计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(1): 94~101, 2005

使用最大熵模型进行中文文本分类

李荣陆 王建会 陈晓云 陶晓鹏 胡运发
(复旦大学计算机与信息技术系 上海 200433)
(lironglu@163.net)

Using Maximum Entropy Model for Chinese Text Categorization

Li Ronglu, Wang Jianhui, Chen Xiaoyun, Tao Xiaopeng, and Hu Yunfa
(*Department of Computing and Information Technology, Fudan University, Shanghai* 200433)

Abstract With the rapid development of World Wide Web, text classification has become the key technology in organizing and processing large amount of document data. Maximum entropy model is a probability estimation technique widely used for a variety of natural language tasks. It offers a clean and accommodable frame to combine diverse pieces of contextual information to estimate the probability of a certain linguistics phenomena. This approach for many tasks of NLP perform near state-of-the-art level, or outperform other competing probability methods when trained and tested under similar conditions. However, relatively little work has been done on applying maximum entropy model to text categorization problems. In addition, no previous work has focused on using maximum entropy model in classifying Chinese documents. Maximum entropy model is used for text categorization. Its categorization performance is compared and analyzed using different approaches for text feature generation, different number of feature and smoothing technique. Moreover, in experiments it is compared to Bayes, KNN and SVM, and it is shown that its performance is higher than Bayes and comparable with KNN and SVM. It is a promising technique for text categorization.

Key words text classification; maximum entropy model; features; *N*-Gram

摘 要 随着WWW的迅猛发展,文本分类成为处理和组织大量文档数据的关键技术.由于最大熵模型可以综合观察到各种相关或不相关的概率知识,对许多问题的处理都可以达到较好的结果.但是,将最大熵模型应用在文本分类中的研究却非常少,而使用最大熵模型进行中文文本分类的研究尚未见到.使用最大熵模型进行了中文文本分类.通过实验比较和分析了不同的中文文本特征生成方法、不同的特征数目,以及在使用平滑技术的情况下,基于最大熵模型的分类器的分类性能.并且将其和Bayes,KNN,SVM三种典型的文本分类器进行了比较,结果显示它的分类性能胜于Bayes方法,与KNN和SVM方法相当,表明这是一种非常有前途的文本分类方法.

关键词 文本分类;最大熵模型;特征;*N*-Gram

中图法分类号 TP391; TP18

1 引 言

随着WWW的迅猛发展,在线文档信息的迅速增加,文档分类成为处理和组织大量文档数据的关键技术.所以,研究利用计算机进行自动文档分类成为自然语言处理和人工智能领域中一项具有重要应用价值的课题.现有的分类方法主要是基于统计理论和机器学习方法的,比较著名的文档分类方法有 Bayes[1], KNN[2], LLSF[3], Nnet[4], Boosting[5]

收稿日期:2003-06-04;**修回日期**:2004-04-30
基金项目:国家自然科学基金项目(60173027)

及 SVM[6]等.

卡内基梅隆大学的 Yang[2,7]使用英文标准分类语料,对常用的多种分类方法进行比较客观的比较后,得出的结论是 KNN 和 SVM 较其他方法有更高的分类准确性和稳定性.KNN 方法是一种基于要求的或懒惰的学习方法,它存放所有的训练样本,直到测试样本需要分类时才建立分类,所以它的分类时间是非线性的. 而且,当训练文档数增加时,其分类时间将急剧增加. SVM 方法本质上是一种两类分类器,对于两类分类,它的时间复杂度是线性的. 但是,如果要使用 SVM 分类器实现多类分类,必须构造多个 SVM 分类器. 一般使用一对剩余(one-vs-rest)[6]的方法来进行多类分类,对于 K 类分类问题,必须构造 $K-1$ 个分类器. 如果类别数目较多时,分类时间同样难以忍受. 而且,SVM 分类器的训练时间也比较长. 本文中,我们使用了自然语言处理中的一种统计模型——最大熵模型进行文本分类,它的训练时间和分类时间都是线性的. 而且,通过实验我们发现,基于最大熵模型的文本分类器的分类准确率与 KNN 和 SVM 分类器不相上下.

有关最大熵的概念可以追溯到很早以前,它反映了人类认识世界的一个朴素原则,即在对某个事件一无所知的情况下,选择一个模型使它的分布应该尽可能均匀. 而现实世界中我们面对的更多的问题是,在已经知道事件的许多先验知识的情况下,如何选择一个合适的模型来对事件做出预测. 最大熵模型正是用来解决这个问题的. 直观地说,最大熵模型就是拟合所有已知事实,保持对未知事件的未知状态. 换而言之,就是给定一些事实集,选择一种模型与现有事实一致,对于未知事件尽可能使其分布均匀.

从 20 世纪 90 年代开始,最大熵方法开始用于大规模真实文本的处理,越来越多的研究人员被这种方法的灵活性和包容性以及优异的处理结果所吸引. 它可以对非常广泛的自然语言现象建立概率模型,可以综合观察到各种相关或不相关的概率知识,对许多问题的处理结果都达到或超过了其他方法. 近年来,最大熵模型被广泛地应用于自然语言处理中,包括分词、词性标注、词义排歧、短语识别、机器翻译等[8~10].

但是,将最大熵模型应用在文本分类中的研究却非常少,而使用最大熵模型进行中文文本分类的研究尚未见到. Adwait[8]在他的博士论文中首次将最大熵模型应用于文本分类,使用 ME DEFAULT 和 ME IFS 两种方法,对基于最大熵模型和基于决策树的分类方法进行了比较. 但是,他在实验中使用的特征是二值的,这对句子层面的应用来说也许是足够的,但是对于文本分类这种基于文档层面应用很难捕获充足的信息. 因为,文档分类中不能仅仅通过词的存在与否来判断它对某一篇文档语义的贡献,更准确的方法是使用词频. 后来,Kamal[11]将词频作为特征函数的值进行了文本分类的研究,并且对基于最大熵模型和 Bayes 模型的分类方法进行了比较.

但是,Adwait 和 Kamal 的研究都存在以下问题:

(1) 没有对特征进行平滑处理;

(2) 没有将基于最大熵模型的分类方法和 KNN,SVM 等常用的分类准确率和稳定性较高的分类器进行比较;

(3) 没有考虑文本特征生成方法对基于最大熵模型的分类方法的影响.

本文中,我们使用 N-Gram 和分词两种中文文本特征生成方法,对基于最大熵模型的文本分类方法和 Bayes,KNN,SVM 三种常用的文本分类方法进行了比较,我们发现它的分类准确率要优于 Bayes 方法,与 KNN 和 SVM 不相上下. 并且,使用了绝对折扣(absolute discounting)[12]的平滑技术对特征进行了平滑处理,我们发现使用平滑技术后,文本分类的准确率在一定程度上有所提高.

2 最大熵模型

最大熵模型是用来进行概率估计的. 假设 a 是某个事件,b 是事件 a 发生的环境(或称上下文),我们想知道 a 和 b 的联合概率,记为 $p(a,b)$. 更一般地,设所有可能发生的事件组成的集合为 A,所有环境组成的集合为 B,我们想知道,对于任意给定的 $a\in A, b\in B$,概率 $p(a,b)$是多少?

我们把这个问题放到自然语言处理的领域来讨论,对于文本分类问题,一个文档分到某个类别可以看成一个事件,文档中出现的词可以看成这个事件发生的环境,我们想知道包含词 b 的文档属于某一类 a 的概率. 很容易想到的方法是通过训练语料进行统计. 给定一个训练集,定义 $A=\{a_1,a_2,\cdots,a_m\}$是文档所属类别集,$B=\{b_1,b_2,\cdots,b_n\}$是文档的特征词集,$num(a_i,b_j)$为训练集中二元组(a_i,b_j)出现的次数,那么我们可以使用如下公式进行概率估计:

$$\tilde{p}(a_i,b_j)=\frac{num(a_i,b_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}num(a_i,b_j)}. \quad (1)$$

这个方法有个很大的问题,即"稀疏事件"(sparse evidence)问题,即便是很大的训练文本,很多二元组(a_i,b_j)仍然没有出现,武断地认为它的概率为0显然是不可取的. 最大熵模型是这样来解决稀疏事件问题的,它使未知事件的概率分布总是尽可能均匀,即倾向于得到最大熵. 例如一个军事、政治和科技的3类文本分类问题,我们得知,出现"飞机"这个词的80%的文档属于军事类别,对于"飞机"这个词在其他两类中的分布未知. 根据最大熵原则,如果给定一个包含"飞机"这个词文档,那么认为文档以0.8的概率属于军事类别,分别以0.1的概率属于其他两类;如果文档中不包含"飞机"这个词,那么认为文档分别以相同的、1/3的概率属于每一个类. 即在符合已知约束的情况下,使未知事件的分布尽可能均匀.

具体来说,根据Shannon的定义,熵的计算公式如下:

$$H(p)=-\sum_{x}p(x)\log_2 p(x). \quad (2)$$

那么,求解满足最大熵原则的概率分布的公式如下:

$$p^*=\arg\max_{p\in P}H(p). \quad (3)$$

如果没有其他任何先验知识,根据熵的性质,式(3)得到最大值的条件是:

$$p(a\mid b)=\frac{1}{|A|}, \quad (4)$$

因为, $\sum_{a\in A}p(a\mid b)=1$.

但是,尽管训练语料中不能给出所有二元组(a_i,b_j)的概率值,但能够给出部分二元组的概率值,或某些概率需要满足的条件. 即问题变成求部分信息下的最大熵或满足一定约束的最优解.

如何表示这些部分信息呢? 研究者引入了特征函数的概念(有时简称为特征). 特征函数一般情况下是一个二值函数$f(a,b)\to\{0,1\}$,例如对于上述的文本分类问题,我们可以定义特征函数为

$$f(a,b)=\begin{cases}1, & (a=\text{事类})\wedge(b=\text{飞机}),\\ 0, & \text{otherwise}.\end{cases}$$

对于特征函数f_i,它相对于经验概率分布$\tilde{p}(a,b)$的期望值为

$$E_{\tilde{p}}f_i=\sum_{a,b}\tilde{p}(a,b)f_i(a,b). \quad (5)$$

特征函数f_i相对于模型$p(a\mid b)$的期望值为

$$E_pf_i=\sum_{a,b}\tilde{p}(b)p(a\mid b)f_i(a,b). \quad (6)$$

我们限制在训练集中,这两个期望值相同,即

$$E_pf_i=E_{\tilde{p}}f_i. \quad (7)$$

我们将式(7)称为约束. 显然,可以定义很多这样的特征函数,它们之间可以是互不相关的,甚至描述问题的角度也可以是完全不同的,刻画问题的粒度也可大可小. 总之,特征函数很灵活地将许多分散、零碎的知识组合起来完成同一个任务. 给定k个特征函数$f_1,f_2,\cdots,f_k$,我们可以得到所求概率分布的k组约束.

$$E_pf_i=E_{\tilde{p}}f_i, \quad (8)$$

其中,$i=1,2,\cdots,k$. 现在,我们的问题就变成了满足一组约束条件的最优解问题,即

$$P=\{p\mid E_pf_i=E_{\tilde{p}}f_i,i=1,2,\cdots,k\},$$
$$p^*=\arg\max_{p\in P}H(p). \quad (9)$$

求解这个最优解的经典方法是拉格朗日乘子算法,本文直接给出结论. 式(9)要求的p^*具有下面的形式:

$$p^*(a\mid b)=\frac{1}{\pi(b)}\exp\Big(\sum_{i=1}^{k}\lambda_if_i(a,b)\Big), \quad (10)$$

其中,$\pi(b)$是规一化因子.

$$\pi(b)=\sum_{a}\exp\Big(\sum_{i=1}^{k}\lambda_if_i(a,b)\Big), \quad (11)$$

λ_i是参数,可以看成特征函数的权值. 如果通过在训练集上进行学习,知道了λ_i的值,就得到了概率分布函数,完成了最大熵模型的构造. 设$|A|$是事件集的大小,k是特征函数的数目,从式(10)我们可以看到,最大熵模型的时间复杂度是$O(k|A|)$.

为了构造最大熵模型,我们必须求出参数λ_i,文本中我们使用了GIS[13]算法. 设N是训练样本集的大小,$|A|$是事件集的大小,算法经过P次迭代后收敛,则整个复杂度是$O(NP|A|)$.

3 基于最大熵模型的文本分类方法

3.1 特征函数的选择

将事件集A当做类别集,将上下文环境集B当做文档集,那么我们就可以使用式(10)求任意一篇文档$b_i\in B$属于任意类别$a_j\in A$的概率$p(a_j\mid b_i)$. 对于不存在兼类的分类问题,只要选择$\arg\max_j p(a_j\mid b_i)$就是文档$b_i$的所属类别;对于存在兼类的分类问题,只要定义一个阈值ε,文档b_i属于所有$p(a_j\mid b_i)>\varepsilon$的类别.

这样,对于文本分类,最重要的问题是特征函数的选择.一般情况下,特征函数是一个二值函数,这对文本分类这种基于文档层面应用来说是不够的.我们选择一个"词-类别"对作为一个特征,使用词频作为特征值.对于词 w 和类别 a',它的特征函数如下所示:

$$f_{w,a}(a,b)=\begin{cases}num(b,w), & a=a',\\ 0, & \text{otherwise},\end{cases} \tag{12}$$

其中,$num(b,w)$表示词 w 在文档 b 中出现的次数.

从式(10)我们知道,最大熵模型的时间复杂度是 $O(k|A|)$,$|A|$ 是事件集的大小,k 是特征函数的数目.对于文本分类问题来说,每一篇文档中的特征是非常稀疏的.经过我们初步统计,一般为小于特征总数的十分之一.所以,使用最大熵模型进行文本分类时间是非常快的.

3.2 平滑(smoothing)技术

对于文本分类问题来说,文档中的特征是非常稀疏的.对于一篇文档,其中大部分特征函数的 $num(b,w)$为 0.对于这种情况可以采用平滑技术来进行处理.现在还没有一种专门针对最大熵模型的平滑技术.针对文本分类问题,本文中我们使用了绝对折扣[12]平滑技术.

绝对折扣平滑技术是指对于模型中观察到的事件进行折扣,减掉一个固定值 d,然后将折扣得来的概率分摊到所有未现事件中.由于特征函数的值是词频,对特征出现次数进行折扣时不涉及到保持概率和为 1 的问题,所以我们只需直接给所有出现次数为 0 的特征一个值 d'即可.使用绝对折扣平滑技术后,特征函数(12)将变为

$$f_{w,a'}(a,b)=\begin{cases}num(b,w), & num(b,w)\neq 0,\\ d', & num(b,w)=0.\end{cases} \tag{13}$$

本文中,我们取 d'的值为 0.1.

4 中文文本特征的生成方法

4.1 文本分类模型

图 1 是一个文本分类的基本模型,从图中可以看到文本分类由训练过程和测试过程构成的.在训练过程中,首先要生成训练文本的特征,得到特征的集合;特征选择算法从文本特征的全集中抽取一个最优的特征子集;这里的"最优"子集是由评价算法来判定的,它根据分类器对由特征子集所表示的训练文本进行分类,并对分类性能进行性能评价.在分类过程中,首先将测试文档用最优特征子集表示,再经分类器分类,得到测试文本所属的类别.

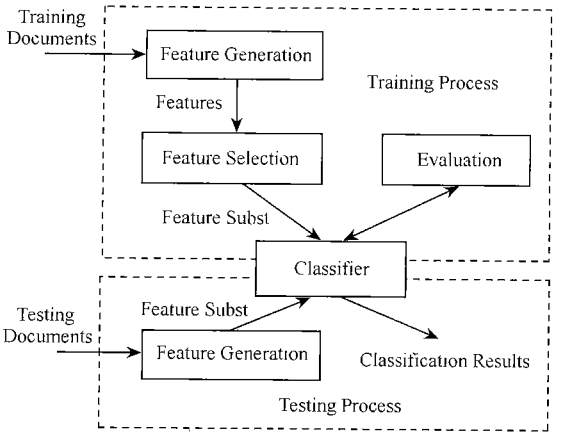

Fig. 1 Model of text categorization.

图 1 文本分类模型

对于中文文本分类和西方语种文本分类,最主要的差别在于文本特征生成模块.对于英文等西方语种,如果使用词作为文本特征,则文本不需要进行分词,因为每个词之间已经使用空格隔开了,而且一般对每个词都要进行词干抽取;如果使用 N-Gram 项作为文本特征,也可以分别以字母或词为单位来生成文本特征.对于中文文本分类,同样可以使用词或 N-Gram 项作为文本特征,但在中文文本中词和 N-Gram 项的粒度与英文文本中有所不同,而且中文的分词问题还没有完全解决.

为了对最大熵模型在中文文本分类中的性能做出评价,我们分别使用分词的方法和 N-Gram 的方法来生成文本的特征.分词我们采用了联想-回溯算法,N-Gram 项的生成方法如第 4.2 节所述.

4.2 N-Gram 项的生成方法

4.2.1 N-Gram 项和中文文本分类

假设训练文档库 D 有 N_D 个文档,每个文档平均包含 N_s 个句子,而句子的平均长度为 L_s,则这个训练文档库包含的 N-Gram 最多达 $N_DN_sL_s(L_s+1)/2$.由此可见,文档中包含的 N-Gram 项非常丰富.这提醒我们,在用 N-Gram 项进行文档分类时必须有所选择.从另一方面来看,文档分类是面向语义的操作,因此,用于文档分类的文档属性应该能够尽可能地表现文档的语义.显然,并不是所有出现在文档中的 N-Gram 项都对分类有用.一个 N-Gram 项对分类的有用性或者说分辨能力,可以从 3 个方面来衡量:频度、分散度和集中度.下面分别给出它们的定义.

定义 1. 在文档 d 中，N-Gram 项 t 的频度用它在 d 中出现的次数 tf 表示.

定义 2. 在文档类 c 中，N-Gram 项 t 的分散度用 c 中包含 t 的文档数目 df 表示. df 越大，则 t 在 c 中越分散；反之越不分散.

定义 3. 在文档集 D 中，N-Gram 项 t 的集中度用 D 中包含 t 的文档类数目 cf 表示. cf 越小，则 t 在 D 中越集中；反之，越不集中.

直观地，对于 N-Gram 项 t，其在文档中的频度越高、在文档类里的分散度越大、在训练文档集内的集中度越强，则它对分类越有用，即分辨度越强. 不过，目前还没有找到很好的数学方法来综合频度、分散度和集中度这 3 个因素，使得选出的文档属性能够获得最优分类效果. 为了减少提取不必要的 N-Gram 项，在提取 N-Gram 项时加如下两个约束条件：

约束 1. 对于预先给定的最小频度值 $min\text{-}tf$，在文档 d 中某一 N-Gram 项 t 被提取的先决条件是它在 d 中的 $tf \geqslant min\text{-}tf$.

约束 2. 对于预先给定的最小分散度值 $min\text{-}df$，在文档类 c 中某一 N-Gram 项 t 被提取的先决条件是它在 c 中的 $df \geqslant min\text{-}df$.

在实验中，我们一般取 $min\text{-}tf$ 和 $min\text{-}df$ 为 2.

4.2.2 N-Gram 项生成算法

一种直接的 N-Gram 项提取方法是只扫描一遍文档，一次性将所有满足上述两个约束条件的 N-Gram 项取出. 由于只需扫描一遍文档，这种方法对于较小的训练文档库是有效的. 但对于大训练库则需要很大的内存空间. 否则，就得在内存和外存之间不断交换中间结果. 这里采用一种分步提取的方法，其基本思想是先提取符合约束条件的 1-Gram 项；从选择得到的 1-Gram 项构造候选 2-Gram 项，剔除其中不符合约束条件的候选项，得到真正需要的 2-Gram 项. 依此方法，提取其他 N-Gram($N=3,4,\cdots$)项.

为了说明 N-Grams 提取算法，先给出一个定义和引理.

定义 4. 子项. 若有 i-Gram 和 j-Gram($i \geqslant j$)，且 j-Gram 包含在 i-Gram 中，则称 j-Gram 为 i-Gram 的子项，记为 $j\text{-Gram} \subseteq i\text{-Gram}$.

性质 1. 若有 i-Gram 满足约束 1 和约束 2，则 i-Gram 的所有子项都满足约束 1 和约束 2.

性质 1 构成了分步提取 N-Gram 项的算法基础.

算法 1. N-Gram 项生成算法[14].

输入：文档库 D，$min\text{-}tf$，$min\text{-}df$ 和 $max\text{-}N$(N 的最大值).

输出：满足约束条件 1 和 2 的 N-Gram 项的集合 $S(N=1\sim max\text{-}N)$.

(1) 求 1-Gram 项集合 S_1：逐个扫描文档库 D 中的文档，提取所有 1-Gram 项，抛弃不满足约束 1 和约束 2 的项，得到 1-Gram 项集合 S_1；

(2) 求 2-Gram 项集合 S_2：将 S_1 中的项两两组合，得到候选 2-Gram 项集合 C_2. 抛弃 C_2 中不满足约束 1 和约束 2 的项，得到 2-Gram 项集合 S_2；

(3) 求 i-Gram 项集合 $S_i(i=3\sim max\text{-}M)$：

① 由 $S_{i-1}(i=2\sim max\text{-}M)$ 求候选 i-Gram 项集合 Ci；

② 对于 S_{i-1} 中的任意两项 t_m，t_n，$t_m(k)$ 和 $t_n(k)(k=1\sim(i-1))$ 分别表示 t_m 和 t_n 中的第 k 个字. 若 $t_m(k+1)=t_n(k)(k=1\sim(i-2))$，则 $C_i = C_i \cup t_m t_n(i-1)$；

③ 抛弃 C_i 中不满足约束 1 和约束 2 的项，得到 i-Gram 项集合 S_i.

使用 N-Gram 项进行文本分类，最基本的要求是所选择的 N-Gram 项能够覆盖文档中的词. 因此，并非 N-Gram 项越多越好. 这就涉及如何选择参数 N 的问题. 根据对中文文档中词的字数构成与分布的统计分析，发现中文文档中主要词条为 1 字、2 字、3 字和 4 字词条，因此，用这些词条可以比较完整地表达文档语义. 这样意味着在用 N-Gram 进行文档分类时，只需取 1-Gram，2-Gram，3-Gram 和 4-Gram 项，即最大的 N 值取 4.

5 实验及结果分析

对于英文文档分类研究，国外有相对标准的训练和测试文档库，这样就可以在共同文档库上比较不同分类方法和系统的性能. 而就中文文档分类而言，目前国内还没有标准、开放的分类文档集可供使用. 所以，我们收集了 20000 余篇新闻网页，通过人工的方式将其分为计算机、交通、环境、经济、医药、军事、政治、体育、艺术、教育共 10 个类. 将某些无法分入这 10 类的网页去掉后，我们一同得到了 16085 篇文档. 我们将这 16085 篇文档分为训练集和测试集两个集合，其中训练集包含了 10723 篇文档，用来进行分类器的学习；测试集包含 5362 篇文档，用来对分类器的性能进行评价.

我们从以下几个方面考察基于最大熵模型的文

本分类器的性能,使用的性能评价指标为国际上常用的,准确率和召回率的盈亏平衡点(precision/recall breakeven point)处分类器的微平均准确率[7].

① 分别使用分词和 N-Gram 的方法生成文本特征时分类器的性能;

② 选不同数量的特征时分类器的性能;

③ 平滑技术对分类器性能的影响;

④ 特征函数对分类其性能的影响;

⑤ 与 Bayes,KNN 和 SVM 分类器的性能比较.

在文本分类的训练阶段,我们分别使用分词和 N-Gram 的方法来生成每一篇文档的特征,使用 χ^2 的方法进行特征选择. 在训练最大熵模型的参数时,我们使用了 GIS 算法,迭代 100 次.

为了对使用不同文本特征生成方法、不同特征数目时基于最大熵模型分类器性能做出评价,我们分别使用分词和 N-Gram 的方法生成文本的特征,在特征属性数目从 300～2500 的情况下,对基于最大熵模型的文本分类方法进行了测试. 分类器的微平均准确率如表 1 所示. 表 1 中,1-Gram 表示只取 1-Gram 项,2-Gram 同此;1/2-Gram 表示既取 1-Gram 项,又取 2-Gram 项,1/2/3-Gram 和 1/2/3/4-Gram 与此类似.

Table 1 Performance Comparison Among Different Feature Generation Methods

表 1 不同属性生成方法下微平均准确率比较表

Number of Features	Word Segmentation	1-Gram	2-Gram	1/2-Gram	1/2/3-Gram	1/2/3/4-Gram
300	90.37	86.20	85.13	86.85	86.74	86.85
500	91.44	86.56	86.52	87.06	87.17	87.17
1000	92.41	86.10	88.24	87.49	87.81	87.59
1500	92.73	86.74	88.89	88.02	88.24	88.02
2000	93.37	86.63	89.63	88.24	88.34	88.24
2500	93.69	86.63	88.98	88.34	88.02	87.91

对于基于最大熵模型的文本分类方法,从表 1 中可以得到如下结论:

(1) 使用分词的方法来生成文本特征要优于 N-Gram 方法. 从表 1 中我们可以看到,无论 N-Gram 方法中 N 取多少,N-Gram 项如何结合,特征数目取多少,使用分词的方法、分类准确率总是高于 N-Gram 的方法.

(2) 随着特征数目的增加,分类准确率逐步提高;当达到一定的数目后,准确率不再升高,反而有所下降,但下降不是很明显.

为了考察平滑技术和特征函数对基于最大熵模型的文本分类技术的影响,在特征数目 300～2500 的情况下,分别使用二值特征函数和基于词频的特征函数(式(12),(13)),对分类器的性能进行了测试和比较. 表 2 为测试结果. 表 2 中,*FBFF* 代表使用基于词频的特征函数;*BVFF* 代表使用二值特征函数;*AD* 代表使用绝对折扣平滑技术;*NS* 代表不使用平滑技术.

Table 2 Affection of Smoothing Technique While Using Different Feature Functions

表 2 不同特征函数下平滑技术对微平均准确率的影响表

Number of Features	Word Segmentation				1/2/3-Gram			
	FBFF		*BVFF*		*FBFF*		*BVFF*	
	AD	*NS*	*AD*	*NS*	*AD*	*NS*	*AD*	*NS*
300	91.98	90.37	87.06	88.56	88.34	86.74	88.45	88.24
500	92.30	91.44	87.27	89.52	89.41	87.17	89.84	88.02
1000	92.73	92.41	88.66	90.48	89.52	87.81	89.84	88.24
1500	92.73	92.73	88.98	90.69	89.20	88.24	90.59	88.45
2000	93.26	93.37	89.52	90.91	89.41	88.34	89.95	88.56
2500	93.16	93.69	88.88	92.19	89.73	88.02	90.16	88.77

对于基于最大熵模型的文本分类方法,从表 2 中可以得到如下结论:

(1) 使用基于词频的特征函数要优于二值特征函数.

(2) 使用平滑技术,在一定程度上可以提高分类准确率,但是有时也会降低分类的准确率,如在表2中,特征数目为2000,使用词频特征函数和分词的方法生成文本特征时,准确率出现下降. 而且,如果使用二值特征函数,使用分词的方法生成文本特征时,平滑技术也造成准确率的下降.

为了对基于最大熵模型的文本分类技术和其他类型的分类器做出比较,我们选择了文本分类中常用的3种方法:Bayes, KNN 和 SVM. 其中,Bayes 使用多项式模型[15];KNN[16]方法 K 值取 50;SVM 方法选择多项式核函数,多项式核的阶数为3,使用一对剩余[6,17]的方法进行多类分类.

Table 3 Performance Comparison Among Different Classification Approaches

表3 不同分类方法下微平均准确率比较表

Number of Features	Bayes		KNN		SVM		ME	
	Word Segmentation	Gram	Word Segmentation	Gram	Word Segmentation	Gram	Word Segmentation	Gram
300	82.76	65.52	90.58	90.36	91.47	91.17	91.98	88.34
500	83.40	61.99	91.76	91.22	92.33	91.59	92.30	89.41
1000	82.33	62.74	92.83	92.18	92.52	91.06	92.73	89.52
1500	82.23	62.63	93.04	92.29	92.87	92.98	92.73	89.20
2000	76.87	63.17	92.18	92.29	93.36	92.77	93.26	89.41
2500	74.73	64.13	92.93	92.29	94.31	92.89	93.16	89.73

从表3中可以得到如下结论:

(1) 基于最大熵模型的文本分类方法优于Bayes方法,这与文献[11]的实验结果基本吻合;

(2) 基于最大熵模型的文本分类方法至少与KNN和SVM方法不相上下.

6 结束语

本文中,我们使用最大熵模型进行了文本分类的研究. 并且就特征生成方法、特征数目、特征函数的选择和平滑技术对基于最大熵模型的分类器的性能影响进行了实验和分析. 实验结果显示,基于最大熵模型的分类器是一种非常有前途的分类器. 但是,我们在实验中也发现,基于最大熵模型的分类器的稳定性比 KNN 方法要差一些,使用不同的训练文档测试结果相差较大,这些问题还有待于我们进一步研究. 而且,实验的规模也有待于进一步扩大.

参 考 文 献

1 D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In: Proc. of the 10th European Conf. on Machine Learning. New York: Springer, 1998. 4~15

2 Y. Yang, X. Lin. A re-examination of text categorization methods. In: The 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in the Information Retrieval. New York: ACM Press, 1999

3 Y. Yang, C. G. Chute. An example-based mapping method for text categorization and retrieval. ACM Trans. on Information Systems, 1994, 12(3): 252~277

4 E. Wiener. A neural network approach to topic spotting. The 4th Annual Symp. on Document Analysis and Information Retrieval, Las Vegas, NV, 1995

5 R. E. Schapire, Y. Singer. Improved boosting algorithms using confidence-rated predications. In: Proc. of the 11th Annual Conf. on Computational Learning Theory. New York: ACM Press, 1998. 80~91

6 T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In: Proc. of the 10th European Conf. on Machine Learning. New York: Springer, 1998. 137~142

7 Y. Yang. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1): 76~88

8 R. Adwait. Maximum entropy models for natural language ambiguity resolution: [Ph. D. dissertation]. Pennsylvania: University of Pennsylvania, 1998

9 R. Adwait. A maximum entropy model for part-of-speech tagging. The Empirical Methods in Natural Language Processing Conference, Philadelphia, USA, 1996

10 Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1): 38~73

11 N. Kamal, L. John, M. Andrew. Using maximum entropy for text classification. The IJCAI-99 Workshop on Information Filtering, Stockholm, Sweden, 1999

12 M. Sven, N. Hermann, Z. Jrg. Smoothing methods in maximum entropy language modeling. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, Phoenix, AR, 1999

13 R. Adwait. A simple introduction to maximum entropy models for natural language processing. Institute for Research in Cognitive Science, University of Pennsylvania. Tech. Rep.: 97-08, 1997

14 Zhou Shuigeng. The research on some key techniques of Chinese text database: [Ph. D. dissertation]. Shanghai: Fudan University, 2000 (in Chinese)
(周水庚. 中文文本数据库若干关键技术研究:[博士学位论文]. 上海:复旦大学, 2000)

15 S. Eyheramendy, D. Lewis, D. Madigan. On the naive Bayes model for text categorization. The 9th Int'l Workshop on Artificial Intelligence and Statistics, Key West, Florida, 2003

16 Li Ronglu, Hu Yunfa. A density-based method for reducing the amount of training data in *k*NN text classification. Journal of Computer Research and Development, 2004, 41(4): 539~545 (in Chinese)
(李荣陆, 胡运发. 基于密度的 *k*NN 文本分类器训练样本裁剪方法. 计算机研究与发展, 2004, 41(4): 539~545)

17 C. Hsu, C. Lin. A comparison on methods for multi-class support vector machines. IEEE Trans. on Neural Networks, 2003, 13: 415~425

Li Ronglu, born in 1976. Currently Ph. D. candidate in Fudan University. His main research interests are text database and natural language processing.
李荣陆,1976 年生,博士研究生,主要研究方向为全文数据库和自然语言处理.

Wang Jianhui, born in 1972. Currently Ph. D. candidate in Fudan University. His main research interests are database and knowledge engineering.
王建会,1972 年生,博士研究生,主要研究方向为信息处理、数据库与知识工程.

Chen Xiaoyun, born in 1976. Currently Ph. D. candidate and lecturer in Fudan University. Her main research interests are data minig and machine learning.
陈晓云,1976 年生,博士研究生,讲师,主要研究方向为数据挖掘和机器学习.

Tao Xiaopeng, born in 1970. Post doctor and associate professor of Fudan University. His main research interests are natural language processing and text database.
陶晓鹏,1970 年生,博士后,副教授,主要研究方向为中文信息处理和全文数据库.

Hu Yunfa, born in 1940. Currently professor and Ph. D. supervisor in Fudan University. His main research interests are data engineering and knowledge engineering.
胡运发,1940 年生,教授,博士生导师,主要研究方向为数据工程与知识工程.

Research Background

With the rapid growth of World Wild Web and the steady accumulation of on line document information, text categorization has become one of the key techniques for handling and organizing text data. As a result, using computers to automatically classify documents has emerged as an important direction in NLP and AI community.

In this paper, we propose the use of maximum entropy techniques for text classification. Maximum entropy is a probability distribution estimation technique widely used for a variety of natural language tasks. Given a set of fact, maximum entropy keeps distribution of the unknown events as uniform as possible by conforming to all facts at hand.

Adwait Ratnaparkhi first introduced maximum entropy to text classification and draw comparison to decision trees. Kamal Nigam used word frequency to compute value of feature function, investigated maximum entropy model for text classification and compared it to the Bayes model. However, no previous work has focused on using maximum entropy model in classifying Chinese documents.

In this paper, *n*-gram and word segmentation are respectively used to generate Chinese text feature. We compare maximum entropy model for text categorization with other three approaches, Bayes, KNN and SVM. A detailed analysis is also made.

Our research is supported by the National Natural Science Foundation of China under Grant No. 60173027.