第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)09—0200—03 文献标识码：A 中图分类号：TP311

# 一种基于特征投票的文本分类方法

焦庆争[1,2]，蔚承建[1]
(1. 南京工业大学信息科学与工程学院，南京 210009；2. 安徽师范大学信息管理中心，芜湖 241000)

**摘 要：**基于特征投票机制设计一种线性文本分类方法，运用信任机制理论分析文档类别对特征的信任关系，给出具体特征信任度的模型，并在 Newsgroup、复旦中文分类语料、Reuters-21578 3 个广泛使用且具有不同特性的语料集上与传统方法进行比较。实验结果表明，该方法分类性能优于传统方法且稳定、高效，适用于大规模文本分类任务。

**关键词：**文本分类；特征投票；经验概率；自然语言处理

# Text Categorization Method Based on Features Vote

JIAO Qing-zheng[1,2], WEI Cheng-jian[1]
(1. College of Information Science and Engineering, Nanjing University of Technology, Nanjing 210009;
2. Information Management Center, Anhui Normal University, Wuhu 241000)

**【Abstract】**This paper discusses a text categorization approach based on features vote, which is linear as well as high efficient. It uses the trust mechanism theory to analyze the trust relation between features and document classes, and gives the model to calculate the trust values. In the comparison experiments, Newsgroup, Fudan Chinese evaluation data collection and Reuters-21578 are used to evaluate the effectiveness of the techniques. Experimental results show the method can improve the performance for text categorization, and is suitable for large-scale text categorization.

**【Key words】**text categorization; features vote; empirical probability; natural language processing

## 1 概述

自动文本分类是一种有监督的学习任务[1]，即根据已分类的训练文档集合，对未分类文档分配类标签。近年来，越来越多的统计理论和机器学习方法用于文本自动分类，文献[1-2]对主要分类方法做了详细论述。虽然文本分类方法很多，但传统分类方法并没有在分类性能和分类效率 2 个层面上取得理想的结果。本文采用一种新的基于特征信任投票的文本分类方法，将文本分类理解为测试文档中的特征对文档类别投票的结果，综合特征对测试文档的投票数(词频)及文档类别对特征的信任值，实现线性的文本分类。

## 2 特征投票机制分析

本文基于特征投票机制文本分类提出如下假设：文本分类中特征既是文本分类的参与者又是文档类别判定评审专家，分类是根据专家的信用度及在测试文档中的投票数决定测试文档类别。基于此假设，在训练语料时，将参与分类的词汇视为专家特征，根据训练语料考察文档类别对特征的信任值。

在测试文档时，分别将各文档类别对特征的信任值与投票数的乘积作为特征对文档类别判定的贡献率，综合各特征的投票即可得到各类别的分类值。分类值排序则是测试文档类别排序，根据单标签、多标签不同分类方式选取不同数量的文档类别。特征投票机制文本分类的关键是计算文档类别对特征的信任值。信任值与类特征频率、文档类别分布、特征平均频率等向量相关。

从直觉考虑，文档类别对特征信任值相对类特征概率( $p(c_i)=TF_i/\sum TF_k$ )单调增加，即 $p(c_i)$越大，文档类别对特征越信任，在判别测试文档为 $i$ 类时，特征对 $i$ 类的贡献也越大；同时，特征信任值与特征平均频率 $\overline{TF}=1/C\times\sum TF_k$ 呈线性增加的关系，例如，在其他条件同等情况下，假如 $\overline{TF}_A$ 为 10，$\overline{TF}_B$ 为 5，那么对特征 $A$ 比对特征 $B$ 更信任。当然，这种简单的直觉观存在很大风险，类特征概率是一个归一结果，它忽略了特征在各类别之间训练的分布对特征信任值的影响，因此，必须对类特征概率的风险进行评估。从类特征概率公式可以发现，$p(c_i)$对特征信任值的调节过于平缓，尤其在类别数较多时，分母基数较大，分子间的数量差距不能充分表达，致使测试文档的分类过于依赖特征投票数。另外，当训练文档集合极不均匀时，类特征概率没有同等的比较条件。因此，本文重新构建特征概率模型。

## 3 基于特征投票的分类模型

为了更好地表达特征在各类别中的分布特征，本文将特征概率分为微观经验概率和宏观经验概率，分别定义如下：

微观经验概率($mip_{ij}$)：将训练集合中非 $i$ 类文档数平衡到与 $i$ 类文档数相同时特征 $j$ 对 $i$ 类的概率，计算如下：

$$mip_{ij}=\frac{TF_{ij}}{TF_{ij}+\dfrac{TF_i-TF_{ij}}{N-N_j}N_j} \tag{1}$$

**基金项目：**国家自然科学基金资助项目(60703071)；安徽省高校省级自然科学研究基金资助重点项目(KJ2009A63)
**作者简介：**焦庆争(1974－)，男，讲师、硕士研究生，主研方向：自然语言处理；蔚成建，教授、博士
**收稿日期：**2009-10-13 **E-mail：**qzjiao@mail.ahnu.edu.cn

当训练文档集合为类别均匀时，计算如下：

$$mip_{ij}=\frac{TF_{ij}}{TF_{ij}+\dfrac{TF_i-TF_{ij}}{k-1}} \tag{2}$$

宏观经验概率($map_{ij}$)：在训练集合原始 $j$ 类与非 $j$ 类文档数分布下，特征 $j$ 对 $i$ 类的概率计算如下：

$$map_{ij}=\frac{TF_{ij}}{TF_j} \tag{3}$$

将微观经验概率作为类特征概率，宏观经验概率为微观经验概率的可靠性评估提供标准基线参考。为了对微观经验概率的可靠性进行评估，本文考察了相关信任模型，重点研究了 Josang 模型[3-5]。Josang 模型中提出了基于主观逻辑(subjective logic)的信任模型，并引入了事实空间(evidence space)和观念空间(opinion space)概念来描述和度量信任关系，以二项事件后验概率的 Beta 分布函数为基础，给出了概率确定性密度函数 *pcdf*，并以此为基础计算实体之间产生的每个事件的概率的可信度。

设概率变量为 $\theta$，$r$ 和 $s$ 分别表示肯定事件和否定事件数，则 *pcdf* 公式表述为

$$f(\theta\mid r,s)=\frac{\Gamma(r+s+2)}{\Gamma(r+1)\Gamma(s+1)}\theta^r(1-\theta)^s \tag{4}$$

本文并不使用 Josang 模型直接计算事件的可信度，而仅利用其概率确定性密度的概念。将特征在文档类中的分布映射到 Josang 二项事件模型可表述为：特征在 $C_i$ 类的频率为肯定事件数，在非 $C_i$ 类的频率为否定事件数。据此将特征微观、宏观经验概率对应参数代入式(4)，得到对应的概率确定性密度。计算如下：

$$mipcdf_{ij}=f(mip_{ij}\mid TF_{ij},(TF_j-TF_{ij})\times N_i/(N-N_i))$$

$$mapcdf_{ij}=f(map_{ji}\mid TF_{ij},TF_i-TF_{ij})$$

概率确定性密度并不是一个标准化的结果，直接参加计算会导致对其过依赖而忽略其他变量的作用，因此，以宏观经验概率的确定性密度为参照，将微观与宏观经验概率的确定性密度相对比较值作为特征可靠度($TR$)，其计算公式如下：

$$TR_{ij}=mipcdf_{ij}/mapcdf_{ij}$$

与此同时，还必须评估文档数不平衡带来的风险。与特征可靠度计算类似，可用下式计算文档数可靠度：

$$Dmipcdf_i=f(0.5\mid N_i,N_i)$$

$$Dmapcdf_i=f(N_i/(N-N_i)\mid N_i,N-N_i)$$

$$DR_i=Dmipcdf_i/Dmapcdf_i$$

综合微观经验概率、特征可靠度、文档数可靠度及特征平均频率，最终得到文档类对特征的信任值：

$$R_{ij}=mip_{ij}\times TR_{ij}\times DR_i\times\overline{TF}$$

在均匀训练文档集合中可直接令特征可靠度为微观经验概率的可靠度，因为此时所有特征和文档数的可靠度皆相同，所以可以省去文档数可靠度的计算过程，提高算法效率。

测试文档线性分类器综合所有特征的投票并以文档类别对特征的信任值为权重累加，得到测试文档分别在不同文档类别上获得的投票结果，对结果排序，选择较大者作为测试文档的类别，见式(5)。

对于单标签分类，投票值最大的类别即为测试类别，对于多标签分类，则需确定阈值，根据阈值选择测试文档类别，阈值是通过训练获得的。

$$DC_i=\sum_j p_{ij}R_{ij}\overline{TF_j}$$

$$DC=\arg\max_{1\leqslant i\leqslant k}DC_i \tag{5}$$

其中，$j$ 为测试文档中特征出现的顺序。

## 4 实验分析

### 4.1 实验方法

为了较全面地测试算法对不同分类语料集的适应性，本文选择 3 个具有典型特征的语料集 Reuters-20Newsgroup (Newsgroup)、复旦大学计算机信息与技术系国际数据库中心提供的中文分类语料集(Fudan)和 Reuters-21578。这 3 个分类语料被国内外学者广泛用于评测文本分类性能，便于评测数据的比较，同时这 3 个语料又各具特色，Newsgroup 为单标签英语平衡语料，Fudan 为单标签中文不平衡语料，Reuters-21578 为多标签不平衡语料。本文使用了各语料提供的所有文档类别及所有文档。关于语料的详细特性可参阅各语料的官方网站。

在 Newsgroup, Fudan 语料集上，采用 LtcTFIDF 加权的朴素贝叶斯方法(性能优于朴素贝叶斯方法)与 KNN 方法进行比照实验。在 Reuters-21578 语料上，本文将它与文献[1]的实验结果进行比较。

本文分类方法是为了构建无需特征选择的分类算法，但比照的 2 种分类方法需要利用特征选择来提高分类性能[6]，因此，本文使用信息增益(Information Gain, IG)统计量进行特征选择，比较算法在不同特征规模下的性能表现。

### 4.2 实验结果比较

(1)Newsgroup 语料实验

3 种算法的微平均 $F1$ 值和宏平均 $F1$ 值都十分接近，为了使图表清晰直观，只用微平均 $F1$ 值表示。如图 1 所示，基于特征投票模型(FVM)与 LtcNB 性能明显高于 KNN，在各自取得最佳性能时，FVM 比 LtcNB 高 1.1%，比 KNN 高 8.6%。

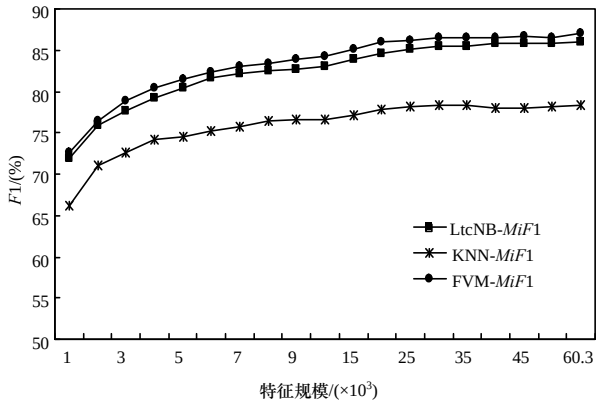


**图 1 NewsGroup 分类语料分类结果**

(2)复旦中文分类语料实验

在此语料上，3 种算法在各自微 $F1$ 取得最佳性能时，FVM 的微 $F1$ 指标与宏 $F1$ 指标分别比 LtcNB 高 9.6%和 6.2%，比 KNN 高 2.4%和 12.4%。特征选择对 LtcNB 的性能影响很大，在选取 1 000 个特征时性能最好。随着特征规模的增加，性能急剧下降，而对于 FVM 与 KNN，特征规模增大总体有

利于分类性能的提高。结果见图 2。

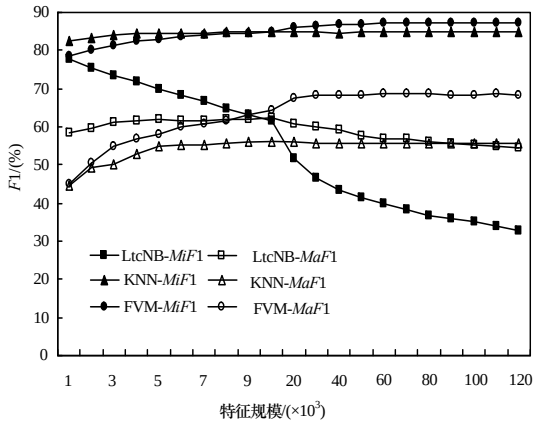


**图 2　复旦中文分类语料分类结果**

(3)Reuters-21578 语料实验

在 Reuters-21578 语料上，文献[1]对主流方法做了文本分类实验，并有 SVM, KNN, LSF, NNET, NB 等实验的具体性能指标值，如表 1 所示。FVM 性能指标为微平均 *F*1 指标达到最佳时的性能指标，特征规模为 2 900。FVM 除微平均精度略低于 SVM 外，各项指标均优于其他方法。

**表 1　Reuters-21578 文本分类结果**

| Method | *Mir* | *Mip* | *MiF*1 | *MaF*1 |
|---|---|---|---|---|
| FVM | 0.886 0 | 0.909 5 | 0.897 6 | 0.578 7 |
| SVM | 0.812 0 | 0.913 7 | 0.859 9 | 0.525 1 |
| KNN | 0.833 9 | 0.883 7 | 0.856 7 | 0.524 2 |
| LSF | 0.850 7 | 0.848 9 | 0.849 8 | 0.500 8 |
| NNET | 0.784 2 | 0.878 5 | 0.828 7 | 0.376 5 |
| NB | 0.768 8 | 0.824 5 | 0.795 6 | 0.388 6 |

综合以上 3 个语料的实验结果，FVM 算法适合不同语种、单或多标签、不平衡语料分类，具有高稳定性，性能优于传统分类方法。尤其在不均匀语料集上性能提高显著，克服了 NB, KNN 等方法在不同语料下性能不稳定的缺点，同时算法为线性分类方法，适合为大规模文本分类，有很强的实用性。

## 5　结束语

与以往基于统计理论的文本分类方法大多从机器学习理论寻求解决途径不同，本文基于文档类别与特征的信任分析，提出了一种基于特征投票的文本分类方法，算法简洁、高效、易于理解，对机器学习吸收其他研究领域理论做了有益尝试。下一步将深入研究可信计算与机器学习之间的关系，为自然语言处理领域的问题寻求更好的解决途径。

## 参考文献

[1] Yang Yiming, Liu Xin. A Re-examination of Text Categorization Methods[C]//Proceedings of ACM SIGIR'99. Berkeley, CA, USA: ACM Press, 1999: 42-49.

[2] Yang Yiming. An Evaluation of Statistical Approaches to Text Categorization[J]. Information Retrieval, 1999, 1(1/2): 69-90.

[3] Josang A. A Model for Trust in Security Systems[C]//Proceedings of the 2nd Nordic Workshop on Secure Computer Systems. Philadelphia, USA: ACM Press, 1997.

[4] Jøsang A, Knapskog S J. A Metric for Trusted Systems[C]// Proceedings of the 21st National Security Conference. Gaithersburg, MD, USA: NIST Press, 1998: 16-29.

[5] Jøsang A, Ismail R. The Beta Reputation System[C]//Proceedings of the 15th Bled Conference on Electronic Commerce. Bled, Slovenia: [s. n.], 2002: 17-19.

[6] Bressan M, Vitria J. On the Selection and Classification of Independent Features[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1312-1317.

编辑　张正兴

（上接第 199 页）

## 参考文献

[1] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[2] 鲁　松，李晓黎，白　硕，等. 文档中词语权重计算方法的改进[J]. 中文信息学报, 2001, 14(6): 8-13.

[3] 唐焕玲，孙建涛，陆玉昌. 文本分类中结合评估函数的 TEF-WA 权值调整技术[J]. 计算机研究与发展, 2005, 42(1): 47-53.

[4] Shankar S, Karypis G. A Feature Weight Adjustment Algorithm for Document Categorization[C]//Proc. of KDD'00. New York, USA: ACM Press, 2000.

[5] 陆玉昌，鲁明羽，李　凡，等. 向量空间法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.

[6] Forman G. BNS Feature Scaling: An Improved Representation over TF-IDF for SVM Text Classification[C]//Proc. of the 12th ACM Conference on Information and Knowledge Management. Napa Valley, CA, USA: ACM Press, 2008: 26-30.

[7] Zhang Yuntao, Gong Ling, Wang Yongcheng. An Improved TF-IDF Approach for Text Classification[J]. Journal of Zhejiang University, 2005, 6A(1): 49-55.

[8] Rocchio J. The SMART Retrieval System: Experiments in Automatic Document Processing[M]. Englewood Cliffs, USA: Prentice-Hall, 1971.

[9] Salton G, Buckley C. Term Weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management, 1988, 24(5): 513-523.

[10] Salton G. Developments in Automatic Text Retrieval[J]. Science, 1991, 253(5023): 974-979.

编辑　张正兴