计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(4): 596~602, 2008

文本分类中特征选择的约束研究

徐 燕[1,2] 李锦涛[1] 王 斌[1] 孙春明[1,2] 张 森[1]
[1](中国科学院计算技术研究所 北京 100080)
[2](华北电力大学计算机系 北京 102206)
(xuyan@ict.ac.cn)

A Study on Constraints for Feature Selection in Text Categorization

Xu Yan[1,2], Li Jintao[1], Wang Bin[1], Sun Chunming[1,2], and Zhang Sen[1]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)
[2](*Department of Computer Science, North China Electric Power University, Beijing* 102206)

Abstract Text categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, TC has become a key technique for handling and organizing text data. One of the most important issues in TC is feature selection (FS). Many FS methods have been put forward and widely used in the TC field, such as information gain (IG), document frequency thresholding (DF) and mutual information. Empirical studies show that some of these (e.g. IG, DF) produce better categorization performance than others (e.g. MI). A basic research question is why these FS methods cause different performance. Many existing works seek to answer this question based on empirical studies. In this paper, a theoretical performance evaluation function for FS methods is put forward in text categorization, Some basic desirable constraints that any reasonable FS function should satisfy are defind and then these constraints on some popular FS methods are checked, including IG, DF and MI. It is found that IG satisfies these constraints, and that there are strong statistical correlations between DF and the constraints, whilst MI does not satisfy the constraints. Experimental results on Reuters 21578 and OHSUMED corpora show that the empirical performance of a feature selection method is tightly related to how well it satisfies these constraints.

Key words feature selection; text categorization; information retrieval; information gain; mutual information

摘 要 特征选择在文本分类中起重要的作用. 文档频率(DF)、信息增益(IG)和互信息(MI)等特征选择方法在文本分类中广泛应用. 已有的实验结果表明, IG 是最有效的特征选择算法之一, DF 稍差而 MI 效果相对较差. 在文本分类中, 现有的特征选择函数性能的评估均是通过实验验证的方法, 即完全是基于经验的方法, 为此提出了一种定性地评估特征选择函数性能的方法, 并且定义了一组与分类信息相关的基本的约束条件. 分析和实验表明, IG 完全满足该约束条件, DF 不能完全满足, MI 和该约束相冲突, 即一个特征选择算法的性能在实验中的表现与它是否满足这些约束条件是紧密相关的.

关键词 特征选择; 文本分类; 信息检索; 信息增益; 互信息
中图法分类号 TP391.3

文本分类是根据文档内容将文档归入一个或多个预先定义的类别, 随着网络的发展, 大量的文档数据涌现在网上, 用于处理海量数据的自动文本分类技术变得越来越重要, 自动文本分类已成为处理和

收稿日期: 2006-06-16; 修回日期: 2007-09-21
基金项目: 国家自然科学基金项目(60473002, 60603094); 北京自然科学基金项目(4051004)

组织大量文档数据的关键技术[1].

文本自动分类的主要困难之一是特征空间的维数很高, 特征数达到上万, 甚至几十万. 如何降低特征空间的维数, 提高分类的效率和精度, 成为文本自动分类中需要首先解决的问题.

特征选择是文本分类的一个重要步骤. 特征选择函数是特征(词条)到实数的一个映射. 实际应用中对训练集中每一个词条计算它的特征选择函数值, 移除函数值小于阈值的词条. 近年来, 大量的统计分类方法和机器学习方法被应用到特征选择领域. 现有的特征选择函数主要有文档频率(DF)、信息增益(IG)、互信息(MI)等等.

已有的实验表明[2-3], IG 是最有效的特征选择方法之一, DF 的效果稍差, 但和 IG 基本相似, 而 MI 相对较差.

是什么原因导致了文本分类中特征选择方法表现的差异呢? 文献[2] 通过实验从不同角度比较发现, DF, IG 的出色表现说明高频词汇确实对文本分类有益, 而 MI 性能差的原因是其特征选择倾向于罕见词.

CTD(categorical descriptor term)特征选择方法应用了 IDF 中的文档频率信息和 ICF 中的类别信息[4]. 实验证明 CTD 可以得到比另外的特征选择方法较好的效果, 特别是在文档集中有较多的重叠主题时.

SCIW(strong class information words)[5] 特征选择方法是一种选择带有强类别信息的词的方法. 例如, "football"通常出现在"sport"类里, 因而, 这种方法主要考虑类别信息, 实验证明这种方法在线性分类器上有较好的准确率.

由 CTD 和 SCIW 可见利用类别信息的特征选择算法能够得到较好的效果.

所以, 通过实验发现, 能使特征选择得到好的效果的影响因素有:

1) 使用高频词;

2) 利用类别信息.

然而, 这个结论来自于实验分析, 即都是基于经验的方法得到的. 本文对特征选择算法进行了定性分析, 定义了一组高性能的特征选择算法应该满足的基本性质.

既然文本分类是分类问题的一种, 那么利用分类信息对分类性能的提高是至关重要的. 如果一个词条 t 在文档中出现与否对分类没有丝毫影响, 那么该词条 t 对文本分类没有意义, 可以把它从特征空间中去除, 这时它的特征选择函数的函数值应该是最小的, 相反地, 如果一个文档的分类完全取决于一个词条 t 出现与否, 那么它的特征选择函数的函数值应该是最大的.

根据这种基本想法, 本文定义了特征选择算法需要满足的基本约束条件. 分析和实验证明, 特征提取算法效果的好坏与它们是否符合约束条件的程度是紧密相联的.

1　特征选择方法

本节我们对最常用的特征选择算法 DF, IG, MI 进行概述, DF 和 IG 在文本分类中表现得较好, 而且 IG 在许多实验中都是表现最好的特征提取算法[2-3].

下面给出的 DF, IG, MI 的定义来自文献[2].

1.1　文档频率

词条的文档频率(document frequency)是指在语料中出现该词条的文档的数目. 只有当某词条在较多的文档中出现时才被保留下来, DF 值低于某个阈值的词条是低频词, 将这样的词条从原始特征空间中移除, 不但能够降低特征空间的维数, 而且还有可能提高分类的精度.

DF 是一种最简单的词约简技术, 由于具有相对于语料规模的线性复杂度, 所以它能够容易地被用于大规模的语料特征选择中.

1.2　信息增益

信息增益被广泛应用在机器学习领域[6]. 它通过一个词条在一篇文章中出现与否来计算对类别的信息增益. 设$\{c_i\}_{i=1}^m$为目标空间中类别的集合, 那么词条 t 对类别的信息增益为

$$G(t)=-\sum_{i=1}^{m}p(c_i)\log\ p(c_i)+p(t)\sum_{i=1}^{m}p(c_i\ |t)\times\log\ p(c_i\ |t)+p(\overline{t})\sum_{i=1}^{m}p(c_i\ |\overline{t})\log\ p(c_i\ |\overline{t}).$$

1.3　互信息

互信息广泛应用于统计语言模型[7-8], 对于类别 c 和词条 t, 它们之间的互信息定义为

$$I(t,c)=\log_2\frac{p(t\wedge c)}{p(t)\times p(c)}.$$

这是单个类别的互信息, 将互信息应用于多个类别有两种常用的方法: 设$\{c_i\}_{i=1}^m$为目标空间的类的集合, 则平均和最大互信息分别为

$$I_{\mathrm{avg}}(t)=\sum_{i=1}^{m}p(c_i)I(t,c_i),$$

$$I_{\max}(t) = \max_{i=1}^{m}\{ I(t, c_i)\}.$$

2 类别约束

首先介绍一些在本节及后续章节中用到的符号. t 表示一个词条在文档中出现, $\overline{t}$ 表示一个词条在文档中不出现, $\boldsymbol{T}=\{t, \overline{t}\}$, C_i 表示一个类别, $\boldsymbol{C}=\{c_i\}_{i=1}^{m}$ 表示类别的集合, f 表示特征选择函数, 用 $f(\boldsymbol{C}, \boldsymbol{T})$ 表示 $\boldsymbol{T}$ 与类别 $\boldsymbol{C}$ 的一种关系程度的值.

$\boldsymbol{C}$ 和 $\boldsymbol{T}$ 可以看做两个离散的随机变量, 如果 $\boldsymbol{T}$ 对 $\boldsymbol{C}$ 的值没有影响, 那么通常认为它们是独立的, 反之亦然.

这里说的独立是概率论中的随机变量之间的独立性. 对于独立的离散随机变量 $\boldsymbol{C}$ 和 $\boldsymbol{T}$, 它们应满足如下公式[9]:

对于每一个组对 (c_i, t) 或 $(c_i, \overline{t})$ $(0\leqslant i\leqslant m)$, 有:

$$p(c_i; t) = p(c_i)\times p(t), \tag{1}$$

$$p(c_i; \overline{t}) = p(c_i)\times p(\overline{t}). \tag{2}$$

如果 $\boldsymbol{C}$ 和 $\boldsymbol{T}$ 相互独立, 也就是说 $\boldsymbol{C}$ 和 $\boldsymbol{T}$ 之间没有任何联系, 那么 $\boldsymbol{T}$ 对于 $\boldsymbol{C}$ 毫无意义, 如果 $\boldsymbol{C}$ 的值完全取决于 $\boldsymbol{T}$, 很明显 $\boldsymbol{T}$ 和 $\boldsymbol{C}$ 之间的联系很紧密, $\boldsymbol{T}$ 对 $\boldsymbol{C}$ 很重要.

所以 $f(\boldsymbol{T}, \boldsymbol{C})$ 可表示 $\boldsymbol{T}$ 与类别 $\boldsymbol{C}$ 的一种依赖程度的量化, 特别地, 应该满足类别对特征的一组约束(term-category constrains, TCC)为:

1) 如果 $\boldsymbol{T}$ 和 $\boldsymbol{C}$ 独立, 那么 $f(\boldsymbol{C}, \boldsymbol{T})$ 的值最小(通常为 0);

2) 如果 $\boldsymbol{C}$ 的值完全取决于 $\boldsymbol{T}$, 那么 $f(\boldsymbol{C}, \boldsymbol{T})$ 的值最大.

另外, 最优的情况应该是随着 $\boldsymbol{C}$ 对 $\boldsymbol{T}$ 依赖程度的增加, $f(\boldsymbol{C}, \boldsymbol{T})$ 的值增加, 但由于这个条件暂时无法精确定量分析, 所以暂不列入这组类别对特征的约束条件中.

由于 $\boldsymbol{C}=\{c_i\}_{i=1}^{m}$ 代表类别的集合, 在一个特定的语料集中它是确定的值, 因而可以用 $f(\boldsymbol{T})$ 代替 $f(\boldsymbol{C}, \boldsymbol{T})$, 因为 $\boldsymbol{T}=\{t, \overline{t}\}$, 所以也可以用 $f(t)$ 来代表特征选择函数. 例如 $G(t)$ 是特征选择函数, 可以用 $G(t)$ 代替 $G(\boldsymbol{C}, \boldsymbol{T})$, 在本文中这几个记法视同等价.

3 特征选择函数分析

本节我们讨论分析第 1 节中提到的常用特征选择函数, 分析它们是否满足上述 TCC(term-category constrains).

如果一个词条 t 只在语料中出现了一、两次, 那么式(1)和式(2)的等式两边都等于或接近于零, $\boldsymbol{T}=\{t, \overline{t}\}$ 和 $\boldsymbol{C}$ 几乎是独立的, $f(\boldsymbol{C}, \boldsymbol{T})$ 的值应该较小, 因而, 一般地, 高频词汇比稀有词汇有更高的 $f(\boldsymbol{C}, \boldsymbol{T})$ 值. 满足 TCC 的特征选择算法不倾向于低频词.

3.1 分析信息增益

根据 Moore[10] 的论述可知:

$$G(\boldsymbol{C}, \boldsymbol{T}) = H(\boldsymbol{C}) - H(\boldsymbol{C}\mid\boldsymbol{T}).$$

在这里, $H(\boldsymbol{C})$ 是 $\boldsymbol{C}$ 的熵, $H(\boldsymbol{C}\mid\boldsymbol{T})$ 是 $\boldsymbol{C}$ 在给定 $\boldsymbol{T}$ 时的平均条件熵. $H(\boldsymbol{C}\mid\boldsymbol{T})$ 的性质如下[10]:

1) $0\leqslant H(C\mid T)\leqslant H(C)$.

2) 当且仅当 C 的值完全由 T 所决定时 $H(C\mid T)=0$.

3) 当且仅当 $\boldsymbol{T}$ 和 $\boldsymbol{C}$ 独立时 $H(C\mid T)=H(C)$.

因而:

1) $0\leqslant G(C, T)\leqslant H(C)$.

2) 当且仅当 $\boldsymbol{C}$ 和 $\boldsymbol{T}$ 独立时 $G(C, T)=0$.

3) 当且仅当 $\boldsymbol{C}$ 的值完全由 $\boldsymbol{T}$ 所决定时 $G(C, T)=H(C)$.

由此可见 $G(\boldsymbol{C}, \boldsymbol{T})$ 完全满足 TCC, 也就是说信息增益满足 TCC.

3.2 分析互信息

根据文献[2-3], 类别 c 和词条 t 它们的概率分别为 $p(c)$, $p(t)$, 则互信息 $I(t, c)$ 的定义如下:

$$I(t, c) = \log_2\frac{p(t, c)}{p(t)\times p(c)} = \log_2\frac{p(t\wedge c)}{p(t)\times p(c)}.$$

一般地, 如果 t 和 c 相交的部分较少, 那么 $p(t, c)$ 将小于 $p(t)\times p(c)$, 则 $I(t, c)<0$, 也就是说, 如果 $p(t, c)$ 接近于 0 时, 一般会有 $I(t, c)<0$, 这样可使 $I_{\text{avg}}(t)<0$.

例如, $\{c_i\}_{i=1}^{2}$ 为目标空间中的类别集合, 令 $p(t)=0.5001$, $p(c_1)=0.5$, $p(t\wedge c_1)=0.0001$. 那么 $p(c_2)=1-p(c_1)=0.5$, $p(t\wedge c_2)=p(t)-p(t\wedge c_1)=0.5001-0.0001=0.5$.

$$I(t, c_1) = \log_2\frac{p(t\wedge c_1)}{p(t)\times p(c_1)} = \log_2\frac{0.0001}{0.5001\times 0.5} = -11.288 < 0.$$

$$I(t, c_2) = \log_2\frac{p(t\wedge c_2)}{p(t)\times p(c_2)} = \log_2\frac{0.5}{0.5001\times 0.5} = 0.9997.$$

$$I_{\text{avg}}(t)=\sum_{i=1}^{2}p(c_i)I(t,c_i)=$$

$0.5\times(-11.288)+0.5\times 0.9997=-5.144.$

$P(t\wedge c_1)=0.0001$ 意味着几乎每个 t 出现的文档都属于 c_2, 并且几乎所有没有出现 t 的文档都属于 c_1, 即 $\boldsymbol{C}$ 的值几乎由 $\boldsymbol{T}$ 的值决定, 但是 $I_{\text{avg}}(t)$ 的值却比较小.

另外, MI 偏向于低频词汇[2] (详细的实验结果可参见本文图 4), 低频词汇的 MI 函数值要高于很多高频词汇. 所以它不满足"不倾向于低频词的 TCC".

3.3　文档频率的分析

文档频率方法没有利用类别信息, 所以它不完全满足 TCC.

例如, 使用四元组 $\langle U, A, V, g\rangle$ 表示文本分类信息系统, $U=\{D_1,\cdots,D_n\}$ 为文档的集合, $A=T\cup C$, $T=\{T_1,\cdots,T_k\}$ 为特征的集合, C 为文档的类别. V 为 T_i 的取值范围. $V=\{0,1\}$, 定义的映射函数 g, $U\to V$:

$$g(D_i)=\begin{cases}0, & t\ \text{dosen}'\text{t occurs in}\ D_i,\\ 1, & t\ \text{occurs in}\ D_i.\end{cases}$$

表 1 给出一个文本分类信息表的例子, 表的行 $D_1, D_2, \cdots, D_6$ 代表每一个文档, 表的列 T_1, T_2, T_3 代表词条, C 代表文档的类别.

Table 1　An Information Table

表 1　文档决策表

D(document)	T_1	T_2	T_3	D(category)
D_1	1	1	1	C_1
D_2	0	1	1	C_2
D_3	0	0	1	C_2
D_4	0	0	0	C_2
D_5	0	0	0	C_2
D_6	0	0	0	C_2

在这个例子里, C 的值完全由 T_1 的值所决定, 根据 TCC, $f(\boldsymbol{C}, \boldsymbol{T})$应该是最大的, 但是根据文档频率 T_3 的值却是最大的, 所以 DF 不完全满足 TCC.

然而, 在现实中, 存在大量的文档, 而且大都数文本分类是多类问题, 每一个类中的文档又都多于一个. Yang 和 Pedersen 发现词条的 DF 值和 IG 值有很强的统计联系[2], 在我们的实验中也发现了这个现象.

所以, 虽然文档频率方法没有利用类别信息, 但在大规模语料集中统计规律符合 TCC.

另外, $f(C, T)$不倾向于低频词, 所以 DF 有很好的效果, 但是比完全符合 TCC 的 IG 差, IG 是最好的特征选择算法之一.

4　实验分析

已有许多统计分类和机器学习技术应用于文本分类中, 我们用其中的两种算法 k-近邻法(kNN)和朴素贝叶斯(Naïve Bayes)方法. 选择 kNN 是因为它是性能较好的分类器[11], 选择 NaïveBayes 方法是因为它是最有效的启发学习算法之一[12].

根据文献[13], 微平均精确率(micro-averaging precision)被广泛用于交叉验证比较. 这里我们用它来比较不同的特征选择算法的效果.

4.1　语料集

实验中我们用 Reuters-21578[2,14] 和 OHSUMED 两个语料集[2]. 对于 Reuters-21578 我们只使用一个类别, 而且每个类别至少有 5 个文档. 这样, 训练集有 5273 篇文档, 测试集有 1767 篇文档, 总共有 29 类满足我们的条件, 经过停用词移除、词干还原等处理后, 有 13961 个词汇.

OHSUMED 是一个医学预料库, 共有 1800 个类别, 14321 个有标题的文档. 实验中我们用这个语料集的一个子集, 共有 7445 篇文档作为训练集, 3279 篇文档作为测试集. 在训练集中共有 11465 个词条和 10 个类别.

4.2　实验结果

图 1 和图 2 分别表示 DF, IG, MI 在 Reuters-21578 语料集上用 kNN 和 NaïveBayes 分类器分类的实验效果, 从图中可以看出, IG 是效果最好的, DF 和它相近, 但稍差, MI 的效果最差.

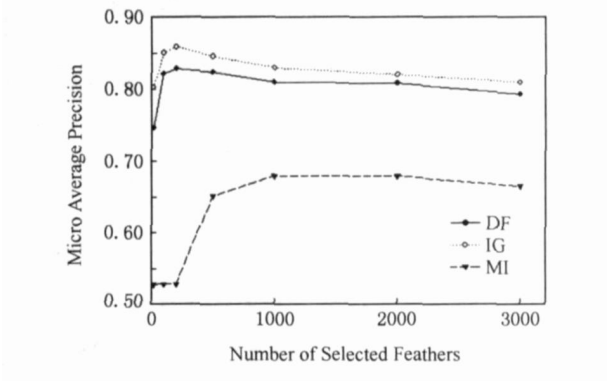

Fig. 1　Average precision of kNN vs. number of selected features on Reauters-21578.

图 1　在语料 Reuters-21578 上使用 kNN 分类器的微平均精确率的性能比较

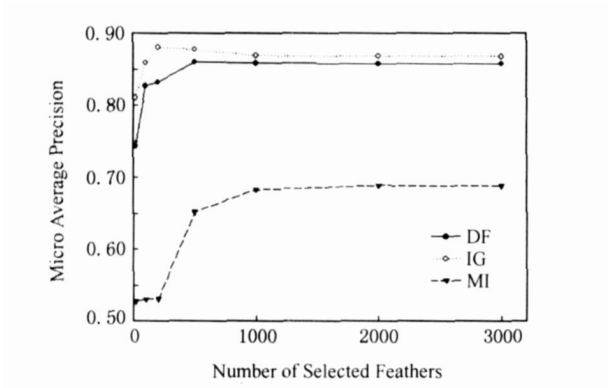

Fig. 2 Average precision of Naïve Bayes vs. number of selected features on Reauters-21578.

图2 在语料 Reuters-21578 上使用贝叶斯分类器的微平均精确率的性能比较

图3显示的是 Reuters 中每个词条的 DF 和 IG 比较的值. 很明显 DF 和 IG 的确有很紧密的统计联系,即 IG 值是大的,DF 的值也大,反之亦然,IG 满足 TCC,可见满足 TCC 的倾向于高频词,DF 在统计上基本符合 TCC.

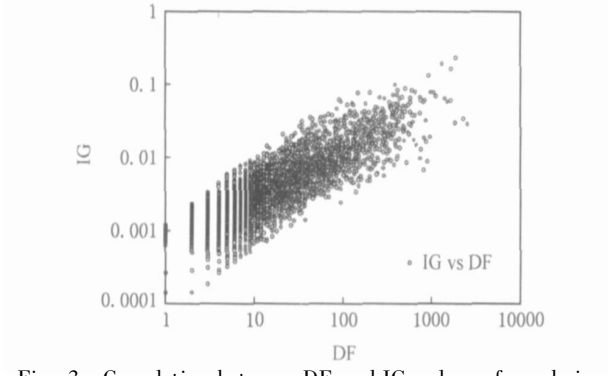

Fig. 3 Correlation between DF and IG values of words in Reuters-21578.

图3 在语料 Reuters-21578 上一个词条的 DF 值与 IG 值之间的关系

图4显示的是 DF 的值与 MI 的值的比较,很明显 MI 有很多负值,很多低频次有很高的 MI 值,而很多高频词的 MI 值是负值. MI 完全不符合 TCC.

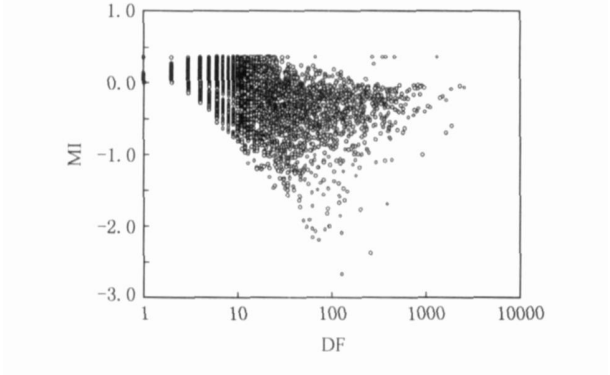

Fig. 4 Correlation between DF and MI values of words in Reuters-21578.

图4 在语料 Reuters-21578 上一个词条的 DF 值与 MI 值之间的关系

图5和图6表示 DF, IG, MI 在 OHSUMED 语料集上用 kNN 和 Naïve Bayes 分类器分类的实验效果,可以看到和图1、图2相同的结果

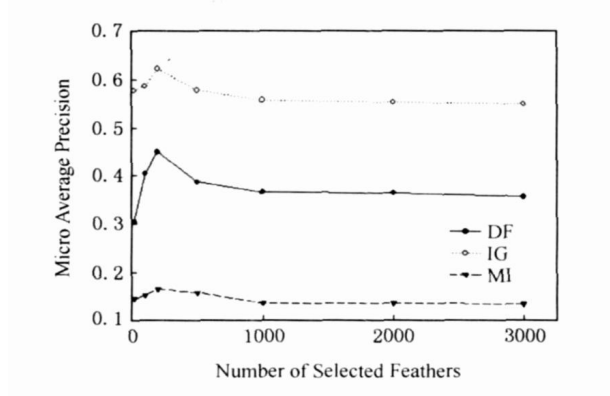

Fig. 5 Average precision of kNN vs. number of selected features on OHSUMED.

图5 在语料 OHSUMED 上使用 kNN 分类器的微平均精确率的性能比较

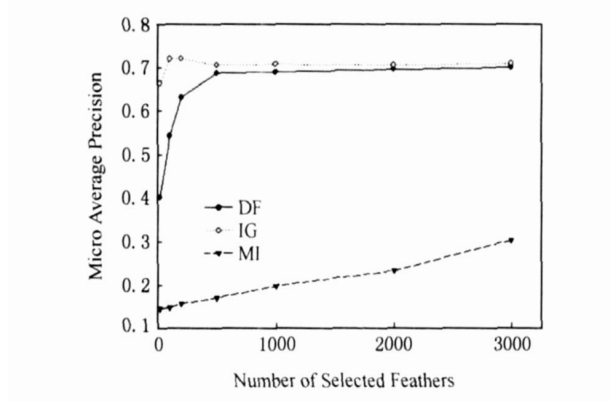

Fig. 6 Average precision of Naïve Bayes vs. number of selected features on OHSUMED.

图6 在语料 OHSUMED 上使用贝叶斯分类器的微平均精确率的性能比较

5 结 论

因为文本分类是一个分类问题,所以类别信息对于特征选择是很重要的. 本文定义了一组关于类别信息的基本约束 TCC,由分析可知 IG 完全满足 TCC,DF 在统计上基本满足 TCC,MI 完全不满足 TCC,因而 DF 也应该有较好的效果. 但是它比 IG 的效果差,IG 是性能最好的特征选择算法之一,MI 的性能相对较差,而我们的实验也验证了这点.

我们的下一步工作是构造满足约束 TCC 的一

组函数,分析这些函数的特点,进一步精确刻画随着 C 对 T 的依赖程度的增加,$f(C, T)$的值增加,从而寻找更优的特征选择函数

参 考 文 献

[1] Shang Wenqian, Huang Houkuan, Liu Yuling, *et al*. Research on the algorithm of feature selection based on gini index for text categorization[J]. Journal of Computer Research and Development, 2006, 43(10): 1688-1694 (in Chinese)
(尚文倩, 黄厚宽, 刘玉玲, 等. 文本分类中基于基尼指数的特征选择算法研究[J]. 计算机研究与发展, 2006, 43(10): 1688-1694)

[2] Y Yang, J O Pedersen. A comparative study on feature selection in text categorization[C]. In: D H Fisher, ed. Proc of the 14th Int'l Conf on Machine Learning (ICML-97). San Franicisco: Morgan Kaufmann, 1997. 412-420

[3] Shan Songwei, Feng Shicong, Li Xiaoming. A comparative study on several typical feature selection methods for Chinese Web page categorization[J]. Journal of the Computer Engineering and Application, 2003, 39(22): 146-148 (in Chinese)
(单松巍, 冯是聪, 李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用, 2003, 39(22): 146-148)

[4] Ying Liu. A comparative study on feature selection methods for drug discovery[J]. Chemical Information Computer Science, 2004, 44: 1823-1828

[5] Stewart M Yang, Xiao-Bin Wu, Zhi-Hong Deng, *et al*. Modification of feature selection methods using relative term frequency[C]. ICMLC-2002, Beijing, 2002

[6] J R Quinlan. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106

[7] Fabrizio Sebastiani. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47

[8] Kenneth Ward Church, Patrick Hanks. Word association norms, mutual information and lexicography[C]. The 27th Annual Meeting on Association for Computational Linguistics (ACL 27), Vancouver, Canada, 1989

[9] S R S Varadhan. Probability Theory[M]. New York: New York University Publisher, 2000

[10] Andrew Moore. Statistical Data Mining Tutorials[OL]. http://www.autonlab.org/tutorials/, 2006-06-16

[11] Y Yang, X Liu. A re-examination of text categorization methods[C]. SIGIR'99, Berkeley, 1999

[12] H Zhang. The optimality of naive Bayes[C]. The 17th Int'l FLAIRS Conference, Miami Beach, 2004

[13] Yiming Yang. An evaluation of statistical approaches to text categorization[J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88

[14] D David Lewis. Reuters-21578 test collection[OL]. http://www.daviddlewis.com/resources/testcollections/reuters21578/, 2004-05-14/2007-02-04

Xu Yan, born in 1968. Received her M. S. degree and Ph. D. degree in computer science from Beijing University of Aeronautics & Astronautics, Beijing, China. Her main research interests include data mining and information retrieval.
徐　燕, 1968年生, 博士, 主要研究方向为数据挖掘和信息检索

Li Jintao, born in 1962. Ph. D. Member of the IEEE Computer Society, professor and Ph. D. Supervisor in the Institute of Computing Technology, the Chinese Academy of Sciences now. His research interests include information retrieval and digital technology.
李锦涛, 1962年生, 博士, 研究员, 博士生导师, IEEE 会员, 主要研究方向为跨媒体检索、数字化技术

Wang Bin, born in 1972. He received his bachelor degree in computer science in 1993 and then accomplished his master thesis in 1996 from Wuhan University. Received his Ph. D. degree in 1999 in the Institute of Computing Technology, the Chinese Academy of Sciences. His main research interests include information retrieval and natural language processing.
王　斌, 1972年生, 博士, 副研究员, 中国计算机学会高级会员, 主要研究方向为信息检索、自然语言处理

Sun Chunming, born in 1982. Master. His main research interests include information retrieval.
孙春明, 1982年生, 硕士, 主要研究方向为信息检索

Zhang Sen, born in 1983. Master. His main research interests include information retrieval and query analysis.
张　森, 1983年生, 硕士, 主要研究方向为信息检索和查询分析

Research Background

Text categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, text categorization has become a key technique for handling and organizing text data. One of the most important issues in TC is feature selection (FS), which is to select the features for TC from the available feature space. Many FS methods have been proposed for TC, including document frequency thresholding (DF), information gain (IG) and Mutual Information (MI). Some comparative experiments show IG to be one of the most effective methods, while DF also performs very well. However MI has relatively poor performance. All these conclusions are based principally on empirical studies. The theoretical reasons why these methods behave in this way remains an unanswered question. In this paper, we put forward a theoretical performance evaluation functions for FS methods in text categorization. We define some basic desirable constraints which regard category information. Analyses and experiments show that the empirical performance of a feature selection method is tightly related to how well it satisfies these constraints. This work is supported by The National Natural Science Fundamental Research Project of China (60473002, 60603094) and by The National Natural Science Fundamental Research Project of Beijing (4051004).

2008年全国软件与应用学术会议(NASAC2008)
征文通知

由中国计算机学会系统软件专业委员会和软件工程专业委员会联合主办、华南理工大学软件学院承办的"2008年全国软件与应用学术会议(NASAC2008)"将于2008年11月6日至8日在广州举行. 会议录用的论文将出版论文集并评选优秀学生论文,择优论文将推荐到核心学术刊物(EI检索源)发表. 欢迎踊跃投稿.

征文范围(但不限于下列内容)

① 需求工程 ② 构件技术与软件复用 ③ 面向对象与软件Agent
④ 软件体系结构与设计模式 ⑤ 软件开发方法及自动化 ⑥ 软件过程管理与改进
⑦ 软件质量、测试与验证 ⑧ 软件再工程 ⑨ 软件工具与环境
⑩ 软件理论与形式化方法 ⑪ 操作系统 ⑫ 软件中间件与应用集成
⑬ 分布式系统及应用 ⑭ 软件语言与编译 ⑮ 软件标准与规范
⑯ 软件技术教育 ⑰ 计算机应用软件

论文要求

① 论文必须未在杂志和会议上发表和录用过
② 论文篇幅限定6页(A4纸)内
③ 会议只接受电子文档PDF或PS格式提交论文. 论文格式的详细要求请访问会议网站http://www.scut.edu.cn/nasac2008/
④ 投稿方式:采用在线投稿:http://www.scut.edu.cn/nasac2008/

重要日期

论文征稿截止日期:2008年7月15日,论文录用通知日期:2008年8月15日

联系方式

联系人:曹晓叶、卢叶莉,广州五山华南理工大学软件学院(510640)

Tel:020-87114028, 020-39380208

NASAC 2008会议网址:http://www.scut.edu.cn/nasac2008/