ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.1, January 2008, pp.82−89
DOI: 10.3724/SP.J.1001.2008.00082

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于区分类别能力的高性能特征选择方法*

徐 燕[1,2+], 李锦涛[1], 王 斌[1], 孙春明[1,2]

[1](中国科学院 计算技术研究所,北京 100080)

[2](华北电力大学,北京 102206)

A Category Resolve Power-Based Feature Selection Method

XU Yan[1,2+], LI Jin-Tao[1], WANG Bin[1], SUN Chun-Ming[1,2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](North China Electric Power University, Beijing 102206, China)

+ Corresponding author: Phn: +86-10-82522199, Fax: +86-10-62600602, E-mail: xuyan@ict.ac.cn

Xu Y, Li JT, Wang B, Sun CM. A category resolve power-based feature selection method. *Journal of Software*, 2008,19(1):82−89. http://www.jos.org.cn/1000-9825/19/82.htm

Abstract: One of the most important issues in Text Categorization (TC) is Feature Selection (FS). Many FS methods have been put forward and widely used in TC field, such as Information Gain (IG), Document Frequency (DF) thresholding, Mutual Information (MI) and so on. Empirical studies show that IG is one of the most effective methods, DF performs similarly, in contrast, and MI had relatively poor performance. One basic research question is why these FS methods cause different performance. Many existing work answers this question based on empirical studies. This paper presents a formal study of FS based on category resolve power. First, two desirable constraints that any reasonable FS function should satisfy are defined, then a universal method for developing FS functions is presented, and a new FS function KG using this method is developed. Analysis shows that IG and KG (knowledge gain) satisfy this universal method. Experiments on Reuters-21578 collection, NewsGroup collection and OHSUMED collection show that KG and IG get the best performance, even KG performs better than the IG method in two collections. These experiments imply that the universal method is very effective and gives a formal evaluation criterion for FS method.

Key words: feature selection; text categorization; information retrieval

摘 要: 特征选择在文本分类中起着重要作用.文档频率(document frequency,简称 DF)、信息增益(information gain,简称 IG)和互信息(mutual information,简称 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.特征选择是选择部分最有区分类别能力的特征,为此,给出了两个特征选择函数需满足的基本约束条件,并提出了一种构造高性能特征选择的通用方法.依此方法构造了一个新的特征选择函数 KG(knowledge gain).分析发现,IG 和 KG 完全满足该构造方法,在 Reuters-21578,OHSUMED 和

* Supported by the National Natural Science Foundation of China under Grant Nos.60473002, 60603094 (国家自然科学基金); the Beijing Natural Science Foundation of China under Grant No.4051004 (北京市自然科学基金)

Received 2006-09-29; Accepted 2006-12-27

NewsGroup这3个语料集上的实验表明,IG和KG性能最好,在两个语料集上,KG甚至超过了IG.验证了提出的构造高性能特征选择函数方法的有效性,同时也在理论上给出了一个评价高性能特征选择算法的标准.

关键词: 特征选择;文本分类;信息检索

中图法分类号: TP181 **文献标识码:** A

文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术[1],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.随着可用电子文档的增长和在线信息的快速膨胀,文本分类技术已经成为处理和组织文本数据的关键技术之一.

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[2].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.为此,特征选择是文本分类的一个非常重要的步骤.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.现有的特征选择函数主要有文档频率(document frequency,简称DF)、信息增益(information gain,简称IG)、互信息(mutual information,简称MI)等等.

已有的实验结果表明[2−4]: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函数性能表现的差异呢?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.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[5].

SCIW(strong class information words)特征选择方法是一种选择带有强类别信息词的方法.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[6].文献[7]使用类别特征域的方法将每个类别中重要的特征提取出来作为重要的特征,取得了较好的实验效果.

由CTD,SCIW和类别特征域可见,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:使用高频词和利用类别信息.然而,这个结论均来自于实验分析,即都是基于经验的方法得到的,并未进行定性的分析.

在文本分类中,特征选择方法性能的评价也均是基于实验的.本文在已有经验的基础上,对特征选择函数进行定性地分析,具体思路如下:

既然文本分类是分类问题的一种,而特征选择是根据某种准则从原始特征中选择最有区分类别能力的特征,因此,如果一个词条 t 在文档中出现与否对该文档分类没有丝毫影响,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除去,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.

根据这个基本想法,本文给出了特征选择算法需要满足的基本条件,在这组基本条件的基础上,给出了一种构造高性能特征选择函数的通用方法,并且按照这种构造方法实际构造了一个特征选择函数KG(knowledge gain).分析表明,IG和KG完全满足该构造方法,在Reuters-21578[8],OHSUMED和NewsGroup这3个语料集上的实验表明,IG和KG性能最好,在两个语料集OHSUMED和NewsGroup上,KG甚至超过了IG.

本文第1节介绍3种常用的特征选择算法.第2节给出特征选择算法需要满足的基本约束条件,将类别对词条的依赖程度进行了形式化分析.第3节提出一种构造高性能特征选择函数的通用方法,并构造新的特征选择函数.第4节在3个通用语料集上进行实验,验证本文提出的方法的有效性.第5节进行总结.

1 常用特征选择方法

在这一节,我们对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之一[2−4].

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,它容易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.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 $\{c_i\}_{i=1}^m$ 为目标空间中的类别的集合,那么,词条 t 对类别的信息增益为

$$IG(t)=-\sum_{i=1}^{m}p(c_i)\log p(c_i)+\ p(t)\sum_{i=1}^{m}p(c_i\mid t)\log p(c_i\mid t)+p(\bar{t})\sum_{i=1}^{m}p(c_i\mid \bar{t})\log p(c_i\mid \bar{t}).$$

1.3 互信息

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

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

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

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

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

2 区分类别的能力

特征选择是根据某种准则从原始特征中选择部分最有区分类别能力的特征.所谓特征的区分类别的能力,直观上说,是通过一个特征在文档中出现与否来判定该文档的类别属性的能力.

首先介绍一些在本文中用到的符号:

t 表示一个词条在文档中出现;

$\bar{t}$ 表示一个词条在文档中不出现,$T=\{t,\bar{t}\}$;

c_i 表示一个类别;

$C=\{c_i\}_{i=1}^m$ 表示类别的集合;

类别分布 $p(C)=(p(c_1),p(c_2),\ldots,p(c_m))$,$p$ 表示概率;

用 $f(C,t)$表示 t 区分类别 C 的能力,即 f 表示特征选择函数.

2.1 基本约束条件

$f(C,t)$表示 t 区分类别 C 的能力,若将 C 和 T 看作两个离散的随机变量,如果 T 对 C 的值没有产生影响,那么,通常认为它们是独立的;反之亦然.由此,$f(C,t)$应该满足如下基本约束条件(term-category constrains,简称 TCC):

(1) 如果 T 和 C 独立,当且仅当 t 的类别区分能力最小,即 $f(C,t)$的值最小(通常为 0);

(2) 如果 C 的值完全取决于 T,当且仅当 t 的区分类别能力最大,即 $f(C,t)$的值最大(有关约束条件的详细内容,将另文加以介绍).

以 TCC 作为出发点,下面将对特征的区分类别的能力进行严格的定义和量化.

2.2 类别区分能力的两个极端情况

上节所述的独立是概率论中随机变量之间的独立性.根据文献[9],离散随机变量 C 和 T 相互独立,当且仅当它们满足如下公式(1)和公式(2):

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

$$P(C=c_i;T=t)=P(C=c_i)\times P(T=t) \tag{1}$$

$$P(C=c_i;T=\bar{t})=P(C=c_i)\times P(T=\bar{t}) \tag{2}$$

将 T 与 C 看成随机变量,由基本约束条件 TCC 可以得到类别区分能力的两个极端情况的定义.

定义 1. $T=\{t,\bar{t}\}$,$C=\{c_i\}_{i=1}^{m}$ 表示类别的集合,如果离散随机变量 C 和 T 相互独立,则称 t 的区分类别能力最小,记为 0.

根据定义,可以得到如下性质:

性质 1. $p(c_i)=p(c_i\mid t)=p(c_i\mid \bar{t})(0\le i\le m)$ 当且仅当 t 的类别区分能力为 0(离散随机变量 C 和 T 相互独立).

证明: $p(c_i)=p(c_i\mid t)=p(c_i\mid \bar{t})(0\le i\le m)$ 当且仅当公式(1)和公式(2)成立,

当且仅当离散随机变量 C 和 T 相互独立,即 t 的类别区分能力为 0. □

定义 2. 如果 C 的值完全取决于 T,t 的类别区分能力最大,那么 $f(C,t)$的值最大.

由 t 的出现与否就能判断文档的类别,这时有如下性质:

性质 2. 对于两类问题,即 $C=\{c_i\}_{i=1}^{2}$, $p(c_i\mid t)=0$且$p(c_i\mid \bar{t})=1$,或$p(c_i\mid t)=1$且$p(c_i\mid \bar{t})=0$ $(0\le i\le 2)$当且仅当离散随机变量 C 由 T 决定,t 的区分类别能力最大.

从区分类别能力的两个极端情况可以发现:t 的区分类别能力可以通过文档的类别分布 $p(C)=(p(c_1),p(c_2),\ldots,p(c_m))$,与文档在 t 出现与否的类别分布 $p(C|t)=(p(c_1|t),p(c_2|t),\ldots,p(c_m|t))$, $p(C\mid\bar{t})=(p(c_1\mid\bar{t}),p(c_2\mid\bar{t}),\ldots,p(c_m\mid\bar{t}))$ 的变化来刻画:没有变化的,说明 t 出现与否对文档的类别不产生影响,t 的类别区分能力为 0,最小;有较大变化的,说明 t 出现与否对文档的类别产生的影响较大,可通过 t 出现与不出现之间的变化推断而得到类别,这时,t 的区分类别能力较大.

下面,通过这个想法,对 t 的区分类别能力进行量化.

2.3 类别区分能力的量化.

通过上述定义可知,C 对 T 的依赖,可通过 C 的类别分布和词条 t 出现或不出现时的类别分布的变化来刻画,该“变化”是一个模糊的概念,需要进一步细化,通过一个关于类别分布的函数 $g(C)$来刻画“变化”.

定义 3. 文档的类别分布函数:若文档集的类别分布为 $p(C)=(p(c_1),p(c_2),\ldots,p(c_m))$,$g(C)$为 $p(c_1),p(c_2),\ldots,p(c_m)$ 的一个函数,则称 $g(C)$为一个关于类别分布的函数.

定义 4. 称 $p(t)g(C\mid t)+p(\bar{t})g(C\mid\bar{t})$ 为文档集在给定 T 时的一个条件类别分布值(条件类别分布).

定义 5. 一个类别分布值与给定 T 时的条件类别分布值之差为 $g(C)-(p(t)g(C\mid t)+p(\bar{t})g(C\mid\bar{t}))$,若它满足基本约束条件 TCC,则称其为 t 的区分类别的能力,记为 $f(C,t)$.

性质 3. $f(C,t)$不唯一,它依赖于类别分布的函数 $g(C)$.

3 高性能特征选择函数的设计与实现

在上节中,我们已经量化了 t 的区分类别的能力,本节在此量化基础上,提出高性能特征选择函数的设计方法,并构造高性能特征选择函数.

3.1 高性能特征选择算法的设计步骤

高性能特征选择算法的设计分以下几个主要步骤:

第 1 步:构造文档集 D 的一个类别分布函数 $g(C)$.

第 2 步:计算文档集 D 在 T 下的条件类别分布: $p(t)g(C\mid t)+p(\bar{t})g(C\mid\bar{t})$.

第 3 步:得到特征选择函数.

- 类别分布值与给定 T 时的条件类别分布值之差: $g(C)-(p(t)g(C|t)+p(\bar{t})g(C|\bar{t}))$.
- 若上式满足基本约束条件 TCC,则称其为 t 的区分类别的能力,即特征选择函数为 $f(C,t)$:

$$f(C,t)= g(C)-p(t)g(C|t)+p(\bar{t})g(C|\bar{t}).$$

例:若 $g(C)=-\sum_{i=1}^{m}p(c_i)\log(p(c_i))$, 则 $f(C,t)=IG(t)$,即 $IG(t)$满足该构造方法.

而 DF 和 $MI(t)$不满足该构造方法.

3.2 设计新的特征选择函数

本节设计两个函数:一个部分满足条件;一个完全满足上文提到的高性能特征选择方法的条件.

独立的离散随机变量 C 和 T 满足公式(1)和公式(2),按照这两个公式,构造一个函数如下:

$$IND(t)=\sum_{i=1}^{m}|P(C=c_i;T=t)-P(C=c_i)\times P(T=t)|+\sum_{i=1}^{m}|P(C=c_i;T=\bar{t})-P(C=c_i)\times P(T=\bar{t})|.$$

该函数满足:如果 T 和 C 独立,那么 $f(C,t)=IND(t)$的值最小(值为 0).它符合 $f(C,t)$需满足的两个基本约束条件中的一个.

下面,我们严格按照上文提到的高性能特征选择算法的设计步骤,设计一个特征选择函数:

第 1 步:设计类别分布函数 $g(C)$:

$C=\{c_i\}_{i=1}^{m}$ 表示类别的集合,C 的类别分布 $p(C)=(p(c_1),p(c_2),\ldots,p(c_m))$,定义 $g(C)$如下:

$$g(C)=\sum_{1\le i<j\le m}p(c_i)\times p(c_j).$$

第 2 步:计算条件分布,即在条件 T 下的类别分布:

$$p(t)\sum_{1\le i<j\le m}p(c_i|t)p(c_j|t)+p(\bar{t})\sum_{1\le i<j\le m}p(c_i|\bar{t})p(c_j|\bar{t}).$$

第 3 步:得到特征选择函数(记作 $KG(t)$):

- 计算分布与给定 T 时的条件分布之差:

$$\sum_{1\le i<j\le m}p(c_i)\times p(c_j)-\left(p(t)\sum_{1\le i<j\le m}p(c_i|t)p(c_j|t)+p(\bar{t})\sum_{1\le i<j\le m}p(c_i|\bar{t})p(c_j|\bar{t})\right).$$

- 上式满足基本约束条件 TCC,

$$KG(t)=\sum_{1\le i<j\le m}p(c_i)\times p(c_j)-\left(p(t)\sum_{1\le i<j\le m}p(c_i|t)p(c_j|t)+p(\bar{t})\sum_{1\le i<j\le m}p(c_i|\bar{t})p(c_j|\bar{t})\right)$$

为构造的特征选择函数.

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,我们采用其中的两种算法: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[10],我们选择 Naïve Bayes 方法是因为它是最有效的启发学习算法之一,分类效果也较好[11].根据文献[12],微平均精确率(microaveraging 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

为了实验结果的普遍性,实验中我们用 Reuters-21578[8],OHSUMED[2]和 NewsGroup[13]这 3 个国际上通用的语料集.

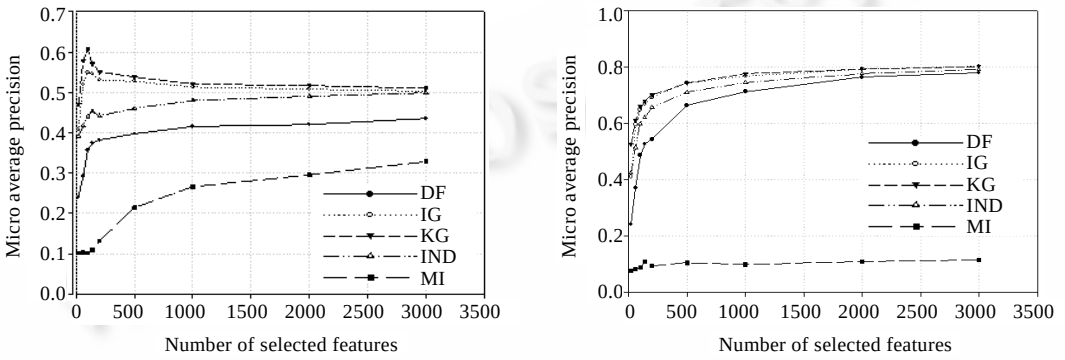
对于 Reuters-21578,我们使用只有 1 个类别,而且,每个类别至少有 5 个文档的文档.这样,训练集有 5 273 篇文档,测试集有 1 767 篇文档,总共有 29 类满足我们的条件.经过停用词移除、词干还原等处理后,有 13 961 个词汇.

OHSUMED 是一个医学语料库,共有 1 800 个类别,14 321 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 7 445 篇文档作为训练集,3 279 篇文档作为测试集.在训练集中共有 11 465 个词条和 10 个类别.

NewsGroup 语料集是由互联网用户在 Usenet 上张贴的 19 997 条消息组成的.这些消息均匀分布在 20 个不同的新闻组中,每个新闻组有 1 000 条消息,每个新闻组对应着一个文本类别.我们取其中的 10 个类别作为实验语料集,经过处理后,共有 31 109 个词条,6 162 篇文档作为训练集,3 838 篇文档作为测试集.

4.2 实验结果

图 1 表示 DF,IG,MI,KG 和 IND(independence)在 NewsGroup 语料集上分别用 *k*NN 和 Naïve Bayes 分类器的分类效果,从图中可以明显看出,IG 和 KG 是效果最好的,KG 甚至超过了 IG,IND 和 DF 稍差,IND 明显超过了 DF,MI 的效果最差,并且当减少到很低的维数时,它们之间的差距比较明显.

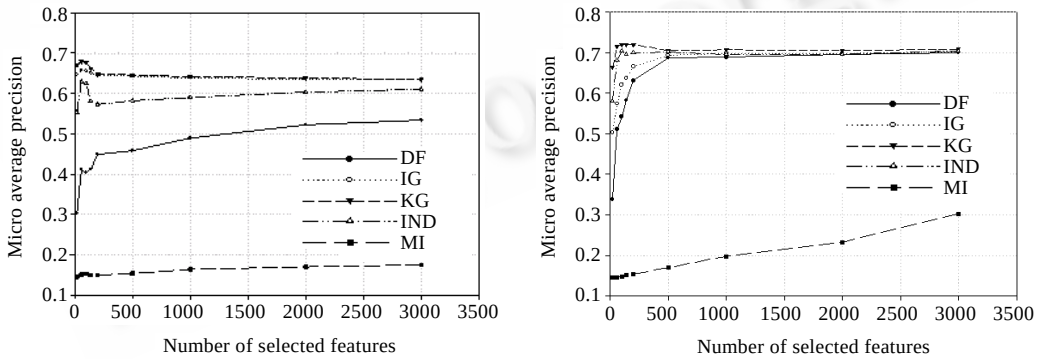


(a) *k*NN (b) Naïve Bayes

Fig.1 Average precision of *k*NN or Naïve Bayes vs. number of selected features on NewsGroup

图 1 在 NewsGroup 语料集上使用 *k*NN 或 Naïve Bayes 的平均精确率

图 2 表示 DF,IG,MI,KG 和 IND 在 OHSUMED 语料集上分别用 *k*NN 和 Naïve Bayes 分类器分类的效果,可以看到与图 1 相同的结果.



(a) *k*NN (b) Naïve Bayes

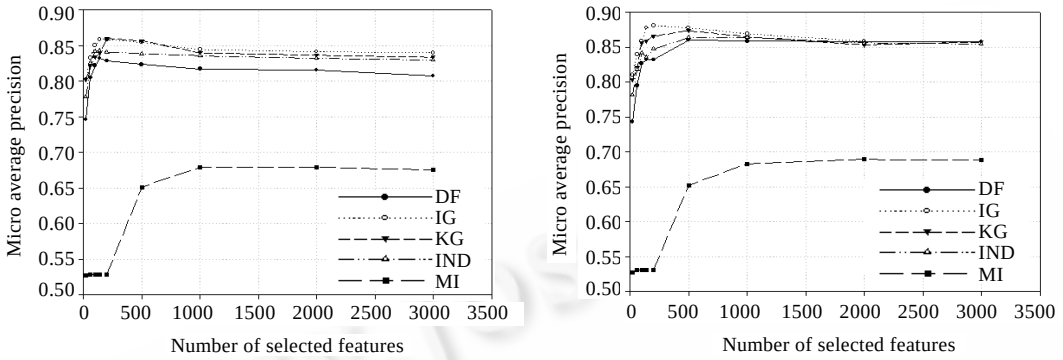
Fig.2 Average precision of *k*NN or Naïve Bayes vs. number of selected features on OHSUMED

图 1 在 OHSUMED 语料集上使用 *k*NN 或 Naïve Bayes 的平均精确率

图 3 表示 DF,IG,MI,KG 和 IND 在 Reuters-21578 语料集上分别用 *k*NN 和 Naïve Bayes 分类器分类的实验效果.从图中可以看出,IG 和 KG 是效果最好的(IG 比 KG 稍好),IND 和 DF 稍差,IND 比 DF 稍好,MI 的效果最差.在 3 个语料集上 IG 和 KG 均是效果最好的,在第 3 节的分析中我们发现,只有 IG 和 KG 满足我们的高性能特征

选择函数构造方法.实验验证了提出的构造高性能特征选择函数方法的有效性,同时也在理论上给出了一个评价高性能特征选择算法的标准.

信息增益 IG 和 KG 均完全满足我们的高性能特征选择算法的设计要求,在我们的实验中也验证了它们是性能最好的特征选择方法.我们构造的 IND 函数满足了约束条件 TCC 的独立性要求(满足一个约束),在实验中,它也有较好的性能,但比 IG 和 KG 要差.



(a) *k*NN (b) Naïve Bayes

Fig.3 Average precision of *k*NN or Naïve Bayes vs. number of selected features on Reuters-21578

图 3 在 Reuters-21578 语料集上使用 *k*NN 或 Naïve Bayes 的平均精确率

5 结 论

特征选择在文本分类中起着重要作用.特征选择是选择部分最有区分类别能力的特征,本文基于 T 和 C 的关系给出了特征的区分类别能力的一个解释,并由此提出了一种构造高性能特征选择的通用方法,依此方法构造了一个新的特征选择函数 KG.分析发现,IG 和 KG 完全满足该构造方法,在 Reuters-21578,OHSUMED 和 NewsGroup 这 3 个语料集上的实验表明:IG 和 KG 性能最好,在两个语料集上,KG 甚至超过了 IG.文章还验证了提出的构造高性能特征选择函数方法的有效性,同时也在理论上给出了一个评价高性能特征选择算法的标准.

References:

[1] Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. Journal of Software, 2006,17(9):1848−1859 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1848.htm

[2] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Fisher DH, ed. Proc. of the 14th Int'l Conf. on Machine Learning (ICML'97). Nashville: Morgan Kaufmann Publishers, 1997. 412−420.

[3] Yang SM, Wu XB, Deng ZH, Zhang M, Yang DQ. Relativeterm-Frequency based feature selection for text categorization. In: Proc. of the 1st Int'l Conf. of Machine Learning and Cybernetics (ICMLC 2002). Beijing, 2002. 1432−1436. http://www.icmlc.org/2002/

[4] Shan SW, Feng SC, Li XM. A comparative study on several typical feature selection methods for Chinese Web page categorization. Journal of the Computer Engineering and Application, 2003,39(22):146−148 (in Chinese with English abstract).

[5] Bong CH, Narayanan K. An empirical study of feature selection for text categorization based on term weightage. In: Proc. of the IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2004). Beijing: IEEE Computer Society Press, 2004. 599−602. http://www.comp.hkbu.edu.hk/IAT04/

[6] Li SS, Zong CQ. A new approach to feature selection for text categorization. In: Ren FJ, Zhong YX, eds. Proc. of the IEEE Int'l Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE). Wuhan: IEEE Press, 2005. 626−630.

[7] Zhao SQ, Zhang Y, Liu T, Chen YH, Huang YG, Li S. A feature selection method based on class feature domains for text categorization. The Journal of Chinese Information Processing, 2005,19(6):21−27 (in Chinese with English abstract).

[8] Reuters21578. 2004. http://www.daviddlewis.com/resources/testcollections/reuters21578/

[9] Varadhan SRS. Probability Theory. New York: Courant Institute of Mathematical Sciences Press, 2000. 1−167.

[10] Yang Y, Liu X. A re-examination of text categorization methods. In: Gey F, Hearst M, Rong R, eds. Proc. of the 22nd ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR'99). Berkeley: ACM Press, 1999. 42−49.

[11] Zhang H. The optimality of naive Bayes. In: Valerie B, Zdravko M, eds. Proc. of the 17th Int'l FLAIRS Conf., American Association for Artificial Intelligence. Miami Beach: AAAI Press, 2004. 562−567.

[12] Yang YM. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1999,1(1/2):67−88.

[13] NewsGroup. 1999. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html

附中文参考文献:

[1] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报,2006,17(9):1848−1859. http://www.jos.org.cn/1000-9825/17/1848.htm

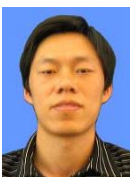
[4] 单松巍,冯是聪,李晓明.几种典型特征选取方法在中文网页分类上的效果比较.计算机工程与应用,2003,39(22):146−148.

[7] 赵世奇,张宇,刘挺,陈毅恒,黄永光,李生.基于类别特征域的文本分类特征选择方法.中文信息学报,2005,19(6):21−27.

徐燕(1968－),女,湖南大庸人,博士,副教授,主要研究领域为文本分类,信息检索,数据挖掘.

王斌(1972－),男,博士,副研究员,主要研究领域为信息检索.

李锦涛(1962－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为跨媒体检索,数字化技术.



孙春明(1982－),男,硕士,主要研究领域为文本分类,信息检索.