基于 NB 的双级分类模型在邮件过滤中的研究

惠 孛 吴 跃 陈 佳
(电子科技大学 成都 610054)

摘 要 使用朴素的贝叶斯(NB)分类模型对邮件进行分类,是目前基于内容的垃圾邮件过滤方法的研究热点。朴素的贝叶斯在参数之间联系不强的时候分类效果简单而有效。但是朴素的贝叶斯分类模型中对特征参数的条件独立假设无法表达参数之间在语义上的关系,影响分类性能。在朴素的贝叶斯分类模型的基础上,我们提出了一种双级贝叶斯分类模型(DLB,Double Level Bayes),既考虑到了参数之间的影响又保留了朴素的贝叶斯分类模型的优点。同时对DLB模型与朴素的贝叶斯分类模型的性能进行比较。仿真实验表明,DLB分类模型在垃圾邮件过滤应用中的效果在大部分条件下优于朴素的贝叶斯分类模型。

关键词 垃圾邮件过虑,朴素的贝叶斯分类模型,双级分类模型

The Research of NB-based DLB Classification Anti-spam

HUI Bei WU Yue CHEN Jia
(University of Electronic Science and Technology of China, Chengdu 610054)

Abstract Classification method using Naïve Bayesian(NB) classifier model which is the context-based spam filter method, is a hot point. The Naïve Bayesian classifier is a simple and effective classification method, but its attribute independence assumption makes it unable to express its semantic dependence. A new classification model is proposed which we call Double Lever Bayes classifier model (DLB). It considers not only the semantic dependence but also the simple and effective which is the excellence of NB classifier model. The performance is also compared between DLB and NB. The conclusion we get from experiment is that the performance using DLB classifier model is better than which using NB classifier model.

Keywords Spam filter, Naïve Bayesian classifier model, DLB model

1 引言

电子邮件是 Internet 上广泛使用的网络功能,但是,在它为人们带来便捷的同时,随之产生的是大量用户不希望收到的邮件。邮件的分类过滤成为人们重要的工作之一。朴素贝叶斯分类模型(Naïve Bayes Classifier Model)在邮件过滤应用的研究中一直扮演着重要的角色,在前人的研究[3,6]中显示,朴素贝叶斯分类模型的条件独立假设过于苛刻,尽管在一些特定条件下的应用中表现出了较好的效率,但是在现实使用中的大多数情况下明显不成立。在朴素贝叶斯分类模型中,特征参数之间的条件独立假设完全割裂了特征参数之间在内容表达上的关系,所以在要求考虑参数间关联的分类应用中显得有些不足。而且贝叶斯分类模型在应用链式规则计算后验概率的时候,随着参数数量的增加,计算规模成几何倍数增加,不适用于我们的应用。为了得到一个行之有效的方法,而同时又保留朴素的贝叶斯分类模型计算简单的优势,我们提出了改进的分类模型——双级贝叶斯(DLB)分类模型,将特征参数分为两级,选取联系较强的参数为第一级,联系较弱的参数为第二级,设定两级参数之间条件独立。这样,既考虑到了特征参数之间的影响,又在计算规模上接近于朴素的贝叶斯分类模型。

2 朴素贝叶斯分类模型在邮件过滤中的应用

贝叶斯分类模型是基于贝叶斯公式提出的一种使用统计方法对文档 d 进行分类的模型,对文档 d 分类就是计算文档 d 属于类集为 $C=\langle c_1,c_2,\cdots,c_m\rangle$ 中的类元素 c_i 的后验概率 $\Pr(c_i|d)$,并比较 $\Pr(c_i|d)$、$\Pr(c_j|d)$的大小($i\neq j$),确定最大的 $\Pr(c_j|d)$,将文档 d 的类标签归确定为 c_i。具有特征参数向量的文档 $d=\langle x_1,x_2,\cdots,x_n\rangle$ 属于类 c_i 的概率,表示为:

$$\Pr(c_i|x_1,x_2,\cdots,x_n)=\frac{\Pr(x_1,x_2,\cdots,x_n|c_i)\cdot\Pr(c_i)}{\sum_{i=1}^{m}\Pr(x_1,x_2,\cdots,x_n)\cdot\Pr(c_i)}$$
$$=\alpha\cdot\Pr(x_1,x_2,\cdots,x_n|c_i)\cdot\Pr(c_i) \quad (1)$$

其中 $\Pr(c_i)$为类 c_i 的先验概率,α 为正则因子。在一般情况下(1)式中 $\Pr(c_i)$是由专业领域的专家根据经验给出,如果不能事先取得 $\Pr(c_i)$的初始值时,通常令所有类的先验概率是相等的,即 $\Pr(c_i)=\frac{1}{m}$。先验概率 $\Pr(c_i)$在计算前可以根据经验事先确定,所以贝叶斯分类的关键是计算参数向量的联合分布。根据贝叶斯链式规则我们可以得到:

$$\Pr(x_1,x_2,\cdots,x_n|c_i)\Pr(c_i)=\Pr(c_i)\Pr(x_1|c_i)\Pr(x_2|x_1,c_j)\cdots Pr(x_n|x_1,x_2,\cdots,x_{n-1},c_i)=\Pr(c_i)\cdot\prod_{j=1}^{n}\Pr(x_j|$$

惠 孛 博士研究生,研究方向:网络计算,网络安全技术。吴 跃 博士生导师,教授,研究方向:网络计算,数据库技术。陈 佳 博士研究生,研究方向:智能计算,数据库技术。

$x_1, x_2, \cdots, x_{j-1}, c_i)$ (2)

贝叶斯分类模型的计算是一个计算规模巨大的运算，特别是在现实运用中随着参数数量的增加，计算规模成几何倍数增长。运用朴素的贝叶斯假设对其进行化简，假设所以的特征参数之间互相条件独立，即 x_i 对 $x_j(i\neq j)$ 的概率没有影响，(2)式简化为：

$$\Pr(x_1, x_2, \cdots, x_n | c_i)\Pr(c_i) = \Pr(c_i) \cdot \prod_{j=1}^{n} \Pr(x_j | c_i) \quad (3)$$

(3)式的计算与参数之间的关系无关，在参数数量增加时，计算规模成线性增长。

在垃圾邮件过滤的应用中，类集元素只有合法邮件类和垃圾邮件类两类 $C=\langle c_l, c_s\rangle$，其中 c_l 表示合法邮件类，c_s 表示垃圾邮件类，同时具有训练样本邮件集，$E=\langle e_1, e_2, \cdots, e_k\rangle$，分类模型通过学习后得到先验概率，计算后验概率的贝叶斯公式变形为：

$$\Pr(c_i | x_2, \cdots, x_n; E) = \frac{\Pr(x_1, x_2, \cdots, x_n | c_i; E) \cdot \Pr(c_i | E)}{\sum_{i=1}^{m} \Pr(x_1, x_2, \cdots, x_n | E) \cdot \Pr(c_i | E)} \quad (4)$$

其中的 E 是强调特征参数的概率是通过样本集 $E=\langle e_1, e_2, \cdots, e_k\rangle$ 的训练而得到的。合法(垃圾)邮件类的概率计算是通过统计合法(垃圾)邮件在样本集 E 中所占的比例而得到的，然后利用样本邮件集 $E=\langle e_1, e_2, \cdots, e_k\rangle$ 进行学习，计算出 $\Pr(c_i)$：

$$\Pr(c_i | E) = \frac{\sum_{j=1}^{|E|} \Pr(c_i | e_j)}{|E|} \quad (5)$$

而一封邮件 e 的特征参数概率 $\Pr(x_j | c_i; E)$ 就是其内容文本所包含的出现在 c_i 类中特征集中的文字出现的频率。使用训练样本，属于 c_i 类的邮件包含参数 x_j 的概率描述为 x_j 出现的次数除以 c_i 类样本所含的所有参数的数量：

$$\Pr(x_j | c_i; E) = \frac{N(x_j)}{\sum_{x \in c_j} N(x)} \quad (6)$$

这样依据公式(5)、(6)计算出 $\Pr(c_i | E)$ 和 $\Pr(x_j | c_i; E)$，计算分类的概率的(4)式可以变换为：

$$\beta = \frac{\Pr(c_s | E) \cdot \prod_{j=1}^{n} \Pr(x_j | c_s; E)}{\Pr(c_l | E) \cdot \prod_{j=1}^{n} \Pr(x_j | c_l; E)} \quad (7)$$

当 $\beta > \lambda$ 时，邮件 e_i 应该归为 c_s 类，作为垃圾邮件被阻塞。通常情况下 λ 取值为 1、9、999，分别表示错误地将一封正常邮件归为 c_s 类的代价是将一份垃圾邮件归为 c_l 类的 1 倍、9 倍、999 倍。

3 双级贝叶斯(DLB)分类模型

3.1 DLB 分类模型

虽然朴素贝叶斯分类模型具有较好的分类性能，但是它的条件独立假设太过局限，在参数之间没有太多关联性的应用上可以取得较好的效果，而在参数之间相互影响较大的时候效果就不太理想。朴素贝叶斯分类模型的简单假设使得一封邮件属于某个类的概率仅仅依赖于构成邮件的单词属于某个类的概率，而不考虑单词之间在语义、语言结构上的相互影响。我们对朴素贝叶斯分类模型进行改进，放松朴素贝叶斯在现实中过于苛刻的独立假设，提高分类器的性能与适应性，提出双级贝叶斯(DLB, Double Level Bayes)分类模型。首先根据特征参数间的相互关系计算 MI 和 I，选取 I 最大的 m 个特征值为第一级特征参数向量，其他的特征值为第二级，两级之间的参数使用朴素的贝叶斯假，这样，既考虑了语义的相关性，又保留了朴素的贝叶斯分类模型的简单高效的特点。

特征属性 x_i, x_j 之间的条件相互信息表示两个特征属性之间的条件依赖关系，MI 的值越大，说明该特征值对文本分类的影响越大：

$$MI(x_i, x_j | C) = \sum_{x, c_k} \Pr(x_i, x_j | c_k) \log \frac{\Pr(x_i, x_j | c_k)}{\Pr(x_i | c_k)\Pr(x_j | c_k)} \quad (8)$$

$$I(X_i) = \frac{1}{n-1} \sum_{j \neq i, j=1}^{n} MI(X_i, X_j | C) \quad (9)$$

对于属性集 $X=\langle X_1, X_2, \cdots, X_n\rangle$，分别计算各个 X_i 和 X_j 的 MI 值，然后利用式(9)对 X_i 计算平均的 MI 值，选择影响最大的 m 个 X_i 成为第一级特征参数向量，其余的成为第二级。

3.2 分类算法

一个朴素贝叶斯分类模型建立是需要经过训练样本的学习来构造的。首先初始化类集，令 $\Pr(c_{legit}) = \Pr(c_{spam}) = \frac{1}{2}$，确定 λ 的值。使用训练样本训练分类器，分别得到 $\Pr(c_i | E)$，$\Pr(x_j | c_i)$。当要对一封新邮件 $e=(x_1, x_2, \cdots, x_n)$ 分类时，按以下步骤进行：

1) 分别计算($\Pr(x_j | c_i)$)

2) 使用公式(7)计算 β

3) 比较 β 与 λ 的大小，当 $\beta > \lambda$ 时，将邮件归为 c_{spam} 类，并阻塞；否则归为 c_{legit} 类，让邮件通过

4) 完成分类

对于改进的双级贝叶斯(DLB)分类模型的分类算法与朴素贝叶斯分类算法的不同之处主要是在学习训练样本上，双级贝叶斯分类算法在训练样本时要使用式(8)计算各个参数的 MI 值，然后使用式(9)计算各个参数对其他参数的影响均值 $I(X_i)$，选取较高的 m 个参数成为第一级特征值向量，其他的参数为第二级特征值向量。对新邮件分类时，利用第一级特征向量分类，如果不能确定是否为垃圾邮件，再使用第二级特征向量分类。

4 实验、性能分析

在邮件过滤任务中，性能的评估通常是使用以下几个指标：过滤的准确率、错误率、漏报率。准确率是分类器正确分类邮件的概率，$R_{acc} = \frac{n_{l\to l} + n_{s\to s}}{N_L + N_S}$；错误率是分类器将合法邮件分类为垃圾邮件的概率，$R_{err} = \frac{n_{l\to s}}{N_L + n_{l\to s}}$；漏报率是分类器将垃圾邮件分类为合法邮件的概率，$R_{miss} = \frac{n_{s\to l}}{N_s + n_{s\to l}}$；$n_i$ 表示分类器分类后的邮件数量，N_L 和 N_S 分别表示合法和垃圾邮件的数量，$l\to l$ 表示将合法邮件分类为合法邮件，$s\to s$ 表示将垃圾邮件分类为垃圾邮件，$l\to s$ 表示将合法邮件分类为垃圾邮件，$s\to l$ 表示将垃圾邮件分类为合法邮件。作为比较的基础，在没有使用分类器的情况下，各种概率计算如下：

$R_{acc} = \frac{n_{l\to l}}{N_L + N_S}$，$R_{err} = \frac{n_{l\to s}}{N_L + n_{l\to s}}$，$R_{miss} = \frac{n_{s\to l}}{N_S + n_{s\to l}} = \frac{N_S}{N_S + N_S}$。

在仿真实验中，我们发现，由于双级贝叶斯(DLB)分类模型在学习训练样本时需要计算它们之间的相互信息影响 MI 和 I，因此训练时间比朴素的贝叶斯分类模型所需的时间要多一些。但是，在分类时，朴素贝叶斯分类模型所计算的特征

向量为$\langle x_1, x_2, \cdots, x_n \rangle$，DLB模型计算特征向量分为两级，$L_1 = \langle x_1, x_2, \cdots, x_k \rangle$和$L_1 = \langle x_{n-k}, \cdots, x_n \rangle$，在运算规模上DLB分类模型要小于朴素贝叶斯分类模型，特别是当n很大时更为明显。表1显示了实验的结果。

表1 实验结果

	λ	R_{acc}(%)	R_{err}(%)	R_{miss}(%)
NB	1	96.8	0.44	1.1
DLB		96.5	0.39	0.8
Baseline (no filter)		53	0	50
NB	9	97.6	0.32	1.5
DLB		98.0	0.26	1.4
Baseline (no filter)		53	0	50
NB	999	7.9	0.25	2.2
DLB		98.4	0.17	2.0
Baseline (no filter)		53	0	50

从实验中我们发现DLB分类模型的学习时间要大于朴素贝叶斯分类模型所需要的时间，但是在大多数条件下，分类的准确率要优于朴素的贝叶斯分类模型，同时其错误率和漏报率要小于朴素贝叶斯分类模型。

总结 本文介绍了朴素的贝叶斯分类模型在邮件过滤中的应用，并且在此基础上提出改进的双级贝叶斯(DLB)分类模型，给出了改进的分类算法，并且进行仿真实验。通过实验结果可以看出新的模型在大部分性能指标上都优于一般的朴素贝叶斯分类模型。我们的下一步研究工作将进一步分析特征参数之间的相互信息对分类效果的影响，并考虑对邮件分类后进行回馈学习对模型分类性能的影响。

参考文献

1 Huang Cecil, Darwiche A. Inference in Belief Networks: A Procedural Guide. Int Journal of Approximate Reasoning, 1996, 15: 255～263

2 Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning, 1997, 29(2-3): 131～163

3 Yerazunis W S. The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past it. January 2004. Presented at the 2004 MIT Spam Conference

4 PaulGraham. com. a Plan for spam. www. paulgraham. com/spam. html

5 PaulGraham. com. Better Bayesian Filtering. www. paulgraham. com/better. html

6 Androutsopoulos I, Paliouras G, Karkaletsis V, et al. Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach. In: Proc. of the Workshop on Machine Learning and Textual Information Access, 4th European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France, 2000. 1～13

7 Hastie T, Tibshirani R, Friedman J[美]著. 范明，柴玉梅等译. 统计学学习基础——数据挖掘、推理与预测. 北京：电子工业出版社，2004

8 石洪波，王志海，等. 一种限定性的双层贝叶斯分类模型. 软件学报，2004, 15(2)

(上接第109页)

垃圾邮件处理中尤为重要。分析其原因是：SAC根据数据本身"抱团"的性质来选取特征项，不但考虑短语出现的位置和顺序，而且特征项的长度仅仅由特征项本身决定。

指标 / 指标	查准率	查全率	分类时间(秒/每邮件)
SAC[PG贝叶斯分类算法]	98.9%	95.1%	0.523
切词[PG贝叶斯分类算法]	68.9%	64.6%	0.271
切词[朴素贝叶斯(1500)]	60.6%	75.4%	0.486
切词[朴素贝叶斯(1800)]	63.3%	77.5%	0.521

图6 切词和SAC下中文垃圾邮件过滤指标对比

因此，经SAC抽取后得到的特征项"分类质量"高，系统的输出指标有了明显的提高。

此外，使用SAC的另外一个优势在于系统不需要切词软件的支持和安装词典库。

应用SAC分类时间略长于其他方法。这一点可以通过增加"时间窗口"的方法加以解决。即对输入的语料增加时间参数的限制，使系统只处理一定时间范围内的语料。这样做与垃圾邮件在一定时间内具有重复性的特点是一致的。

结束语 本文针对中文垃圾邮件过滤指标低的问题，提出了一种新的基于后缀数组聚类的中文邮件特征项抽取方法SAC，并将其应用于PG贝叶斯算法，实现中文垃圾邮件的过滤。实验表明，系统的查准率和查全率得到显著提高。下一步，我们需要进一步缩短算法的分类时间，使得算法能够在大规模邮件过滤上获得更好的性能。

参考文献

1 Sahami M, Dumais S, Heckerman D, et al. A Bayesian Approach to Filtering Junk E-mail. In: AAAI Workshop on Learning for Text Categorization, Madis on, Wisconsin, 1998. 55～62

2 Graham P. Better Bayesian filtering. URL: http://paulgraham. com/better. html, 2003

3 Graham P. A Plan for Spam. URL. http://paulgraham. com/spam. html, 2002

4 Segal R, Crawford J, Kephart J, et al. SpamGuru: An Enterprise Anti-Spam Filtering System. In: Proceedings of First Conference on Email and Anti-Spam (CEAS), Mountain View, CA. 2004. URL: http://www. ceas. cc/papers-2004/126. pdf

5 李国栋，李卫. 基于文本分类技术的垃圾邮件识别系统. 微电子学与计算机，2004, 21(6): 143～146

6 刘新斌，李俊. 一种基于N-gram组合的中文垃圾邮件过滤方法. 微电子学与计算机，2004, 21(12): 85～91

7 Zamir O, Etzioni O. Grouper: A dynamic clustering interface to web search results. Eighth International World Wide Web Conference, TorontTo, 1999

8 Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. furst edition. In: New York, USA: published by the press syndicate of the university of Cambridge, 1997. 90～91

9 Ukkonen E. On-line construction of suffix-trees. Algorithmica, 1995, 14(3): 249～260

10 Manber U, Myers G. Suffix arrays: A new method for on-line string searches. In: Proceedings of the First Annual ACM_ SIAM Symposium on Discrete Algorithms, 1990. 319～327

11 Abouelhoda M, Kurtz S, Ohlebusch E. Replacing Suffix Trees with Enhanced Suffix Arrays. Journal of Discrete Algorithms, 2004, 2(1): 53～86