A_2\cdots A_n)}{P(A_1)P(A_2)\cdots P(A_n)}$。

3.2 TAN 分类器的词间相关性分析

将事件相关性应用于贝叶斯分类模型,得以下改进贝叶斯分类模型:

$$C^*=\arg\max(C_j)\left\{Corr(t_1,t_2,\cdots,t_n|C_j)*\prod_{k=1}^{n}P(t_k|C_j)*P(C_j)\right\} \quad (4)$$

$Corr(t_1,t_2,\cdots,t_n|C_j)$表示文档特征词集在类 C_j 中的相关度,从以上对式(2)的分析可知在训练集直接计算 $Corr(t_1,t_2,\cdots,t_n|C_j)$不可行,为此需要一种文档相关度估计方法。引入贝叶斯网络的属性在父节点集上的条件独立性假设,对 TAN 分类模型的联合概率公式作如下转换:

$$P(t_1,t_2,\cdots,t_n)=\prod_{k=1}^{n}P(t_k|\Pi_{t_k})=\prod_{k=1}^{n}\frac{P(t_k,\Pi_{t_k})}{P(\Pi_{t_k})}=$$

$$\prod_{k=1}^{n}\frac{P(t_k)P(t_k,\Pi_{t_k})}{P(t_k)P(\Pi_{t_k})}=\prod_{k=1}^{n}P(t_k)Corr(t_k,\Pi_{t_k})=$$

$$\prod_{k=1}^{n}Corr(t_k,\Pi_{t_k})\prod_{k=1}^{n}P(t_k) \quad (5)$$

Π_{t_k} 为 t_k 在 TAN 分类器树形结构上的父节点,转换公式表明 TAN 分类器事实上属于式(4)的一种特例,根据其父节点选择的条件互信息函数,TAN 分类器选择了与 t_k 强相关词的相关度计算词集相关度,即 $Corr(t_1,t_2,\cdots,t_n|C_j)=\prod_{k=1}^{n}Corr(t_k,\Pi_{t_k})$。由于 TAN 分类器只选择了一个与特征词强相关的父节点,不能较全面地表示特征词与词集的相关程度,有时会产生较大误差,造成分类效果反而下降,而不限定父节点个数的贝叶斯网络结构又存在学习难度太大的缺点。为此,提出基于词集相关度估计的贝叶斯分类模型。

3.3 基于词间相关性的改进贝叶斯分类

通过对 TAN 分类器的词间相关性分析可知,特征词集相关度可通过两两词相关度估算,表示为如下公式:

$$P(t_1,t_2,\cdots,t_n|C_j)=f(\prod Corr(t_i,t_k|C_j))*\prod_{k=1}^{n}P(t_k|C_j) \quad (6)$$

式(6)中 f 函数表示用于估计文档相关度计算的对两两词相关度的运算,在式(5)中 f 即为所有特征词与其父节点相关度的乘积,将其扩展后得到改进贝叶斯分类模型为:

$$C^*=\arg\max(C_j)\left\{[\prod_{Corr>\lambda}Corr(t_i,t_k|C_j)]^{\frac{\alpha}{n}}\times\prod_{k=1}^{n}P(t_k|C_j)\times P(C_j)\right\} \quad (7)$$

式(7)中$[\prod_{Corr>\lambda}Corr(t_i,t_k|C_j)]^{\frac{\alpha}{n}}$即特征词集相关度的估计式,表示所有大于阈值 λ 的两两相关度乘积的 $\dfrac{\alpha}{n}$ 次方,α 为指数调节参数,取值为 1。阈值 λ 可以取固定值,一般大于 1,这里采用分类过程中动态调整的方法确定阈值 λ,分类模型描述如下。

设待分类文档 X 表示为 n 维特征词集 $X=\{t_1,t_2,\cdots,t_n\}$,两两特征词对组成的集合记为 $X^2=\{t_it_j|i,j=1,\cdots,n\}=\{ts_i|i=1,\cdots,\dfrac{n(n-1)}{2}\}$,$ts_i$ 为第 i 个词对。在分类器学习阶段计算各个类上词概率 $P(t_i|C_j)$,$t_i\in X$,以及两两特征词对的相关度 $Corr(ts_i|C_j)$,$ts_i\in X^2$。本文实现分类算法描述如下。

(1)根据朴素贝叶斯分类,找到文档在各个类上的概率值最大的三个类,分别记为 C_1、C_2、C_3。

(2)对 C_1、C_2、C_3 上的词相关度作降序排序,得词相关度集 $CORR_1=\{Corr(ts_{i_1}|C_1)\}$、$CORR_2=\{Corr(ts_{i_2}|C_2)\}$、$CORR_3=\{Corr(ts_{i_3}|C_3)\}$。令 i_1 初值为 1,作以下循环步骤。

(3)令阈值 $\lambda=Corr(ts_{i_1}|C_1)$,若 $\lambda\leqslant 1$,$C^*=C_1$,转(7),否则转(4)。

(4)按式(7)分别计算类 C_1、C_2、C_3 的后验概率 $P(C_1|X)$、$P(C_2|X)$、$P(C_3|X)$。

(5)若 $P(C_2|X)>P(C_1|X)$或 $P(C_3|X)>P(C_1|X)$,则 $C^*=C_2$ 或 C_3,转(7),否则转(6)。

(6)i_1 值加 1,转(3)。

(7)判文档 X 属于类 C^*。

以上改进的贝叶斯分类算法之所以在朴素贝叶斯概率值最大的三个类中进行,是因为通过实践发现,若朴素贝叶斯分类产生错分的情况,实际所属类一般仍为概率值最大的前几类,尤其是属于概率值第二类的情形占较大比例。取前三类进行改进,一方面可减少计算,另一方面文档若实际所属类的概率值相比其他类偏小,则该文档很可能属于噪声样本或专家错判情形,可不考虑。

4 实验结果与分析

4.1 实验数据集

本文实验在 Reuters-21578[7]数据集上进行,去除了其中具有多个类别的文档,选择了其中的 10 个类共 4 400 篇文档组成数据集,如表 1 所示。

表 1 Reuters-21578 上的实验文档集

Class Name	Document Count	Training Document Count on LEWISSPLIT	Test Document Count on LEWISSPLIT
acq	2 362	1 666	696
crude	408	287	121
trade	361	286	75
money-fx	307	220	87
interest	285	203	82
money-supply	161	133	28
ship	158	122	36
sugar	143	118	25
coffee	116	94	22
gold	99	79	20

实验基于 DF+IG 方法进行特征词提取,有研究表明组合的特征提取方法可以获得更好的分类效果[8]。表 1 中的训练文本和测试文本为 David D.Lewis 所作的划分。先在 Lewis 划分的训练和测试集上进行测试,而后又采用 10-折交叉测试的方法进行了实验,实验评测指标为查准率、查全率、F1 及其宏平均和微平均值[9],实验结果及分析如下。

4.2 实验结果与分析

实验中,分别使用两种数据集划分方法进行测试,表 2 为使用 Lewis 划分的测试结果,表 3 为基于 10-折交叉测试方法测试的结果。实验数据显示使用这两种方法测试的改进贝叶斯分类在各类别上的指标以及宏平均和微平均指标均有提高,说明了本文提出的词间相关度计算及分类算法在提高贝叶斯分类器性能上的应用是有效的。

表 2 基于 LEWISSPLIT 划分的 Reuters 数据集实验结果

Class	Naive Bayes			Improved Bayes		
	Precision	Recall	F1	Precision	Recall	F1
acq	0.995	0.931	0.962	0.991	0.945	0.968
crude	0.744	0.983	0.847	0.779	0.992	0.873
trade	0.562	0.973	0.712	0.640	0.973	0.772
money-fx	0.607	0.782	0.683	0.622	0.851	0.718
interest	1.000	0.402	0.574	0.973	0.439	0.605
money-supply	0.957	0.786	0.863	0.957	0.786	0.863
ship	0.957	0.611	0.746	1.000	0.639	0.780
sugar	1.000	0.800	0.889	1.000	0.800	0.889
coffee	0.870	0.909	0.889	0.833	0.909	0.870
gold	0.824	0.700	0.757	0.929	0.650	0.765
Macro-Average	0.851 4	0.787 8	0.818 3	0.872 4	0.798 4	0.833 7
Micro-Average		0.871 6			0.888 4	

表 3 Reuters 数据集上的 10 折交叉验证测试结果

	Naive Bayes			Improved Bayes		
	Precision	Recall	F1	Precision	Recall	F1
Macro-Average	0.880 1	0.860 9	0.870 2	0.892 6	0.880 9	0.886 6
Micro-Average		0.891 1			0.910 7	

5 结论与展望

本文基于词间相关性分析,提出了由文档相关度估计而改进贝叶斯分类的方法。通过对 TAN 分类器的词间相关性分析,指出 TAN 分类器实质是基于文档相关度估计的一种特殊情形。在此基础上,提出了较为一般化的基于文档特征词两两相关度运算的文档相关度估计公式及其用于改进朴素贝叶斯文本分类的算法。在标准数据集 Reuters-21578 上,采用多种方法生成和处理实验数据集,进行了较为完备的测试,4.2 节中实验数据显示该方法能有效提高分类性能,而更多的实验表明该方法对于改进朴素贝叶斯分类性能具有较好的稳定性。

本文提出的基于文档特征词相关度计算的改进贝叶斯分类模型中,关于文档表示只考虑了特征词在文档中出现与否,未将词频考虑在内,而对于线性多相关系数的定义及其在分类中的应用,尚未有较好的实现方法,这些在今后的工作中,有待进一步深入研究。

参考文献:

[1] Langley P,Iba W,Thompson K.An analysis of Bayesian classifiers[C]//Proceedings of the Tenth National Conference on Artificial Intelligence.Menlo Park,USA:AAA I Press,1992:223-228.

[2] Fried N,Geiger D,Goldszmidt M.Bayesian network classifiers[J].Machine Learning,1997,29(2/3):131-163.

[3] Ramoni M,Sebastiani P.Robust Bayes classifiers[J].Artificial Intelligence,2001,125(122):209-226.

[4] Cheng J,Greiner R.Comparing Bayesian network classifiers[C]//Laskey K B,Prade H.Proc of the 15th Conf on Uncertainty in Artificial Intelligence.San Francisco:Morgan Kaufmann Publishers,1999:101-108.

[5] Susumu T.A study on multi relation coefficient among variables[J].Proceedings of the School of Information Technology and Electronics of Tokai University,2004,4(1):67-72.

[6] Bocchieri E,Mark B.Subspace distribution clustering hidden Markov model[J].IEEE Transactions on Speech and Audio Processing,2001,9(3):264-275.

[7] Lewis DD.Reuters-21578 text categorization test collection distribution 1.0[EB/OL].(1997-09).http://www.daviddlewis.com/resources/testcollections/reuters21578.

[8] 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004,18(1):27-33.

[9] Sebastiani F.Machine learning in automated text categorization[J].ACM Computer Survey,2002,34(1):1-47.