

文章编号:1003-0077(2004)03-0017-07

## 中文文本分类中的特征选择研究\*

周茜,赵明生,扈

(清华大学 电子工程系,北京 100084)

**摘要:**本文介绍和比较了八种用于文本分类的特征选择方法,其中把应用于二元分类器中的优势率改造成适用于多类问题的形式,并提出了一种新的类别区分词的特征选择方法,结合两种不同的分类方法:文本相似度方法和 Naïve Bayes 方法,在两个不同的数据集上分别作了训练和测试,结果表明,在这八种文本特征选择方法中,多类优势率和类别区分词方法取得了最好的选择效果。其中,当用 Naïve Bayes 分类方法对各类分布严重不均的 13890 样本集作训练和测试时,当特征维数大于 8000 以后,用类别区分词作特征选择得到的宏 F1 值比用 IG 作特征选择得到的宏 F1 值高出 3%~5% 左右。

**关键词:**计算机应用;中文信息处理;文本分类;特征选择;类别区分词

**中图分类号:**TP391 **文献标识码:**A

## Study on Feature Selection in Chinese Text Categorization

ZHOU Qian, ZHAO Ming-sheng, HU min

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

**Abstract:** This paper introduces and compares eight feature selection methods in text categorization. Among the eight methods, Multi-Class Odds Ratio (MC-OR), a variant of Odds Ratio which is often used in binary classification, and a new feature selection method based on Class-Discriminating Words (CDW) are proposed. Combined with the classic VSM classifier based on cosine similarity and the Naïve Bayes classifier, training and test are carried out on two text sets with different class distribution. As the results indicate, MC-OR and CDW gain the best selecting effect.

**Key words:** computer application; Chinese information processing; text categorization; feature selection; class-discriminating words

## 1 引言

文本分类 (Text Categorization) 是指依据文本的内容,由计算机根据某种自动分类算法,把文本判分为预先定义好的类别。文本分类是信息存储和信息检索中的重要课题。互联网的飞速发展又给文本分类提供了新的应用平台。网页分类是文本分类在网页文本集合上的应用,它在信息过滤、基于个性化的信息服务等方面有着重要用途。网页自动分类具有如下优点:不需要人工干预,节省大量人力物力,更新快,而且分类速度较快,精度较高,满足实际应用要求。

文本分类大致可分为三个步骤:文本的向量模型表示,文本特征选择和分类器训练。数量巨大的训练样本和过高的向量维数是文本分类的两大特点。这两个特性决定了文本分类问题

\* 收稿日期:2003-11-03

基金项目:国家自然科学基金资助项目 (60003014;60171037)

作者简介:周茜 (1980—),女,硕士研究生,研究方向为信息检索、中文信息处理。

是一个运算时间和空间复杂度很高的学习问题。为了兼顾运算时间和分类精度两个方面,我们不得不进行特征选择,力求在不损伤分类性能的同时达到降维的目的。

在文本分类中,常用的特征选择方法有基于阈值的统计方法,如文档频率方法(DF)<sup>[2]</sup>,信息增益方法(IG)<sup>[2]</sup>,互信息方法(MI)<sup>[2]</sup>,CHI<sup>[2]</sup>方法,期望交叉熵<sup>[3]</sup>,文本证据权<sup>[3]</sup>,优势率<sup>[3]</sup>,基于词频覆盖度<sup>[4]</sup>的特征选择方法等,以及由原始的低级特征(比如词)经过某种变换构建正交空间中的新特征的方法,如主分量分析<sup>[5]</sup>的方法等。基于阈值的统计方法具有计算复杂度低,速度快的优点,尤其适合做文本分类中的特征选择,在本文中我们将集中研究和比较8种基于阈值的统计方法。关于文本分类中的特征选择问题,比较有代表性的是 Yang Yiming<sup>[2]</sup>和 Dunja Mladenic<sup>[3]</sup>的工作。前者针对平面文本分类问题,分析和比较了DF,IG,MI和CHI等5种方法,结合LLSF和KNN分类器,得出IG和CHI方法效果相对较好的结论。而后者针对等级文本分类问题,分析和比较了信息增益,期望交叉熵,文本证据权及优势率等方法,结合Naïve Bayes分类器,实验结果表明二元优势率是最好的选择方法。由于两者针对不同类型的分类问题,不同的实验数据集,采用不同的分类器,因此得出了不太一致的实验结论。为了综合研究各种特征选择方法的选择性能,在相同的平面文本分类问题上对各种选择方法进行实验和比较是必要的。

在分类算法的选择上,目前存在各种各样的文本分类算法,如文本相似度法<sup>[6]</sup>(也称向量空间法),Naïve Bayes方法<sup>[3,6]</sup>,K-最近邻算法<sup>[7]</sup>(K-Nearest Neighbor),Neural Network方法<sup>[8]</sup>,SVM方法<sup>[9]</sup>等。文本相似度方法和Naïve Bayes方法是应用最多的两种方法,它们具有分类机制简单,处理速度快的优点。

在前人工作基础上,我们不仅研究和比较了信息增益,期望交叉熵等六种常用的特征选择方法,而且把应用于二元分类器中的优势率改造成适用于多类问题的形式,并提出了一种新的类别区分词的特征选择方法。根据词的类间后验概率分布进行区分性定义,把某个词最可能出现的那一类和其他类别区分开来,这种区分性越大,那么该词就越可能是某一类的核心特征,由此选出那些强类别意义的分类特征。结合文本相似度方法和Naïve Bayes分类器,在两个类别样本分布不同的网页集上作训练和测试,结果表明:改造的多类别优势率和类别区分词的方法取得了最好的特征选择效果。

## 2 向量空间模型

在文本分类领域,最常用的文本表示模型是 G. Salton 在 1975 年提出的向量空间模型(Vector Space Model),其基本思想是把文本  $d_i$  看作向量空间中的一个  $n$  维向量  $(t_{i1}, w(t_{i1}), t_{i2}, w(t_{i2}), \dots, t_{in}, w(t_{in}))$ , 其中  $t_{i1}, t_{i2}, \dots, t_{in}$  为表示该文本的  $n$  个特征,  $w(t_{ik}), k = 1, 2, \dots, n$  是该文本对应第  $k$  个特征的权重,一般取为词频的函数。对于中文文本来说,由于词是语义的最小单位,因此一般选择词作为特征。各维特征通常表示成词频  $tf(t_k)$  和反文档频率  $idf(t_k)$  的函数,即有:  $w(t_{ik}) = tf(t_{ik}) \times idf(t_{ik})$ 。其中  $tf(t_{ik})$  表示词  $t_k$  在第  $i$  篇文档中出现的次数,而  $idf(t_{ik}) = \log(N / df(t_k))$ ,  $N$  为文档集中的全部文档数,而  $df(t_k)$  表示出现词  $t_k$  的文档数。为了计算方便,通常还要对向量进行归一化。

作为网页分类的第一步,我们对中文网页集进行基于词典的分词处理,由于所选用的通用词典共有 116921 个词条,因此把每个网页表示为 116921 维的原始向量。因为词典中的很多词在网页中不出现,该网页向量的很多维特征值为 0,即是说该向量极度稀疏。而且原始特征词中的很多词对分类毫无意义,甚至还会引入分类噪声,降低分类精度。比如“如果”“但是”这些

在文章中起结构作用的虚词,不表示实际意义,在每篇文章中出现概率大致相等,对分类来说是“平凡词”,应该从特征集中去掉。于是,在进行分类器训练之前,我们必须进行特征选择,选出那些对分类有帮助的词,从而大大压缩特征空间,为后续的分类节省运算时间和存储空间。

### 3 特征选择方法

常用的文本特征选择方法有:文档频率(DF)<sup>[2]</sup>、信息增益(IG)<sup>[2]</sup>、互信息(MI)<sup>[2]</sup>、 $\chi^2$ 统计量(CHI)<sup>[2]</sup>、期望交叉熵<sup>[3]</sup>、文本证据权<sup>[3]</sup>、优势率<sup>[3]</sup>等。这些方法的基本思想都是对每一个特征(在这里是中文词),计算某种统计度量值,然后设定一个阈值  $T$ ,把度量值小于  $T$  的那些特征过滤掉,剩下的即认为是有效特征。除了介绍 IG, MI, 期望交叉熵等经典的特征选择方法之外,这里还介绍了一种改造的优势率方法和一种新的类别区分词的选择方法。

对于特征词  $t$ ,各种选择标准的含义如下:

1) 文档频数(Document Frequency),即是特征  $t$  在文本集中出现的文档数。

2) 信息增益(Information Gain):

$$IG(t) = - \sum_{i=1}^m P(C_i) \log P(C_i) + P(t) \sum_{i=1}^m P(C_i | t) \log P(C_i | t) + P(\bar{t}) \sum_{i=1}^m P(C_i | \bar{t}) \log P(C_i | \bar{t})$$

3) 互信息(Mutual Information):  $MI(t) = \sum_{i=1}^m P(C_i) \log \frac{P(t | C_i)}{P(t)}$

4) CHI:  $\chi^2(t, C_i) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$ ,

$$X_{avg}^2 = \sum_{i=1}^m P(C_i) \chi^2(t, C_i)$$

其中  $A$  是特征  $t$  和第  $i$  类文档共同出现的次数,  $B$  是特征  $t$  出现而第  $i$  类文档不出现的次数,  $C$  是第  $i$  类文档出现而特征  $t$  不出现的次数,  $D$  是第  $i$  类文档和特征  $t$  都不出现的次数。

5) 期望交叉熵(Expected Cross Entropy):  $CE(t) = P(t) \sum_{i=1}^m P(C_i | t) \log \frac{P(C_i | t)}{P(C_i)}$

6) 文本证据权(Weight of Evidence for Text):

$$WET(t) = P(t) \sum_{i=1}^m P(C_i) \left| \log \frac{P(C_i | t) (1 - P(C_i))}{P(C_i) (1 - P(C_i | t))} \right|$$

在以上各式中,  $P(C_i)$  表示第  $i$  类文档在文档集中出现的概率,  $P(t)$  表示词  $t$  出现的概率,  $P(\bar{t}) = 1 - P(t)$  表示词  $t$  不出现的概率,  $P(C_i | t)$  表示在出现词  $t$  的情况下,文档属于第  $i$  类的概率。  $P(C_i | \bar{t})$  表示词  $t$  不出现时,文档属于第  $i$  类的概率。

7) 优势率(Odds Ratio)原本用于二元分类器,定义如下:

$$OR(t) = \log \frac{P(t | C_{pos}) (1 - P(t | C_{neg}))}{(1 - P(t | C_{pos})) P(t | C_{neg})}$$

其中:  $C_{pos}$  表示正例集的情况,  $C_{neg}$  表示负例集的情况。为了适用于多类别的情况,我们提出一种多类别优势率(Multi-Class Odds Ratio)的变体形式如下:

$$MC-OR(t) = \sum_{i=1}^m P(C_i) \times |OR(t, C_i)| = \sum_{i=1}^m P(C_i) \left| \log \frac{P(t | C_i) (1 - P(t | C_{else}))}{P(t | C_{else}) (1 - P(t | C_i))} \right|$$

其中,  $C_{else}$  表示除第  $i$  类外的所有类别,即把当前的第  $i$  类当作正例集,而把所有其他类别

合起来作为负例集,从而有  $P(t|C_{else}) = \frac{P(t) - P(t, C_i)}{1 - P(C_i)}$ 。

#### 8) 类别区分词 (Category - Discriminating Word)

上述这些选择方法有一个共同的特点:并不按类别计算统计值,选出的是那些全局意义上的“强类别意义”的词,这些词可能有着多类的指示意义。对于不兼类的文本分类问题来说,选用这些词作为分类特征,将使得某些文本向量位于两类的分界线附近,自动分类极易发生错误。于是我们可以发现这样一种现象,有些词的单类类别意义非常明显,比如“军舰”,“软着陆”,“阿拉法特”,“景泰蓝”等等,它们几乎就只出现在某一类文档之中。比如我们如果要把所有文档分为国际、环保、经济、军事、科教、生活、时政、文娱这八大类,那么“军舰”在文章中的出现就使我们有理由猜测该文章属于军事类,同理,出现“软着陆”的文档极有可能属于经济类。这些词有着极强的类别指示意义,类别区分性相当好,我们称之为“类别区分词”。我们猜测,如果根据词出现的统计信息,选出对应每类的“类别区分词”作为分类的特征表示,那么有可能在大大缩减特征空间的同时,选出那些最具类别指示意义因而也最利于分类的特征。

因此,我们设计“类别区分词”的选取方法如下:

首先,定义词  $t_1$  的类间概率分布如下:

$$Distribute(t) = (P(C_1 | t_1), P(C_2 | t_1), \dots, P(C_n | t_1))$$

其中,  $P(C_i | t_1) = \frac{P(t_1 | C_i) P(C_i)}{P(t_1)}$  为 Bayes 后验概率,  $P(t_1) = \sum_{i=1}^m P(C_i) P(t_1 | C_i)$

$$而, P(t_1 | C_i) = \frac{1 + \sum_{k=1}^{d_i} tf(t_{1k})}{|V| + \sum_{j=1}^{|V|} \sum_{k=1}^{d_j} tf(t_{jk})}, tf(t_{jk}) \text{ 表示词 } t_j \text{ 在 } C_i \text{ 类的第 } k \text{ 篇文档中出现的次数。}$$

$|V|$  为总词数,  $d_i$  表示  $C_i$  类的总文档数。

其次,定义区分词挑选标准  $CDW(t) = \text{Max1} - \text{Max2}$ , 其中  $\text{Max1}$  为  $P(C_i | t_1)$ ,  $i = 1, 2, \dots, m$  中的最大值,  $\text{Max2}$  为次大值。

最后,设置一个阈值  $T$ ,  $T$  为 0 到 1 之间的数,  $CDW(t)$  值大于  $T$  的那些词作为类别区分词被挑选出来。

这种做法的直观意义在于,用词与类别之间的后验概率来衡量该词的类别指示意义。而用后验概率最大值和次大值之间的差距来衡量该词的类间区分性。最大类别后验概率越大,与其余类别后验概率之间的差越大,那么该词关于某一类的指示意义就越强。 $T$  的取值一般在 0.3 以上 ( $T$  的取值与训练样本集有关),  $T$  越大,选出词的类别区分性越强,但这样的词也越少。由于训练样本集的数量和覆盖广度有限,这种方法选出的类别区分词不会很多。要求的类别区分性越强,这样的词总数越少,但特征太少,表征模式的能力就会大大下降。因此,我们需要在类别区分性和结果词的数量上做一个折衷。

## 4 文本分类的文本相似度方法和 Naïve Bayes 方法

### 1) 文本相似度法

文本相似度方法其实是一种基于样本相似度的质心分类法。根据待分类的测试样本  $d_i$  和各类类中心向量的余弦相似度,把该测试样本判分为相似度最大的那一类。即有:

$$C = \max_j \cos(d_i, V_j) = \frac{d_i \times V_j}{|d_i| |V_j|} = \frac{\prod_{l=1}^n w(t_{il}) w(t_{jl})}{\prod_{l=1}^n w(t_{il})^2 \prod_{l=1}^n w(t_{jl})^2}$$

## 2) Naïve Bayes 方法

Naïve Bayes 分类器的一个基本前提是各特征之间的独立性假设,即假定文本中各个特征项属于特定类别的概率相互独立。分类器通过计算待分类样本属于各类的后验概率,把该待分类样本判分为后验概率最大的那一类。Naïve Bayes 分类器的判分准则:

$$C = \max_i P(d | C_i) P(C_i) = \max_i P(C_i) \prod_k P(t_k | C_i)^{N(t_k, d)}$$

$$\text{其中 } P(t_k | C_i) = \frac{1 + \sum_{l=1}^{d_i} f(t_{kl})}{|V| + \sum_{j=1}^{d_i} f(t_{jl})}, f(t_{jl}) \text{ 表示词 } t_j \text{ 在 } C_i \text{ 类的第 } l \text{ 篇文档中出现的次数, } |V| \text{ 为总词数, } d_i \text{ 表示 } C_i \text{ 类的总文档数。} N(t_k, d) \text{ 表示词 } t_k \text{ 在文档 } d \text{ 中出现的次数。}$$

## 5 实验结果与结论

为了比较上述八种特征选择方法,我们在两个数据集上作了测试。数据集 1 包含八类共 16000 个网页的样本集,来自人民网([www.people.com.cn](http://www.people.com.cn)) 2001 年 1 月到 2003 年 1 月的新闻语料,涵盖国际、经济、军事、环保、科教、社会时政、生活、文娱八大类。各类样本数相等,均为 2000 个。训练集和测试集取 4:1 的比例。即训练集中有 12800 个样本,而测试集中有 3200 个样本。数据集 2 来自 2003 年 3 月在北京大学举办的“中文网页分类竞赛”中给出的训练网页集,共包含 11 类 13890 个网页,各类网页数从最少的 138 到最多的 2841 个,在各类中的网页数分别为:人文与艺术 496,新闻与媒体 138,商业与经济 1024,娱乐与休闲 1846,政府与政治 368,社会与文化 1353,教育 364,自然科学 2255,社会科学 2160,计算机与因特网 1045,医疗与健康 2841,网页在各类的分布极不均匀。训练集和测试集同样取 4:1 的比例。从分类结果我们可以看出,这种网页集类间分布不均的情况对最终的分类结果有一定影响。

为了综合考虑分类精度和召回率,全面衡量分类系统性能,我们使用宏 F1 值作为评价指标,计算如下:

$$\text{Macro-F1} = \frac{1}{m} \sum_{i=1}^m \frac{N_i}{N} \times F1_i = \frac{1}{m} \sum_{i=1}^m \frac{N_i}{N} \times \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

其中,  $N_i$  为第  $i$  类的测试文档数,  $N$  为测试文档总数。  $\text{precision}_i$  和  $\text{recall}_i$  分别为第  $i$  类的正确率和召回率,共有  $m$  个类别。

对于数据集 1,图 1 和图 2 分别给出了采用文本相似度方法和 Naïve Bayes 分类方法时,用上述 7 种特征选择方法(除了 Mutual Information 方法)选出 500 ~ 20000 维特征时,得到的相应分类性能。横坐标为特征维数 Features,纵坐标为宏 F1 值。

同样,在数据集 2 上用文本相似度方法和 Naïve Bayes 方法,用 7 种特征选择方法作特征选择的效果如图 3,图 4 所示。

在上述各图中没有出现应用 MI 方法进行特征选择的分类结果,是因为 MI 效果太差,在两个数据集上,用文本相似度方法和 Naïve Bayes 两种方法做分类的情况下,在特征维数低于 20000 维时,得到的宏 F1 值都不超过 60%。Yang Yiming 曾对此给出了解释,她认为这是由于 MI 方法在选择特征时,偏爱那些出现频率低的词<sup>[2]</sup>。为了重点比较其他 7 种特征选择方法的

效果,故在图中不再画出 MI 作特征选择的效果。

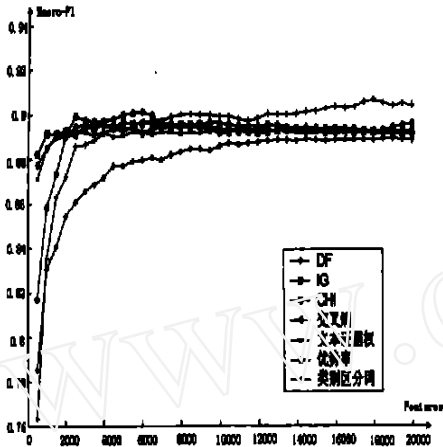


图 1 数据集 1,采用文本相似度方法

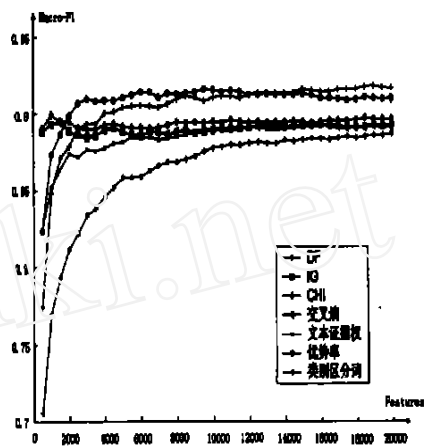


图 2 数据集 1,采用 Naïve Bayes 方法

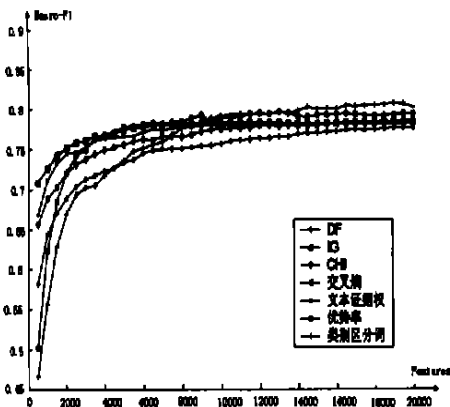


图 3 数据集 2,采用文本相似度方法

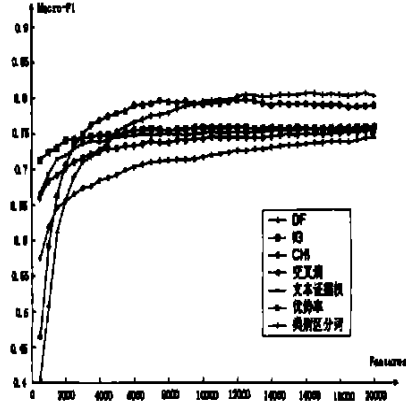


图 4 数据集 2,采用 Naïve Bayes 方法

由上述诸图,我们可以得出以下结论:

1) 改进的多类别优势率和类别区分词,这两种方法效果最好。在不同的数据集和不同的分类方法下,这两种方法的选择效果一直优于其他几种方法。比如在图 1 中,用文本相似度方法对 16000 网页集进行分类训练和测试时,当特征维数大于 6000 维,类别区分词作特征选择的效果最好。

2) 从实验结果来看,类别区分词和多类优势率的效果最好,IG 和期望交叉熵其次,文本证据权和 CHI 再次,DF 效果最差。由于只用到了特征词的文档频率信息,因此 DF 的选择效果最差也就不足为奇了。通常我们把 DF 方法作为比较的基准。

3) 当数据集是均匀分布时(如图 1 和图 2 所示的 16000 数据集),CHI 作特征选择的效果略优于 IG 和期望交叉熵,而当数据集的类别分布极为不均时(如图 3 和图 4 所示的 13890 数据集),当特征维数低于 10000 维时,IG 和期望交叉熵比 CHI 有着明显的优势,当特征维数高于 10000 维时,CHI,IG,期望交叉熵趋于相同的效果。

4) IG 和期望交叉熵的曲线基本重合,说明这两种方法做特征选择时,有着相似的效果。把 IG 公式重写为如下形式:

$$IG(t) = P(t) \prod_{i=1}^m P(C_i | t) \log \frac{P(C_i | t)}{P(C_i)} + P(\bar{t}) \prod_{i=1}^m P(C_i | \bar{t}) \log \frac{P(C_i | \bar{t})}{P(C_i)}$$

第一部分就是期望交叉熵,不同的是 IG 还考虑了特征不出现情况下的信息贡献,而从本次实验结果来看,这部分的贡献很小。

5) 在图 1 和图 2,图 3 和图 4 之间作比较,我们发现当采用文本相似度方法时,各种特征选择方法的效果相差不大,而当采用 Naïve Bayes 方法时,多类别优势率和类别区分词的方法比其他方法有着明显的优势。如图 4 所示,当用 Naïve Bayes 分类方法对各类分布严重不均的 13890 样本集作训练和测试时,当特征维数大于 10000 以后,用类别区分词作特征选择得到的宏 F1 值比用 IG 作特征选择得到的宏 F1 值高出 5 % 左右。

6) 类别分布不同的数据集,在采用相同分类机制和相同特征选择方法进行训练和测试时,有着不同的分类结果。比较图 1 和图 3,图 2 和图 4,我们发现各类样本均匀分布的 16000 样本集最高能达到 90 % 左右的宏 F1 值,而各类样本分布极为不均的 13890 样本集只能达到最高 80 % 左右的宏 F1 值。这说明样本集的选取对文本分类的绝对结果有着相当大的影响,但同时也注意到,一些相对的结果(如上述结论 1 - 5)在不同样本集的测试结果中仍成立。

## 6 结束语

本文介绍和比较了八种文本分类中的特征选择算法,其中把用于二元分类中的优势率改造成适用于多类问题的形式,并提出了一种类别区分词的特征选择方法。结合文本相似度方法和 Naïve Bayes 分类方法,在类别分布均匀和类别分布极度不均的两个数据集上作了训练和测试,结果表明多类优势率和类别区分词是两种有效的特征选择方法。下一步工作,我们将结合较复杂的分类方法如 SVM,以及 Boosting 和 Bagging 组合分类方法等,研究特征选择如何与这些分类方法更好的结合从而提高分类精度和效率,并研究其他基于 Wrapper 模型的特征选择方法在文本分类中的应用。

## 参 考 文 献:

- [1] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.
- [2] Yang Yiming, Pederson J O. A Comparative Study on Feature Selection in Text Categorization [A]. Proceedings of the 14th International Conference on Machine learning[C]. Nashville: Morgan Kaufmann, 1997: 412 - 420.
- [3] Mladenic D., Grobelnik M. Feature Selection for unbalanced class distribution and Naïve Bayes [A]. Proceedings of the Sixteenth International Conference on Machine Learning [C]. Bled: Morgan Kaufmann, 1999: 258 - 267.
- [4] 王梦云,曹素青. 基于字频向量的中文文本自动分类系统[J]. 情报学报,2000,19(6): 644 - 649.
- [5] Y. Yang. Noise reduction in a statistical approach to text categorization [A]. Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) [C]. Seattle: ACM Press, 1995: 256 - 263.
- [6] 范焱,郑诚,等. 用 Naïve Bayes 方法协调分类 Web 网页[J]. 软件学报,2001,12(9): 1386 - 1392.
- [7] 刘斌,黄铁军,程军,高文. 一种新的基于统计的自动文本分类方法[J]. 中文信息学报,2002,16(6): 18 - 24.
- [8] 梁久祯,兰东俊,唐. 基于先验知识的网页特征压缩与线性分类器设计 [A]. 第十二届全国神经计算学术大会论文集 [C]. 北京:人民邮电出版社,2002,494 - 501.
- [9] Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features [A], In: European Conference on Machine Learning (ECML) [C]. Berlin: Springer, 1998,137 - 142.