[文章编号] 1003－4684(2017)04-0051-04

一种用于微信信息分类的改进贝叶斯算法

张颖江，库凯琳
（湖北工业大学 计算机学院，湖北 武汉 430068）

[摘 要] 微信的快速普及加快了信息的传播，随之而来的广告、诈骗等信息严重困扰人们的生活。针对朴素贝叶斯对信息分类时考虑所有特征并将特征赋予相同权值两方面的缺陷，提出一种用于微信信息分类的改进贝叶斯算法。采用改进的互信息进行特征选择，提取关键特征，通过改进 TFIDF 对特征加权，优化朴素贝叶斯的分类性能。实验结果表明，改进的贝叶斯算法能有效选择关键特征属性，提高微信信息分类的精准度。

[关键词] 贝叶斯；微信信息；特征提取；特征加权；信息分类
[中图分类号] TP391.1 [文献标识码] A

随着微信已成为人们日常交流和沟通的一种重要方式，微信平台的信息安全问题急需解决。目前对微信信息的监管主要是通过设置黑名单的形式，即大量收集传播垃圾信息的微信用户 ID，并将其加入黑名单来阻断信息的传播。但由于微信用户量大，增长速度快等特点，传统的设置黑名单的方式很难从源头上杜绝垃圾信息的产生。并且实施周期长，工作量大，效果显微。

微信信息的处理本质上是对文本信息的处理。常见的文本分类器包括决策树(Decision Tree)[1]分类器、朴素贝叶斯(Naive Bayesian)[2]分类器、支持向量机(Support Vector Machine)[3]等。朴素贝叶斯分类器具有训练和分类速度快的特点，许多学者对其进行深入研究并提出了一些改进方法。邓桂骞提出一种条件属性相对于决策属性的相关性和重要性的属性权值计算方法[4]；李静梅提出一种通过 EM 算法(期望值最大算法)，自动增加训练量，得到较为完备的训练库，提高朴素贝叶斯的分类精度[5]；赵文涛，孟令军等人针对朴素贝叶斯算法下溢问题，对算法基本公式进行优化改进，提出一种新的 CIT-NB 算法并通过实验验证其分类性能远优于朴素贝叶斯分类器[6]，以上改进方法都是针对朴素贝叶斯属性权值相同进行的改进。

针对朴素贝叶斯属性选择和属性加权两方面的缺陷，提出采用互信息对文本特征属性进行选择，并针对传统互信息的缺陷提出改进措施，对选取的特征属性采用改进的 TF-IDF 加权[7]，将改进后的贝叶斯算法用于微信信息分类。与传统贝叶斯算法、基于 TF-IDF 的朴素贝叶斯算法相比，该算法在查准率和查全率方面都有显著提升。

1 朴素贝叶斯分类

朴素贝叶斯(NB)分类是一种十分简单的分类算法，其基本思想可概括为：对给出的待分类项，求解出在该分类项出现的情况下各类别出现的概率，并认为该分类项属于类别中概率最大的类，朴素贝叶斯有如下定义：

设 $\boldsymbol{A}=\{a_1,a_2,a_3,\cdots,a_m\}$ 为一个待分类样本，每个 a_i 为 $\boldsymbol{A}$ 的一个特征属性。有类别集合 $\boldsymbol{B}=\{b_1,b_2,b_3,\cdots,b_n\}$，计算 $p(b_1\mid a)$，$p(b_2\mid a)$，$p(b_3\mid a)$，$p(b_4\mid a)$，如果

$$p(b_k\mid a)=\max\{p(b_1\mid a),p(b_2\mid a),p(b_3\mid a),\cdots,p(b_n\mid a)\} \tag{1}$$

则 $a\in b_k$。根据贝叶斯定理：

$$p(b_i\mid a)=\frac{p(a\mid b_i)P(b_i)}{p(a)} \tag{2}$$

$p(a)$ 对所有类均为常数，最大化后验概率 $p(b_i\mid a)$ 可转化为最大化先验概率 $p(a\mid b_i)P(b_i)$。若训练数据集有许多属性和元组，假设各属性相对类别条件独立，即：

$$p(a\mid b_i)p(b_i)=p(a_1\mid b_i)p(a_2\mid b_i)\cdots p(a_m\mid b_i)p(b_i)=p(b_i)\prod_{j=1}^{m}p(a_j\mid b_i) \tag{3}$$

[收稿日期] 2016－06－06
[第一作者] 张颖江(1959－)，男，北京人，湖北工业大学教授，研究方向为计算机网络安全
[通信作者] 库凯琳(1991－)，男，湖北黄冈人，湖北工业大学硕士研究生，研究方向为信息处理与网络安全

为避免(3)式中出现乘积为0的结果,需要对公式进行平滑处理,常用的方法是对式(3)进行拉普拉斯平滑:

$$p(b_i) \approx \frac{1+N(b_i)}{N_C+\sum_k N(b_k)} \tag{4}$$

$$p(a_i \mid b_j) \approx \frac{1+N_{ij}}{M+\sum_k N(c_{kj})} \tag{5}$$

其中:$N(b_i)$ 表示 b_i 的文档个数,$\sum_k N(b_k)$ 表示集中文档总个数,N_{ij} 表示特征属性 a_i 在类别 b_j 的文档中出现的次数,M 表示特征词个数,$\sum_k N(c_{kj})$ 表示 b_j 类中所有词出现的次数。

计算过程中由于 $p(a_i \mid b_j)$、$p(b_i)$ 的值都很小,两者相乘的结果偏小会导致精度下降,常用的做法是取对数进行运算,可减少计算开销并提高了计算结果的精度。

朴素贝叶斯在分类时有两个条件较为苛刻:(1)条件独立性假设:NB分类假设样本各特征之间相互独立,该假设在实际应用中往往是不成立的,从而影响了分类的正确性。(2)各属性权值都设为1:NB分类算法认为样本各属性具有相同的权值,但是在实际分类中,不同类别的样本属性出现的概率并不相同,将所有属性的权值设为1,影响NB算法的精准性。本文针对以上NB算法的局限性提出一种改进的贝叶斯算法用于微信信息的分类,并将实验结果进行对比,给出结论。

2 改进的贝叶斯算法

2.1 特征提取

微信信息可看作为文本特征向量,原始的特征向量空间由出现在微信信息中的所有词组成,如果将出现的所有词都作为特征向量的话,那么特征向量空间通常会维度过高,若直接对这种高维度的样本空间进行训练和分类,需要很大的计算开销,且无用的特征信息参与计算会导致分类不精准。通常在对样本进行训练前,在不影响分类精准性的前提下尽可能剔除无用的特征属性,这个过程称为降维。在特征提取前,首先对特征向量空间人工降维,去掉一些常用的数词、量词、语气词等,这类词对分类结果没有影响,但是会造成很大的计算开销。

本文采用MI互信息(Mutual Information)[8]进行特征提取,MI计算词元 a_i 与类别 b_i 之间的相关度作为特征提取的标准,定义如下:

$$\mathrm{MI}(a,b_i) = \log \frac{p(a \cap b_i)}{p(a) \times p(b_i)} \tag{6}$$

其中 $p(a)$ 为单词出现的概率,$p(b_i)$ 为类别 b_i 出现的概率,$p(a \cap b_i)$ 为两者同时出现的概率。式(6)可简化为:

$$\mathrm{MI}(a,b_i) \approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{7}$$

N 表示训练集中样本总数,A 表示词元 a 与类别 b_i 同时出现的次数,B 为词元 a 出现类 b_i 不出现的次数,C 为类 b_i 出现但词元 a 不出现的次数,式(7)表示的词 a 与各类别 b_i 之间的互信息。

若存在 m 个类别,则特征项 a 与 m 个类别分别有 m 个互信息,常用的方法是求平均互信息,平均互信息的计算公式为:

$$\mathrm{MI}(a) = \sum_{i=1}^{m} p(b_i) \log \frac{p(a \cap b_i)}{p(a) \times p(b_i)} \tag{8}$$

传统互信息在特征提取时,有两个突出的缺陷:(1)不考虑词频,更趋向于低频词汇;(2)忽略负相关特征,特征与类别的负相关部分会导致该特征的权值降低[8]。针对以上两个缺陷带来的分类性能较低的问题,本文采用改进的互信息进行特征提取。

首先,在特征提取时考虑词频信息。一般认为:词频越高、集中度越强的词对文本分类作用越大,而传统互信息在进行特征选择时通常是词频较小的词获得较大的互信息,与实际预期的情况不符,因此将词频、集中度等因素考虑进去采用改进的互信息进行特征提取。为此引入特征词 a 对于类别 b_i 的先验概率 $p(a \mid b_i)$,表示特征词 a 在 b_i 中出现的频度,其参数公式表示为 $X = p(a \mid b_i)$,同时引入后验概率 $p(b_i \mid a)$,表示特征词 a 出现同时又属于类别 b_i 的概率来表示其集中度,其参数公式表示为:

$$Y = p(b_i \mid a)$$

其次,在计算互信息时会出现负数的情况,特征项 a 与类别 b 不相关时,互信息为0,当特征项 a 在类别 b 中很少出现时,互信息为负值,这称为特征项 a 与类别 b 负相关。在计算中负相关会降低分类效果,但在实际情况中,一个特征若出现在少数类别中,这对分类具有较大的区分性。分析式(6)可知,在 $p(a \cap b_i)$ 较小而 $p(a)$ 较大时,$\mathrm{MI}(a)$ 取值为负,且 $\mathrm{MI}(a)$ 的值随 $p(a)$ 变大及 $p(a \cap b_i)$ 变小而变小,但在实际情况中,特征 a 的区分性随着 $p(a)$ 变大及 $p(a \cap b_i)$ 变小而变大。针对此情况,本文对互信息取绝对值进行计算,该做法相对更为合理。综合传统互信息提取特征的缺陷来考虑,本文提出一种改进的互信息方案来提取特征,新的互信息计算公式可以表示为:

$$\mathrm{MI}^*(a,b) = X \cdot Y \cdot \left| \log \frac{p(a \cap b_i)}{p(a) \times p(b_i)} \right| \tag{9}$$

通过限定阈值,就可以对互信息选取特征的缺陷进行很好地补充。

2.2　特征加权

传统贝叶斯算法中认为所有的特征属性具有相同的权值(即将所有的属性权值设为 1)，但在实际应用中，应赋予特征属性不同的权值。如果某个特征在某个文本中重复出现，而在别的文本中很少出现，就可以认为该特征具有较好的区分性，应赋予其较大的权值。特征权值的计算方法有很多，词频权值、布尔权值、TFIDF 都是常用的计算权值的方法[9]。TFIDF[10]用词频乘以逆文档来表示特征项的权值。TF 表示特征项 t 在文档 d 中出现的频率，IDF 表示特征项 t 在整个文档集中的分布量化[11]。TFIDF 公式：

$$w_{ik} = TF \cdot IDF = TF \cdot \frac{1}{DF} = tf_{ik} \cdot \log \frac{N}{n_k} \quad (10)$$

其中，w_{ik} 表示文档 i 中第 K 维的向量值，tf_{ik} 表示文档 i 中第 K 个特征值的 TF 值，N 表示文本集的文档数，n_k 表示文本集中出现该特征项的文本数。将 TFIDF 归一化后得

$$W_{ik} = \frac{tf_{ik} \times \log(\frac{N}{n_k} + L)}{\sqrt{\sum_{i=1}^{n} (tf_{ik})^2 \times [\log(\frac{N}{n_k} + L)]^2}} \quad (11)$$

TFIDF 的局限性表现在：若某一特征在某一类别文档中大量出现，而在其他类别文档中很少出现，或者某个特征只在某一类别的少量文档中大量出现而在别的文档中很少出现，这样的特征应该具有很好的区分性，应该赋予较大的权值，但实际上在式(11)中却不能体现出来，相反 IDF 更趋向于赋予少量出现的特征较大的权值，这与实际情况并不相符。TFIDF 考虑特征与文档之间的信息，而忽略特征与类别的关系。为此，本文将特征的类别信息考虑进 TFIDF 进行特征加权，提出一种信息特征加权函数 TFIDF*。改进的 TFIDF* 函数引入特征类函数 K，其中

$$K_i = \frac{c_{\max}}{n_k} \quad (12)$$

该函数考虑词条 i 在某一具体类中的文档出现的概率。改进的 TFIDF* 函数定义如下：

$$\text{TFIDF}^* = tf_{ik} \cdot \log \frac{N}{n_k} \cdot \frac{c_{\max}}{n_k} \quad (13)$$

式中，n_k 表示包含词条 i 的文档总数，$c_{\max}$ 表示包含特征 i 最多的类中文档数量。特征类函数设计的思想就是将特征与文本类别结合考虑，弥补 TFIDF 只考虑文本特征与数量的不足。如果特征在某一类文本中出现的次数较多，那该特征就可以很好地代表该类别，该类加权的结果值就大，因此，特征 i 在某一类中越重要，特征类函数权值也就越大。

3　实验结果与分析

为验证改进算法的可行性和有效性，设计实验采用本文改进的互信息对特征属性进行选择，并对选择后的特征采用改进 TF-IDF 赋予不同权值，通过贝叶斯公式计算概率得到最终的分类结果。实验数据集主要采用 python 网络爬虫和人工搜索的方式收集。数据来源是一些微信公众号推送的消息，共收集测试信息包含体育、健康、教育、科技、军事、文化等 6 类信息共 2264 条，采用其中的 1509 条作为训练集，其余的 755 条作为测试集用于测试。各信息分布情况见表 1。

表 1　数据集

类别	训练集	测试集
广告	255	127
房产	289	146
色情	324	160
招聘	247	124
诈骗	188	95
文化	206	103

本文采用查全率 R 和查准率 P 作为评测指标，计算公式如下：

$$R = \frac{\text{信息所属的正确类别数目}}{\text{该信息所属的所有类别数目}}$$

$$P = \frac{\text{信息所属的正确类别数目}}{\text{判断为该信息所属的所有类别数目}}$$

查全率和查准率从两个不同方面反映了分类器的性能。

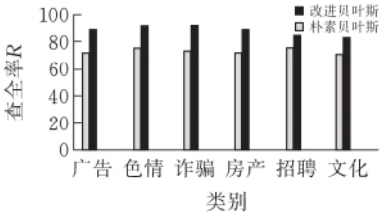

图 1　查全率

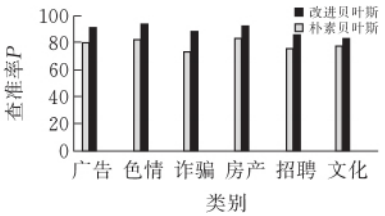

图 2　查准率

从图 1 和图 2 的测试结果中可以看出：采用朴素贝叶斯对样本进行分类，其分类效果明显低于改进后的贝叶斯的性能。分析可知其原因主要有以下几个方面：本文在朴素贝叶斯分类的时候采用的是未改进的互信息方法挑选特征值，同时采用朴素贝叶斯的思想，认为所有的属性权值相同，全部设为

1，导致其分类结果不尽人意，而本文提出的改进的贝叶斯采用改进的互信息提取特征，考虑互信息的负相关影响，将词频因素考虑进特征提取，同时使用改进的 TFIDF 对提取出的特征加权，从测试结果可以看出分类性能高于改进之前。

为验证本文提出的改进贝叶斯算法的稳健性，采用不同数量的训练集作为样本进行训练，观察训练样本数量对分类准确性的影响，实验结果见图 3。

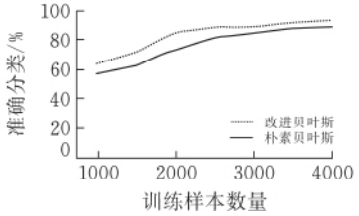

图 3 样本数量对分类结果的影响

从图 3 可以看出，随着训练样本的不断增加，改进的贝叶斯分类器趋于稳定，并保持较高的分类性能。而朴素贝叶斯分类器在训练样本增加之后性能也得到提高，但整体分类性能没有改进贝叶斯分类性能高。可以看出，朴素贝叶斯在大样本训练集中具有较好的分类性能。而改进贝叶斯在小样本区间也具有较好的分类性能。

4 结语与展望

采用改进的贝叶斯分类器对微信信息进行分类，使用改进的互信息提取特征，然后使用改进的 TFIDF 对特征进行加权，最终较朴素贝叶斯取得了更好的分类效果。改进的贝叶斯算法在小样本区间也表现出较好的分类性能，可以将本方法及后续改进方法用于垃圾邮件检测和入侵检测系统中。

[参 考 文 献]

[1] Carreras X, Marquez L. Boosting trees for anti-spam email filtering[J]. 2001, 3(4):1306-1311.

[2] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An evaluation of naive bayesian anti-spam filtering [J]. Tetsu-to-Hagane, 2000, cs.cl/0006013(2):9-17.

[3] Cortes C, Vapnik V. Support vector networks[J].Machine Learning.1995,20(1):273-329.

[4] 邓桂骞，赵跃龙.一种优化的贝叶斯分类算法[J].计算机测量与控制，2012，20(1)：199-202.

[5] 李静梅，孙丽华.一种文本处理中的朴素贝叶斯分类器[J].哈尔滨工程大学学报，2003，24(1)：71-74.

[6] 赵文涛，孟令军.朴素贝叶斯算法的改进与应用[J].测控技术，2016，35(2)：143-147.

[7] Salton G, Yu C T. On the construction of effective vocabularies for information retrieval[J]. Acm Sigplan Notices, 1973, 10(1):48-60.

[8] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data.[C]// Acm Sigkdd Explorations Newsletter,1931:80-89.

[9] 张保富，施化吉，马素琴.基于 TFIDF 文本特征加权方法的改进研究[J].计算机应用软件，2011，28(2)：17-20.

[10] Salton G, Introduction to Modern Information Retriveval [M].New York:McGraw Hill Book Company, 1983.

[11] 鲁松，李晓黎，白硕.文档中词语权重计算方法的改进[J].中文信息学报，2000，14(6)：8-20.

[12] 施聪莺，徐朝军，杨晓江.TFIDF 算法研究综述[J].计算机应用，2009(29)：167-180.

An Optimized Bayesian Classification Algorithm for WeChat Information

ZHANG Yingjiang, KU Kailin

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430000, *China*)

Abstract: The prevalence and development of WeChat has facilitated the dissemination of information. However, it also comes with advertising and fraudulent information, which brings a lot of troubles to people's lives. For the two important factors that naive bayes considering all the features and characteristics are given the same weight. We put forward an improved bayes algorithm. Selecting features by improved mutual information and weighting by improved TFIDF to optimize the bayesian classification performance. Experimental results show that the improved bayes algorithm can effectively select key features and improved classification precision

Keywords: bayes; WeChat information; feature extraction; feature weighting; information classification

[责任编校：张岩芳]