第 31 卷 第 17 期 计 算 机 工 程 2006 年 9 月
Vol.31 № 17 Computer Engineering September 2006
• 专题论文 • 文章编号：1000—3428(2006)17—0063—03 文献标识码：A 中图分类号：TP391

Bayes 文本分类器的改进方法研究

鲁明羽
(大连海事大学计算机科学与技术学院，大连 116026)

摘 要：在文本分类领域，Bayes 分类器是一种常用且效果较好的、基于概率的分类器，具有较严密的理论基础。该文对朴素 Bayes 文本分类器进行了分析，提出了利用权值调整机制改善其分类性能的方法，以及在缺乏大量训练文本的情况下，利用 EM 算法进行非监督 Bayes 分类的方法，并讨论了如何运用启发式方法确定 Bayes 网络结构，在更贴近真实环境的情况下进行文本分类。

关键词：文本分类；朴素 Bayes 分类模型；权值调整；EM 算法

Research on Improvement of Bayesian Text Classifier

LU Mingyu
(College of Computer Science and Technology, Dalian Maritime University, Dalian 116026)

【Abstract】 Bayesian classification model is common, powerful for text categorization task. It is based on probability and is of religious theoretic basis. The paper makes analysis to the simple and common naive Bayesian categorization model, and presents an approach to improve performance of Bayesian classification model using weight adjustment and an approach to make non-tutor Bayesian categorization using EM algorithm when lacking mass training texts, and discusses how to fix the framework of Bayesian network using heuristic methods so as to make text classification in real circumstance.

【Key words】 Text categorization; Naïve Bayesian categorization model; Weight adjustment; EM algorithm

1 概述

朴素Bayes(Naive Bayes)分类器假设各个单词之间条件独立，所有单词结点只有唯一的父结点，即类结点C[1]。其工作原理如下：

在训练集合中确定每个文本的类，然后据此计算训练文本中的单词(也可以选择词条)的概率估计。这样，分类器的参数就由先验类概率值和基于类的单词条件概率组成。严格地说，每个类C_j都有一个相对其它所有类的文件频度$P(C_j)$。对于单词表V中的每个词Wt, $P(Wt|C_j)$表示分类器预期单词Wt在类C_j的文件中发生的频度。在标准有导师学习的Bayes分类器中，分类器参数完全由已标注的训练集文件确定。下面给出$P(Wt|C_j)$和$P(C_j)$的计算公式。

为避免$P(Wt|C_j)$和$P(C_j)$为 0，需要对其做一些数学上的调整，如*Laplace*处理。$N(Wt,di)$表示单词Wt在文件di中出现的次数，$TF(W,C)$表示特征W在C类文本中出现的频度。对于训练集合中的文件di，定义当di属于类C_j时，$P(C_j|di)=1$，否则$P(C_j|di)=0$。这样，单词Wt在类C_j中的概率估计为

$$P(Wt|C_j)=\frac{1+TF(Wt,Cj)}{|V|+\sum_s TF(Ws,Cj)}=\frac{1+\sum_{di\in D}N(Wt,di)P(Cj|di)}{|V|+\sum_{s=1}^{|V|}\sum_{di\in D}N(Ws,di)P(Cj|di)} \quad (1)$$

类C_j的先验概率参数$P(C_j)$计算公式如下：

$$P(C_j)=\frac{1+\sum_{di\in D}P(Cj|di)}{|C|+|D|} \quad (2)$$

式中，$|C|$为类 C 中的元素数目。

对于测试集中的无标注文件，利用已训练好的分类器，可以求出文件属于类C_j的概率。用$W_{di,k}$表示文件di中的第k个单词，有

$$P(C_j|d)\propto P(Cj)P(di|Cj)\propto P(Cj)\prod_{W\in V}P(W|C)^{TF(W,di)}\propto P(Cj)\prod_{k=1}^{|di|}P(W_{di,k}|Cj) \quad (3)$$

对于 Naive Bayes 公式，值得特别提出的有两点：(1)LAPLACE 处理是必需的，否则分子为 0 会使分类效果大幅度降低；(2)Naive Bayes 分类器有布尔型和频度型两种，上面介绍的是频度型公式，其中$P(W|C)$的计算公式可以由贝叶斯公式加以数学处理推导出来。推导时用词频公式来计算$P(C)$、$P(C|W)$和$P(W)$。布尔型公式并不考虑一个单词在文本中出现的次数，而只考虑单词是否在文本中出现，其公式为

$$P(W|C)=\frac{1+N(doc(W)|C)}{2+|Dc|}$$

而

$$P(C|d)\propto P(C)\prod_{W\in V}P(W|C) \quad (4)$$

式中，$N(doc(W)|C)$为 C 类文本中出现特征 W 的文本数，$|Dc|$为 C 类文本所包含的文本数。这个公式也可以由贝叶斯公式加以数学处理推导出来，只不过推导时用文档公式来计算$P(C)$、$P(C|W)$和$P(W)$。

实验表明，一般情况下频度型公式比布尔型公式能得到更好的分类效果，究竟好多少要视具体数据集而定，但可以肯定的是，同一个词在一个文本中出现的次数越多，效果差别越大。

2 朴素贝叶斯模型中的单词权重调整

由于此前我们在向量空间法中用评估函数对单词加权取

基金项目：国家自然科学基金资助项目(60473115)
作者简介：鲁明羽(1963—)，男，博士、教授、博导，主研方向：数据挖掘，文本挖掘，网络挖掘，机器学习
收稿日期：2006-06-08 **E-mail：**lumingyu@tsinghua.org.cn

得了成功[2]，很自然地想到能否把加权的思想延伸到其它分类模型中。采用权重调整方法对朴素贝叶斯分类器进行改进，在实验中得到了较理想的结果。

朴素贝叶斯模型中计算文档d属于类C_j概率的公式为

$$P(Cj \mid d) \propto P(Cj)\prod_{k=1}^{|di|} P(W_{di,k} \mid Cj) \tag{5}$$

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

用评估函数加权后，上述公式变为

$$P(Cj \mid d) \propto P(Cj)\prod_{k=1}^{|di|} P(W_{di,k} \mid Cj)^{f(W_{di,k})} \tag{6}$$

式中，$f(W_{di,k})$为单词$W_{di,k}$的评估函数。文献[3]详细讨论比较了$f(W_{di,k})$的各种计算公式，本文中不再重复。根据观察，$f(W_{di,k})$采取改进后的文本证据权是一种稳定而高效的策略。$f(W_{di,k})$越小，单词$W_{di,k}$在朴素贝叶斯模型中起的作用就越小，当$f(W_{di,k})$为 0 时，$P(W_{di,k} \mid Cj)$实际上就不起作用。

表 1 和表 2 分别列出了训练文本个数为 3 000 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。2 个表中的实验文本均为中文，目标类别为：国际，经济，体育，文教和政治，特征选择后保留单词比例均为 30%。

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 的优越性就越体现出来，当训练文本的个数达到 3 000 时，朴素 Bayes 已明显优于向量空间法。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。

表 1 实验数据 1

分类精度(%)		朴素贝叶斯法	
		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

表 2 实验数据 2

分类精度(%)		朴素贝叶斯法	
		基于单词频数	基于文档频数
无特征选择		83.83	83.83
信息增益	特征选择	77.54	78.67
	权值调整	83.83	77.61
期望交叉熵改进型	特征选择	76.21	7761
	权值调整	85.55	68.61
互信息	特征选择	36.52	35.86
	权值调整	85.45	85.36
文本证据权改进型	特征选择	76.21	77.54
	权值调整	86.12	85.74
几率比改进型	特征选择	78.32	77.56
	权值调整	84.21	82.21
CHI	特征选择	78.32	77.54
	权值调整	76.23	79.81

3 利用 EM 算法进行无导师 Bayes 学习

前面列举的文本挖掘算法，除了文本聚类以外，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。但是在实际运用中的问题是未必能找到大量已经正确标注类别的训练集合。而另一方面，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 WWW 中下载巨量的无标注文本。因此利用少量有类标记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，具有极大的研究意义。

EM算法(最大期望算法)是一种经典的统计算法[4]。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。

我们提出的算法的基本思路是：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。一开始，利用初始有类标数据文本，像处理标准 Naive bayes 一样设置参数估计。E 步骤对每个文件用公式

$$\begin{aligned} P(Cj|d) &\propto P(Cj)P(di \mid Cj) \propto P(Cj)\prod_{W\in V} P(W \mid C)^{TF(W,di)} \\ &\propto P(Cj)\prod_{k=1}^{|di|} P(W_{di,k} \mid Cj) \end{aligned} \tag{7}$$

求取文档di属于某一类的概率值$p(C_j|di)$。

而 M 步骤利用 E 步骤的结果，根据

$$\begin{aligned} P(Wt \mid C_j) &= \frac{1+TF(Wt,Cj)}{|V|+\sum_s TF(Ws,Cj)} \\ &= \frac{1+\sum_{di\in D} N(Wt,di)P(Cj \mid di)}{|V|+\sum_{s=1}^{|V|}\sum_{di\in D} N(Ws,di)P(Cj \mid di)} \end{aligned} \tag{8}$$

$$P(C_j) = \frac{1+\sum_{di\in D} P(Cj \mid di)}{|C|+|D|} \tag{9}$$

求取新的分类器参数。

反复重复 EM 步骤，直到结果收敛。由于初试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我们希望得到的结果。EM 算法将给整个数据集合标上正确和完整的类标。

4 利用启发式方法确定 Bayes 网络结构

朴素Bayes分类器的特征独立性假设在实际应用中往往是不成立的。很多研究者都试图建造更强有力的Bayes分类器，让文本的特征结点除了类结点外，还可以拥有其他父结点，使其更好地将文本中存在的特征依赖关系模型化[1,5,6]。但已经从数学上得到证明，即使我们限制特征结点只能拥有除类结点外的两个父结点，问题也会是NP难的[7]。

一种解决方案是放弃常用的那种代价昂贵的搜索网络结构空间以获取局部最优网的方法，而采用一种启发式方法，以可行的计算量，构造次优的Bayes网络结构。这样有可能在特征数的多项式时间内为每个网络结点找到多个父结点。与树型Bayes网络TAN[8]相比，启发式构造Bayes网络的算法能表示更复杂的依赖关系，时间复杂度却与TAN相似，所以更受到重视。

可以定义 k 依赖 Bayes 分类器是一个允许每个特征结点至多有 k 个父结点、而类结点没有父结点的 Bayes 网络，即

$$P(X) = \prod_i P(Xi \mid \Pi(Xi)) = \prod_i P(Xi \mid \{C, Xdi\}) \tag{10}$$

式中，Xdi 是最多有 k 个结点的集合。

朴素 Bayes 分类器为零依赖 bayes 分类器，而无限制 Bayes 分类器为 n-1 依赖 Bayes 分类器（n 是特征数）。

本文试图寻求一种学习 Bayes 网络分类器的算法，放宽特征独立性假设，将朴素 Bayes 的每个单词结点彼此独立的假设，变为允许每个单词结点拥有 k 个其他单词结点作为父结点。这种算法能够允许有效地学习限制特征依赖结构的模型，用启发式方法为每个结点发现其 k 个父结点，更好地解决文本分类问题。问题的关键是寻找合适的启发式测度。

可以利用单词间的互信息来确定父子结点的关系。由于父子结点关系是单向的，因此可以先把单词结点按其与类结点的互信息分值按降序排序，只有小序号的结点可以为大序号结点的父结点，反之则不行，实际上也就是与类关系越大的结点越可能在 Bayes 网络中占据上层的位置。

Mehran Sahami 在文献[9]中给出了一个算法。以 Mehran Saham 的算法为基础，再利用前述的单词权值调整方法加以处理改进，实现了启发式 Bayes 网络文本分类器，效果比较理想。我们对文献[9]算法的另一个改进之处在于，通过实验发现，对于那些与类结点互信息值很低的单词结点，求取它们的近似父结点往往效果不大。这时把类结点作为它们的唯一父结点，分类精度往往会更高。

表 3 给出了训练集大小为 2 910 时启发式 Bayes 网络文本分类器的效果。与表 1 对比可看出分类精度有明显提高。

表 3 实验数据 3

分类精度(%)		启发式贝叶斯网络法	
		基于单词频数	基于文档频数
无特征选择		86	85
信息增益	特征选择	/	/
	权值调整	86	85
期望交叉熵改进型	特征选择	/	/
	权值调整	86	85
互信息	特征选择	/	/
	权值调整	87	86
文本证据权改进型	特征选择	/	/
	权值调整	88	86
几率比改进型	特征选择	/	/
	权值调整	88	87
CHI	特征选择	/	/
	权值调整	65	84

5 结束语

本文在对朴素 Bayes 分类器进行深入分析的基础上，提出了利用权值调整改善朴素 Bayes 分类器性能的方法，并利用 EM 算法，探索在缺乏大量训练文本的情况下进行无导师 Bayes 分类的有效方法，此外还研究了如何运用启发式方法确定 Bayes 网络结构，试图放宽 Bayes 分类器的特征独立性假设，在更贴近真实环境的情况下进行文本分类。上述工作均进行了实验验证，结果表明是十分有效的。

目前我们已经实现了一个文本分类系统SCETCS，其中集成了k-近邻、改进的朴素Bayes和SVM等多种分类方法，可以对纯文本和网页文本进行分类，并在数量超过 10 000 的文本集上进行了测试，取得了比较理想的效果[10]。今后我们将不断完善该系统，力图加入更多有效的文本分类方法，并尝试Boosting和Bagging等组合分类方法，进一步提高系统分类精度。

参考文献

1 石洪波, 王志海, 黄厚宽等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报, 2004, 14(2): 193-199.

2 鲁明羽, 李　凡, 庞淑英等. 基于权值调整的文本分类改进方法[J]. 清华大学学报, 2003, 43(4): 513-515.

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.

4 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 北京: 高等教育出版社, 1998.

5 Geiger D. An Entropy-based Learning Algorithm of Bayesian Conditional Trees[C]. Proc. of the 8th Annual Conference on Uncertainty in Artificial Intelligence, Stanford, California, 1992: 92-97.

6 Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning, 1997, 29(2/3): 131-163.

7 Chickering D M. Learning Bayesian Networks is NP-complete[M]. Springer Verlag, 1995.

8 石洪波, 王志海, 黄厚宽. 一种基于 TAN 的文本分类方法[J]. 广西师范大学学报, 2003, 21(1): 81-85.

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

10 Mingyu L, Keyun H, Yi W, et al. SCETCS: Towards Improving VSM and Naive Bayes Classifier[C]. Proc. of the 2nd IEEE International Conference on Systems, Man and Cybernetics, 2002: 465-469.

（上接第 62 页）

对稳定的水平上。今后，我们将会在过滤系统处理速度的提高以及特征词典的更新方面作进一步的研究工作。

参考文献

1 Spam Statistics[Z]. http://bloodgate.com/spams/stats.htm, 2004

2 Keizer G. Spam Costs World Businesses $50 Billion[EB/OL]. http://www.techweb.com/article, 2005.

3 第15次中国互联网络发展状况统计报告[R]. http://download.xinhuanet.com/it/document/cnnic15.doc.

4 Lan Mingjun, Zhou Wanlei. Spam Filtering Based on Preference Ranking[C]. Proceedings of the 5th International Conference on Computer and Information Technology, 2005.

5 Drucker H, Wu Donghui. Support Vector Machines for Spam Categorization[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048-1054.

6 Zhang M, Ma S P, Song R H. DF or IDF? On the Use of Primary Feature Model for Web Information Retrieval[J]. Journal of Software, 2005, 16(5): 1012-1020.

7 Moffat A, Davis R, Wilkinson R, et al. Harman D, Ed. Retrieval of Partial Documents[C]. Proc. of the 2nd Text Retrieval Conf. National Institute of Standards and Technology Special Publication, Gaither Sburg, 1994: 181-191.

8 Zhang Huaping, Liu Qun. Model of Chinese Words Rough Segmentation Based on N-shortest-paths Method[J]. Journal of Chinese Information Processing, 2002, 16(5): 1-7.

9 Sun Xia, Zheng Qinghu. Method of Special Domain Lexicon Construction Based on Raw Material[J]. Mini-Micro Systems, 2005, 26(6): 1088-1092.

10 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. Addison Wesley, 1999.