ISSN 1000-0054 清华大学学报（自然科学版）2003 年 第 43 卷 第 4 期 21/35
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2003, Vol. 43, No. 4 513-515,520

基于权值调整的文本分类改进方法

鲁明羽[1,2]， 李 凡[1]， 庞淑英[1,3]， 陆玉昌[1]， 周立柱[1]

（1. 清华大学 计算机科学与技术系，北京 100084；2. 烟台大学 计算机学院，烟台 264005；3. 昆明科技大学 计算中心，昆明 650093）

摘 要：文本分类是文本挖掘的基础与核心，可广泛应用于传统的情报检索和 Web 信息的检索与挖掘等。提出了一种利用权值调整思想对向量空间法(VSM)和朴素 Bayes 分类器(NBC)进行改进的文本分类方法，并探讨了利用 EM 算法进行无导师 Bayes 分类的方法，设计和实现了一个中英文文本分类系统 CZW。3 组实验数据表明，用某些评估函数调节单词权值可有效提高 VSM 和 NBC 等文本分类模型的精度，并且训练文本规模越大，改进的效果越明显。NBC 的分类精度最高可达 86%。

关键词：文本分类；权值调整；VSM；Bayes 分类器

中图分类号：TP 301 **文献标识码**：A

文章编号：1000-0054(2003)04-0513-03

Improved text classification methods based on weighted adjustments

LU Mingyu[1,2], LI Fan[1], PANG Shuying[1,3],
LU Yuchang[1], ZHOU Lizhu[1]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Computer Institute, Yantai University, Yantai 264005, China;
3. Computer Center, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Text classification is the key to text mining which is used extensively in traditional information searches, web information queries and web mining. A text classification method was developed using a weighted adjustment measure to improve the vector space model (VSM) and the naive Bayesian classifier (NBC). The EM algorithm was then used for non-tutor Bayesian learning and a Chinese/English text classification system was developed. Three sets of test results show that the weighted adjustment measure using scoring functions can improve the precision of text classification models such as VSM and NBC with the effect increasing with increasing size of the training text set. The maximum NBC precision is 86%.

Key words: text classification; weight adjustment; VSM; Bayesian classifier

目前已有的多种文本分类方法中，较为常用的是向量空间法(VSM)和朴素 Bayes 分类器(NBC)。前者简单直观，处理速度快，但其分类精度较低[1]；后者具有较严密的理论基础[2]，但如何利用大量未标注文本进行学习值得进一步研究。

本文提出一种利用特征筛选中的评估函数计算单词权值、代替 IDF 函数进行权值调整的方法，可有效提高向量空间法和朴素 Bayes 分类器的性能；比较分析了采用各种不同评估函数进行权值调整的优劣，并探讨了在缺乏大量训练文本的情况下，利用最大期望(EM)算法进行无导师 Bayes 分类的方法。所研制的中英文文本分类系统 CZW 集成了向量空间法、K 近邻分类、朴素 Bayes 分类以及 K 依赖 Bayes 网络分类等多种方法，并且体现了对向量空间法和朴素 Bayes 分类器的改进。

1 VSM 的分析和改进

向量空间法[1]的基本思想是用词袋法表示文本，将每个词条作为特征空间坐标系的一维，将文本看作是特征空间中的一个向量，用两个向量之间的夹角来衡量两个文本之间的相似度。

它采用下面的 TF-IDF 函数来计算单词 W 的权重 weight (W)：

$$\text{weight}(W) = \text{TF} \times \text{IDF} = \text{TF} \times \lg[|D|/\text{DF}(W)],$$

其中：词频 TF 为单词 W 在该文本中出现的次数，$|D|$ 代表训练集中文本总数，$\text{DF}(W)$ 为出现了 W 的文档数，而 $\text{IDF}=\lg[|D|/\text{DF}(W)]$ 称为逆文本频数。该算法又称为 TF-IDF 法。

TF-IDF 函数中的 IDF 函数本质上是一种试图抑制噪音的加权。然而，如果简单地认为文本频数少的单词就重要，显然过于武断。因为一个文本中对分

收稿日期：2002-01-21
基金项目：国家“九七三”重点基础研究项目 (G1998030414)
作者简介：鲁明羽(1963-)，男(汉)，吉林，副教授。
E-mail: mylu99@mails.tsinghua.edu.cn
通讯联系人：陆玉昌，教授，E-mail: lyc@tsinghua.edu.cn

类有用的单词只占一小部分,而大部分单词与要判别的类无关,属于"噪音单词"。两个文本之间的夹角在很大程度上是由这些噪音单词的词频差异而非有用单词的词频差异决定。这些噪音完全可能淹没有用信息,从而导致以 TF-IDF 为坐标系测度的分类方法精度极低。

文本特征选择中的各种评估函数是从信息论中延伸出来的,用于给单个单词打分,能够很好地反映单词与各类别之间的相关程度。常用的评估函数包括信息增益、信息熵、互信息、文本证据权等等,如果以它们代替 IDF 函数,对单词进行权值调整,就有希望得到高质量的向量空间法[3]。

例如,几率比是一个用于二元分类的优秀测度,能够成功地找出与目标类最相关的特征集合。CZW 系统中对其进行了改造,使其可以运用于多元分类问题中。具体方法是:先对每个类用几率比求出所有单词分值,即对应每个类求出一个类向量,向量中的元素是单词在此类中的重要性,用几率比衡量;然后每个类向量都进行归一化;最后把所有类向量加起来,得到一个新向量,此向量中的每个分量的值就是此分量对应单词的权重。这种新的评估函数可称为几率比改进型评估函数。

基于上述思想,对各种评估函数调节单词权值的效果进行了测试。表 1 列出了部分实验数据。从表 1 可以看到,用某些评估函数调节单词权值,能使向量空间法的分类精度得到明显提高。

表 1 VSM 实验数据(2 910 个训练文本)

方法		分类精度/%(向量空间法质心分类)	
		基于单词频数	基于文档频数
无特征选择		76	78
信息增益	特征选择	76	77
	权值调整	71	69
期望交叉熵改进型	特征选择	76	78
	权值调整	76	76
互信息	特征选择	76	77
	权值调整	80	80
文本证据权改进型	特征选择	77	75
	权值调整	79	82
几率比改进型	特征选择	76	77
	权值调整	83	79
CHI	特征选择	75	76
	权值调整	65	76

注:数据来源:易宝中文新闻;测试文本集合个数:1 000;特征选择后保留单词比例:30%;语种:中文;类别:国际、经济、体育、文教、政治。

2 NBC 中的单词权重调整

由于在向量空间法中用评估函数对单词加权取得了成功,因此可以考虑能否把加权的思想延伸到其他分类模型中。CZW 采用权重调整方法对朴素 Bayes 分类器进行改进,在实验中得到了较理想的结果。

朴素 Bayes 模型[4,5]中计算文档 d 属于类 C_j 概率的公式为

$$P(C_j|d) \propto P(C_j)\prod_{k=1}^{|d_i|}P(W_{d_i,k}|C_j),$$

其中:k 表示一篇文档中单词的位置;$W_{d_i,k}$ 表示文档 d_i 中的第 k 个单词。

用评估函数加权后,上述公式变为

$$P(C_j|d) \propto P(C_j)\prod_{k=1}^{|d_i|}P(W_{d_i,k}|C_j)^{f(W_{d_i,k})},$$

式中 $f(W_{d_i,k})$ 为单词 $W_{d_i,k}$ 的评估函数。根据观察,$f(W_{d_i,k})$ 采取改进后的文本证据权是一种稳定而高效的策略。$f(W_{d_i,k})$ 越小,单词 $W_{d_i,k}$ 在朴素 Bayes 模型中起的作用就越小,当 $f(W_{d_i,k})$ 为 0 时,$P(W_{d_i,k}|C_j)$ 实际上就不起作用。文[3]中详细地讨论比较了 $f(W_{d_i,k})$ 的各种计算公式。

表 2 和表 3 分别列出了训练文本个数为 2 910 和 10 000 时,用评估函数加权的方法改进朴素 Bayes 分类器的实验结果。一个值得注意的现象是,当训练集合个数较少时,朴素 Bayes 效果较差,但训练数据越多,朴素 Bayes 的优越性就越能得到体现。当训练文本的个数接近 3 000 时,朴素 Bayes 已明显优于向量空间法。

表 2 NBC 实验数据(2 910 个训练文本)

方法		分类精度/%	
		基于单词频数	基于文档频数
无特征选择		83	83
信息增益	特征选择	77	74
	权值调整	83	81
期望交叉熵改进型	特征选择	78	75
	权值调整	84	83
互信息	特征选择	37	38
	权值调整	84	83
文本证据权改进型	特征选择	77	76
	权值调整	83	82
几率比改进型	特征选择	72	73
	权值调整	84	81
CHI	特征选择	77	73
	权值调整	58	82

注:数据来源:易宝中文新闻;测试文本集合个数:1 000;特征选择后保留单词比例:30%;语种:中文;类别:国际、经济、体育、文教、政治。

表 3 NBC 实验数据(1 万个训练文本)

方法		分类精度/%	
		基于单词频数	基于文档频数
无特征选择		84	84
信息增益	特征选择	78	79
	权值调整	84	78
期望交叉熵改进型	特征选择	76	78
	权值调整	86	69
互信息	特征选择	37	36
	权值调整	85	85
文本证据权改进型	特征选择	76	78
	权值调整	86	86
几率比改进型	特征选择	78	78
	权值调整	84	82
CHI	特征选择	78	78
	权值调整	76	80

注：数据来源：易宝中文新闻；测试文本集合个数：1 045；特征选择后保留单词比例：30%；语种：中文；类别：国际、经济、体育、文教、政治。

另一方面，训练数据越多，用评估函数加权的效果也越明显。可以清楚地看到，当训练文本的个数达到 10 000 时，评估函数能使分类精度有相当大的提高。

3 利用 EM 算法进行 Bayes 学习

前面提到的各种文本分类算法多数都属于有导师学习，需要大量的训练集合才能得到正确的文本分类器。但是在实际运用中面临的一个问题是，未必能找到大量已经正确标注类别的训练集合。然而，未标注的文本集合却是极其丰富的，一个简短的脚本语言程序就可从网络上下载巨量的无标注文本。因此，利用少量有类标的文本集合和大量无类标的文本集合作为训练集，进行无导师学习，具有重要的研究意义。

EM 算法[6]是一种经典的统计算法。当某一数据模型丢失了某些数据时，EM 算法利用当前模型的不完整数据通过反复计算，对缺失数据获得最大的后验概率估计，从而提高模型性能。如果在分类时缺少足够的训练文本，那么采用 EM 算法是一种可行的解决方案。

算法的基本思路是首先利用由少量训练文本组成的原始训练集合，将它们输入前述的标准 Bayes 网络中，然后用这些原始训练集初始化 Bayes 分类器的参数，再利用 EM 算法改造 Bayes 分类器，进行无导师学习，对大量的无标注文本进行处理，从而进一步优化 Bayes 分类器的参数，提高其分类性能。

EM 算法的基本流程是反复执行 E 步骤和 M 步骤。首先，利用初始有类标数据文本，像处理标准朴素 Bayes 一样设置参数估计。E 步骤对每个文件用公式

$$P(C_j|d_i) \approx P(C_j)P(d_i|C_j) \approx P(C_j)\prod_{W\in V}P(W|C_j)^{\mathrm{TF}(W,d_i)} \approx P(C_j)\prod_{k=1}^{|d_i|}P(W_{d_i,k}|C_j),$$

求取文档 d_i 属于某一类的概率值 $P(C_j|d_i)$。

而 M 步骤利用 E 步骤的结果，根据

$$P(W_t|C_j)=\frac{1+\mathrm{TF}(W_t,C_j)}{|V|+\sum_s \mathrm{TF}(W_s,C_j)}=\frac{1+\sum_{d_i\in D}N(W_t,d_i)P(C_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{d_i\in D}N(W_s,d_i)P(C_j|d_i)},$$

$$P(C_j)=\frac{1+\sum_{d_i\in D}P(C_j|d_i)}{|C|+|D|},$$

求取新的分类器参数。式中 W_t 为本次 M 步骤处理的单词，W_s 为单词集合中的任一单词，$|V|$ 表示单词集合中元素的数目，$|C|$ 表示类别集合中元素的数目。

反复重复 E 步骤和 M 步骤，直到结果收敛。由于初始化时原始训练集合对每个类都撒下了种子，因此 EM 算法找到的局部最大值就是希望得到的结果。EM 算法将给整个数据集合标上正确和完整的类标。

整个算法的流程如下。

输入 少量有类标的训练文本集合 T，大量无类标文本集合 d_{nt}，类别序列 C。

处理流程

1) 对于训练文本集合 T 中每个文本，已知其文本类标期望值 $P(C|d_{\mathrm{nt}})$为 0 或 1。

2) 反复执行下面的 E 步骤与 M 步骤，直到收敛。

M 步骤：对已知文本类标期望值 $P(C|D)$的文本中的每个单词 W，求其对于每个类的最大概率估计 $P(W|C)$，得到分类器参数设置。

E 步骤：用 M 步骤求出的分类器参数设置 $P(W|C)$，对每个文本计算新的文本类标期望值 $P(C|d_{\mathrm{nt}})$。

输出 一个训练好的 Bayes 分类器。

EM 算法的主要缺陷是当处理大量模型参数时，其计算量太大，收敛很慢，而文本挖掘恰恰需要大量的词条作为模型参数。如何减少计算量，提高 EM 算法的效率，是一个值得深入研究的课题。

(下转第 520 页)

参考文献 (References)

[1] JIN Qin, SI Luo, HU Qixiu. A high-performance text-independent speaker identification system based on BCDM [A]. Proc of the Fifth Inter Conf on Spoken Language Processing [C]. Sydney, Australia. 1998.

[2] 牟晓隆, 胡起秀, 吴文虎. 与文本无关的复合策略说话人辨识系统 [J]. 清华大学学报, 1997, **37**(3): 16-19.
MOU Xiaolong, HU Qixiu, WU Wenhu. Text-independent speaker identification system based on multiple strategies [J]. *J Tsinghua Univ*, 1997, **37**(3): 16-19. (in Chinese)

[3] SI Luo, HU Qixiu. Two-stage speaker identification system based on VQ and NBDGMM [A]. Proc of the Sixth Inter Conf on Spoken Language Processing [C]. Beijing, 2000.

[4] 杨行峻, 迟惠生. 语音信号数字处理 [M]. 北京: 电子工业出版社, 1995.
YANG Xingjun, CHI Huisheng. Speech Signal Digital Processing [M]. Beijing: Publishing House of Electronic Industry, 1995. (in Chinese)

[5] 何致远. 说话人确认和辨认的研究与实现 [D]. 北京: 清华大学, 2002.
HE Zhiyuan. Research and Implementation of Speaker Verification and Identification [D]. Beijing: Tsinghua University, 2002. (in Chinese)

[6] 何致远, 胡起秀, 姚志宏. 基于HMM的数字串提示文本的说话人确认 [A]. 第九届全国多媒体技术学术会议论文集 [C]. 北京, 2000. 215-219.
HE Zhiyuan, HU Qixiu, YAO Zhihong. Digit string prompt speaker verification based on HMM [A]. Proc of the 9th National Conf of Multimedia Technology [C]. Beijing, 2000. 215-219. (in Chinese)

[7] Fakotakis N, Sirigos J. A high performance text independent speaker recognition system based on vowel spotting and neural nets [A]. Proc Inter Conf on Acoustics, Speech and Signal Processing [C]. Atlanta, USA. 1996. 661-664.

[8] Furui S. Recent advances in speaker recognition [J]. *Lecture Notes in Computer Science*, 1997, **1206**: 237-252.

[9] Li Qi, Juang Biinghwang, Lee Chinhui, et al. Recent advancements in automatic speaker authentication [J]. *IEEE Robotics and Automation Magazine*, 1999, **3**: 24-34.

[10] Furui S. Cepstral analysis technique for automatic speaker verification [J]. *IEEE Trans on Acoustics, Speech and Signal Processing*, 1981, **29**(2): 254-272.

(上接第 515 页)

4 结束语

文本分类原型系统 CZW 集成了质心分类、K 近邻分类、朴素 Bayes 分类、K 依赖 Bayes 网络分类等多种有代表性的文本分类模型,实现了 8 种特征评估函数。与许多商用文本分类系统相比较,CZW 在分类精度方面表现出了良好的性能,效果十分理想。

加入更多的文本分类方法,并尝试 Boosting 和 Bagging 等组合分类方法,可以进一步提高系统的分类精度。此外,权值调整方法还可以应用于网页文本的分类、聚类及文本摘要等问题。

参考文献 (References)

[1] Joachims T. A probabilistic analysis of the rocchio algorithm with TF-IDF for text classification [A]. Proc of 14th Inter Conf on Machine Learning ICML97 [C]. San Francisco, CA: Morgan Kaufmann Publishers, 1997. 143-151.

[2] YANG Yiming. An Evaluation of Statistical Approach to Text Classification [R]. Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University, 1997.

[3] 李凡, 鲁明羽, 陆玉昌. 文本特征选择新方法的研究 [J]. 清华大学学报, 2001, **41**(7): 98-101.
LI Fan, LU Mingyu, LU Yuchang. Research about new methods of feature extraction from text [J]. *J Tsinghua Univ*, 2001, **41**(7): 98-101. (in Chinese)

[4] Mladenic D, Grobelnik M. Feature selection for unbalanced class distribution and Naive Bayes [A]. Proc of 16th Inter Conf on Machine Learning ICML-99 [C]. San Francisco, CA: Morgan Kaufmann Publishers, 1999. 258-267.

[5] YANG Yiming, Pedersen J O. A Comparative Study on Feature Selection in Text Classification [EB/OL]. http://citeseer.nj.nec.com/yang97comparative.html, 1997.

[6] Sahami M. Using Machine Learning to Improve Information Access [D]. Computer Science Department, Stanford University, 1999.

基于权值调整的文本分类改进方法

作者： 鲁明羽， 李凡， 庞淑英， 陆玉昌， 周立柱

作者单位： 鲁明羽(清华大学,计算机科学与技术系,北京,100084;烟台大学,计算机学院,烟台,264005)， 李凡,陆玉昌,周立柱(清华大学,计算机科学与技术系,北京,100084)， 庞淑英(清华大学,计算机科学与技术系,北京,100084;昆明科技大学,计算中心,昆明,650093)

刊名： 清华大学学报(自然科学版) ISTIC EI PKU

英文刊名： JOURNAL OF TSINGHUA UNIVERSITY(SCIENCE AND TECHNOLOGY)

年，卷(期)： 2003,43(4)

被引用次数： 17次

参考文献(6条)

1. Joachims T A probabilistic analysis of the rocchio algorithm with TF-IDF for text classification 1997

2. Yang Yiming An Evaluation of Statistical Approach to Text Classification 1997

3. 李凡;鲁明羽;陆玉昌 关于文本特征抽取新方法的研究[期刊论文]-清华大学学报(自然科学版) 2001(07)

4. Mladenic D;Grobelnik M Feature selection for unbalanced class distribution and Naive Bayes[外文会议] 1999

5. Yang Yiming;Pedersen J O A Comparative Study on Feature Selection in Text Classification 1997

6. Sahami M Using Machine Learning to Improve Information Access 1999

本文读者也读过(4条)

1. 陆玉昌. 鲁明羽. 李凡. 周立柱 向量空间法中单词权重函数的分析和构造[期刊论文]-计算机研究与发展 2002,39(10)

2. 朱靖波. 陈文亮. ZHU Jing-bo. CHEN Wen-liang 基于领域知识的文本分类[期刊论文]-东北大学学报（自然科学版） 2005,26(8)

3. 袁方. 杨柳. 张红霞 基于k-近邻方法的渐进式中文文本分类技术[期刊论文]-华南理工大学学报（自然科学版） 2004,32(z1)

4. 张启蕊. 张凌. 董守斌. 谭景华. ZHANG Qirui. ZHANG Ling. DONG Shoubin. TAN Jinghua 训练集类别分布对文本分类的影响[期刊论文]-清华大学学报（自然科学版） 2005,45(9)

引证文献(17条)

1. 庄世芳 一种基于Ontology的中文Web文本聚类算法的研究[期刊论文]-福建电脑 2008(6)

2. 谭冠群. 丁华福 支持向量机方法在文本分类中的改进[期刊论文]-信息技术 2008(1)

3. 鲁明羽 Bayes文本分类器的改进方法研究[期刊论文]-计算机工程 2006(17)

4. 朱秀华 BP神经网络在网页自动分类中的应用[期刊论文]-现代情报 2009(5)

5. 吴迪. 张亚平. 殷福亮. 李明 基于类别分布差异和VPRS特征选择的文本分类方法[期刊论文]-电子与信息学报 2007(12)

6. 高鲁. 吴建明. 张雪胭. 罗成 评估指标权值调整的平衡处理及自动实现研究[期刊论文]-山西电子技术 2006(2)

7. 胡清华. 谢宗霞. 于达仁 基于粗糙集加权的文本分类方法研究[期刊论文]-情报学报 2005(1)

8. 许增福. 梁静国. 田晓宇 基于FVSM和自组织映射网络的Web文本自动分类方法[期刊论文]-哈尔滨工业大学学报 2004(9)

9. 饶丽丽. 刘雄辉. 张东站 基于特征相关的改进加权朴素贝叶斯分类算法[期刊论文]-厦门大学学报（自然科学版

) 2012(4)

10. 康进峰. 王国营. 梁春迎. 谭晓贞 用于色情网页过滤中的KNN算法改进[期刊论文]-计算机安全 2009(9)

11. 张彰. 樊孝忠 一种改进的基于VSM的文本分类算法[期刊论文]-计算机工程与设计 2006(21)

12. 鲁明羽. 张红. 付克明. 陆玉昌 WebME--一个大型网络挖掘环境系统[期刊论文]-哈尔滨工业大学学报 2004(9)

13. 侯凡. 周明全. 耿国华. 李杰 基于粗糙集的文本分类方法在网络科技资源应用集成环境中的应用[期刊论文]-计算机应用与软件 2009(3)

14. 吴志峰 基于概念特征的中文文本分类研究[学位论文]硕士 2005

15. 吴科 基于向量空间模型的中文文本分类的研究[学位论文]硕士 2004

16. 钟配蓉 基于Web挖掘的文本预处理研究及应用[学位论文]硕士 2006

17. 刘涛 用于文本分类和文本聚类的特征选择和特征抽取方法的研究[学位论文]博士 2004

引用本文格式：鲁明羽. 李凡. 庞淑英. 陆玉昌. 周立柱 基于权值调整的文本分类改进方法[期刊论文]-清华大学学报(自然科学版) 2003(4)