

# 文本分类中基于对数似然比 测试的特征词选择方法\*

李国臣

山西大学计算机科学系 太原 030006

**摘要** 本文将对数似然比测试用于文本分类中的特征词选择。与传统的频度、集中度和分散度等多种统计指标的测试独立进行的方法相比较,这种方法利用协方差矩阵协调了各个统计指标之间的联系,从而将它们有机地统一为一个整体。实验显示,这种特征词选择方法优于传统的频度测试、集中度测试和分散度测试独立进行的特征词选择的方法。

**关键词** 文本分类 特征词选择 对数似然比测试

## A Log-Likelihood-Ratio-Test-Based Feature Word Selection Approach in Text Categorization

Li Guochen

Department of Computer Science Shanxi University Taiyuan 030006

Email: ligc@deer.sxu.edu.cn

**Abstract** The paper uses the Log-Likelihood-Ratio-Test-Based feature words selection approach in the field of text categorization. In comparison with the traditional method, that is, each of the frequency test, salience test and distributioness test is conducted independently, the proposed approach uses covariance matrix to coordinate the associations among the variant statistics so that all of them are integrated into a whole. The experiments show that the approach is superior to the traditional approach.

**Key words** Text categorization Feature Selection Log Likelihood Ratio Test

### 一、引言

随着电子文本的大量涌现,电子文本的自动处理显得日益迫切。作为信息检索的一个辅助技术,文本分类的研究成为一个重要的研究课题。文本自动分类的研究主要集中在两个方面。第一个方面是特征单元的确定问题,即基于什么类型的特征(如字、字串、词等)去分类。

\* 本文于 1998 年 9 月 24 日收到

第二个方面是特征选择问题,即在特征单元确定后如何去选择这些特征。英文文本自动分类的研究主要有[1,4,5]。汉语文本的自动分类研究如下。关于特征单元的确定问题,最初的汉语文本自动分类是基于字的,[7]研究了基于二元同现字串的文本分类技术,试验结果表明基于二元同现汉字字串的分类方法的准确率高于基于字的分类方法。随着汉语自动分词技术的日益成熟,基于词的汉语文本自动分类成为一个研究热点。[10]的研究表明,基于词的分类方法优于基于字和基于二元同现字串的分类方法,基于词次(Word Times)分类方法优于基于词形(Word Tokens)的分类方法。关于特征选择问题,传统的特征选择方法是基于频度的,[8]提出了基于集中度和分布度的特征选择方法。以上的特征选择方法均从单一或片面的测试指标出发选择特征,选出的特征能够高度符合该测试指标,然而不能满足其他测试指标的要求(称为特征的“过度拟合”问题),所以利用这些特征选择方法选出的特征并不是一般意义上的特征。如何将以上各种测试指标有机地结合起来从而在选择特征时能够综合考虑各种测试指标是一个有意义的研究课题。针对这个问题,本文提出了一种基于对数似然比测试的特征选择方法,该方法在进行特征选择时将频度、集中度和广度等几种统计测试指标用对数似然比测试统一起来,使得选出的特征能够在频度、集中度和广度等方面达到整体测试的最优,从而避免了基于单一或片面测试指标所造成的单个特征的“过度拟合”问题。本文以词为特征单元,研究基于对数似然比测试的特征词选择方法。第2节介绍文本分类的基本概念;第3节详细介绍基于对数似然比测试的特征词选择方法;第4节介绍相关的特征词选择实验。

## 二、文本分类的向量空间模型

文本分类可以描述为这样一个问题:对于每个新到的电子文本,计算机自动判断它与系统规定的各个文本类别之间的相关性,从而给每个新到的文本指派一个类别。向量空间模型是文本分类研究最常用的模型,在这个模型中,首先将每个类别和具体的文本分别抽象为特征(如:字、字串、词等)向量,然后通过计算文本向量和类别向量之间的相关性对文本进行分类。这个过程可分为以下三个子问题进行,即类别学习、文本标引和相关性计算。

### 2.1 类别学习和文本标引

给定经过人工分类的文本集合  $C_1, C_2, \dots, C_n$  作为训练集,通过一个或一组可以反映词在一个类别  $C_i$  中的分布情况的统计量(如:频度、集中度、分布度等)来选择该类别的局部特征词集合  $W_i = \{w_{i1}, \dots, w_{im}, \dots, w_{il_i}\}$ ,其中  $w_{im}$  为第  $m$  个特征词,  $l_i$  表示该类别的特征词数。所有类别的局部特征词集合的并集  $W_1 \cup W_2 \cup \dots \cup W_n$  构成文本全集的全局特征词集合  $W$ 。将每个类别  $C_i$  映射到  $W$  上,并加权,就构成类别  $C_i$  的特征词向量(也称类别向量)  $C_i = (t_{i1}, t_{i2}, \dots, t_{ik}, \dots, t_{is})$ ,其中  $t_{ik}$  表示全局特征词  $w_k$  在类别  $C_i$  中的权重,  $S$  是全局特征词集合的规模;将文本  $d_j$  映射到  $W$  中,并加权,就构成文本  $d_j$  的特征词向量(也称文本向量)  $d_j = (t_{j1}, t_{j2}, \dots, t_{jk}, \dots, t_{js})$ 。其中  $t_{jk}$  表示全局特征词  $w_k$  在文本  $d_j$  中的权重。类别特征词的选择和在全局特征词向量空间中的表示称为类别学习,文本特征词在全局特征词向量空间中的表示称为文本标引。

### 2.2 相关性计算

设类别  $C_i$  的特征词向量为  $C_i = (t_{i1}, t_{i2}, \dots, t_{ik}, \dots, t_{is})$ ,文本  $d_j$  的特征词向量为  $d_j = (t_{j1}, t_{j2}, \dots, t_{jk}, \dots, t_{js})$ ,相关性计算可以判断文本  $d_j$  与类别  $C_i$  在多大程度上相关,从而确定文本  $d_j$  的类别归属。文本  $d_j$  与类别  $C_i$  的相关性可以通过二者之间的相似性来衡量。其中,

相似性用特征词向量空间中向量  $d_j$  和向量  $C_i$  之间夹角的余弦表示如下：

$$Sim(C_i, d_j) = \frac{\sum_{k=1}^{l_j} t_{ik} \cdot t_{jk}}{\sqrt{\sum_{k=1}^{l_j} t_{ik}^2 \cdot \sum_{k=1}^{l_j} t_{jk}^2}}$$

通过相关性计算,可确定文本  $d_j$  的类别为在特征词向量空间中与文本向量的相似度最大的那个类别  $C_m$ ,即:

$$C_m = \arg \max_i Sim(C_i, d_j)$$

从以上文本分类的基本概念可知,只有对各类文本进行了特征词选择之后,被选择的特征词才能构成类别向量,在此基础上,才可能进一步构造特征词向量空间以及进行文本与类别的相关性计算,从而确定文本的类别。因此,特征词选择是文本分类中的一个关键问题,以下将详细讨论这个问题。

### 三、基于对数似然比测试的特征词选择方法

#### 3.1 基于对数似然比测试的特征词选择方法的提出

特征词选择依赖于多种测试指标,最常用的测试指标包括频度、集中度和分散度等<sup>[1,8]</sup>。频度是最常用的特征选择测试指标。采用频度指标的特征词选择方法认为,在某一类文本中出现次数越多的词越能代表这类文本,因此选择在同一类文本中出现频度最高的若干词作为该类文本的特征词;采用集中度指标的特征词选择方法认为,一个有标引价值的词,应该集中出现在某一类文本中,而不是均匀地分布在各类文本中;采用分散度指标的特征词选择方法认为,在某类文本中均匀出现的词对该类文本应具有较高的标引价值;相反,如果某个词只集中出现在该类别的个别文本中,而在该类别的其他文本中很少出现,则该词的标引价值相对就要小多了。

以上介绍的测试指标分别从各个不同的侧面对特征词的选择进行限制。如果能够在进行特征词选择时全面考虑这些测试指标,则可以使选出的特征词达到整体测试的最优,从而对于文本分类具有更好的鉴别作用。这是信息理论的观点。然而,到目前为止,综合利用多种测试指标的特征词选择方法还很不成熟。一般的特征词选择方法将这几种测试指标单独使用或用一些经验的阈值简单地统一在一起。例如,测试方法可能是:如果某个词  $w_i$  在类别  $C_j$  中的频度 大于阈值  $T$ ,集中度 大于阈值  $T$ ,分散度 大于阈值  $T$ ,则把  $w_i$  选为  $C_j$  的特征词(本文将这种依次对各种测试指标进行限制的方法称为“串行”的特征词选择方法)。由于各种测试指标的判别是独立进行的,因此这种“串行”的特征词选择方法存在以下问题:(1)不同测试指标的协调问题。例如,如果首先用频度过滤掉一些低频词,然后再从剩余的词中过滤掉一些集中度较小的词,则一些集中度较高而频度较低的词可能没有选为特征词,而这些词对于分类可能有较大的作用;(2)不同阈值的一致性。这种选择方法的阈值都是经验启发而来,对于多个测试指标的不同阈值确定之间很难保持一种平衡,例如需要同时调整频度阈值  $T$ 、集中度阈值  $T$  以及分散度阈值  $T$  以使选出的特征词数满足期望的数量级。

针对以上的问题,本文提出将对数似然比测试应用于特征词选择的方法。

#### 3.2 基于对数似然比测试的特征词选择方法

设对于类别  $C_j$ ,词  $w_i$  的频度、集中度和分散度等多种统计测试指标构成向量  $V_{ij} = ($

, ) , 其中 、 、 分别表示  $w_i$  在  $C_j$  中的频度、集中度和分散度, 所谓特征词选择就是根据  $V_{ij}$  来判别  $w_i$  是否为  $C_j$  的特征词。这个问题可以看作是一个分类问题, 即判断  $w_i$  是属于类别  $C_j$  (表示  $w_i$  是  $C_j$  的特征词) 还是属于类别  $\overline{C_j}$  (表示  $w_i$  不是  $C_j$  的特征词)。

本文利用如下的对数似然比测试来判别  $w_i$  是否是类别  $C_j$  的特征词。

$$LLR = \log \frac{p(C_j | V_{ij})}{p(\overline{C_j} | V_{ij})} = \log \frac{p(V_{ij} | C_j) p(C_j)}{p(V_{ij} | \overline{C_j}) p(\overline{C_j})} \quad (1)$$

其中  $LLR$  是词  $w_i$  对于类别  $C_j$  的对数似然比,  $p(V_{ij} | C_j)$  和  $p(V_{ij} | \overline{C_j})$  分别是  $V_{ij}$  在类别  $C_j$  和类别  $\overline{C_j}$  的补集中的密度函数,  $p(C_j)$  和  $p(\overline{C_j})$  分别是类别  $C_j$  和类别  $\overline{C_j}$  的先验概率。如果  $LLR > 0$  (或大于某个阈值), 则判断它为  $C_j$  的特征词。

在给定训练集的条件下, 先验概率  $p(C_j)$  和  $p(\overline{C_j})$  的估计容易实现, 这里重点介绍联合概率密度函数  $p(V_{ij} | C_j)$  和  $p(V_{ij} | \overline{C_j})$  的估计问题。本文将对数似然比测试中的联合概率密度函数  $p(V_{ij} | C_j)$  和  $p(V_{ij} | \overline{C_j})$  看作是多元正态分布的复合, 即

$$p(V_{ij} | C_j) = \prod_{m=1}^M (2^{-D/2} | \Sigma_m |^{-1/2} \exp[-\frac{1}{2} (V_{ij} - u_m)^T \Sigma_m^{-1} (V_{ij} - u_m)]) \quad (2)$$

其中  $M$  是复合次数,  $u_m$  是第  $m$  次复合的均值向量,  $\Sigma_m$  是第  $m$  次复合的协方差矩阵,  $D$  是统计指标数目, 这里  $D=3$ 。  $p(V_{ij} | \overline{C_j})$  的计算公式类似(2)。

可以看出, 词  $w_i$  在类别  $C_j$  上的密度函数由均值向量  $u_m$  和协方差矩阵  $\Sigma_m$  所决定。均值向量  $u_m$  和协方差矩阵  $\Sigma_m$  的估计可以利用 [3, 6] 给出的估计算法。

公式(2) 给出的密度函数公式利用协方差矩阵协调了各个统计指标之间的联系, 从而将频度、集中度和分散度有机地统一为一个整体。

### 3.3 基于对数似然比测试的特征词选择算法

设特征词候选集合为  $W$ , 类别  $C_j$  的特征词选择算法描述如下。

- (1) 估计类别先验概率  $p(C_j)$  和  $p(\overline{C_j})$  ;
- (2) 对于每个  $w_i \in W$ ,
  - 计算  $w_i$  在  $C_j$  中的频度、集中度和分散度, 并构成向量  $V_{ij} = ( , , )$  (具体公式在 4.2 节给出);
  - 根据上一步的结果计算联合概率密度函数  $p(V_{ij} | C_j)$  和  $p(V_{ij} | \overline{C_j})$  (公式 2);
  - 计算  $w_i$  对于  $C_j$  和的对数似然比  $LLR$ 。若  $LLR > T$ , 则将  $w_i$  选为  $C_j$  特征词

## 四、实验与评测

实验分别利用频度、集中度和分散度三种测试串行进行(方法 1)和基于对数似然比测试(方法 2)这两种方法进行特征词选择, 通过实验结果比较两个方法。实验语料类别有工业、农业、医学、法律和军事五种。每类有 600 个文本, 每个文本只属于一个类别。实验分为封闭测试和开放测试两种类型。

### 4.1 实验步骤

自动文本分类实验的步骤如下:

(1) 对训练文本和测试文本进行自动分词;

(2) 各个文本类别的局部特征词的选择:对于各类文本的训练集,分别用以下两种方法抽取每类文本的局部特征词集合。

对于任一词  $w_i$ , 任一类别  $C_j$ ,

方法 1(串行方法):如果频度  $> T$ ,集中度  $> T$ ,分散度  $> T$ ,则判定  $w_i$  为  $C_j$  的特征词;

方法 2(对数似然比测试):用对数似然比测试抽取每个类别的特征词;

(3) 将每个类别的局部特征词集合的并集作为全局特征词集合,在上分别利用以下两种方法建立每个类别的局部特征向量。

对于任意  $w_i$ , 任一类别  $C_j$ ,

方法 1(串行方法): $w_i$  在类别  $C_j$  的特征向量中的权值为  $(+ + )/3$ ;

方法 2(对数似然比测试): $w_i$  在类别  $C_j$  的特征向量中的权值为其对数似然比 LLR;

(4) 文本自动标引:对于全局特征词集合中的每个词  $w_i$ , 对于任一文本  $d_j$ , 用特征加权公式  $n_{ij}/N_j$  对  $d_j$  进行标引,其中  $n_{ij}$  是  $w_i$  在文本  $d_j$  中的出现次数, $N_j$  是文本  $d_j$  的规模。

(5) 相似度计算。

## 4.2 集中度和分散度的计算

在本文的研究中,统计指标集中度和分散度的计算方法如下。

(1) 局部特征词选择的统计量——集中度:一个对某类文本具有表征作用的词,应该集中出现在该类文本中,而不是均匀地分布在各类文本中。

设词  $w_i$  在类别  $C_j$  中的出现频度为  $n_{ij}$ ,则其集中度表示如下:

$$= \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \quad m_{ij} = \frac{n_{ij}}{\sum_{l=1}^l \sum_{i=1}^k n_{il}}$$

其中  $k$  是词集的规模, $l$  是文本类别总数。

(2) 局部特征词选择的统计量——分散度:设词  $w_i$  是某个文本类别  $C_j$  的特征词,则  $w_i$  应该均匀出现在  $C_j$  类的所有文本中,而不是出现在  $C_j$  类的一个或少数几个文本中。

设训练集中类别为  $C_j$  的文本有  $d_{j1}, \dots, d_{jr}, \dots, d_{jt}$ ,词  $w_i$  在文本  $d_{jr}(1 \leq r \leq t)$  中的出现频度为  $y_{ij}^r$ ,则其分散度表示如下:

$$= \sum_{r=1}^t (x_{ij}^r)^2, \quad (x_{ij}^r)^2 = \frac{(y_{ij}^r - m_{ij}^r)^2}{m_{ij}^r}, \quad m_{ij}^r = \frac{y_{ij}^r}{\sum_{i=1}^k \sum_{r=1}^t y_{ij}^r}$$

## 4.3 实验结果

实验分别以每个类别 50 篇、100 篇、150 篇、200 篇、250 篇、300 篇、350 篇、400 篇、450 篇、500 篇为训练集,测试文本分类的准确率,结果如图 1 和图 2 所示。

结论:从以上实验结果可以看出,无论是封闭测试还是开放测试,基于对数似然比测试的特征词选择方法(方法 2)都好于几种测试指标串行进行的方法(方法 1),而且随着训练规模的扩大,这种优势逐渐显示出来。这个实验结果也符合理论上的分析。

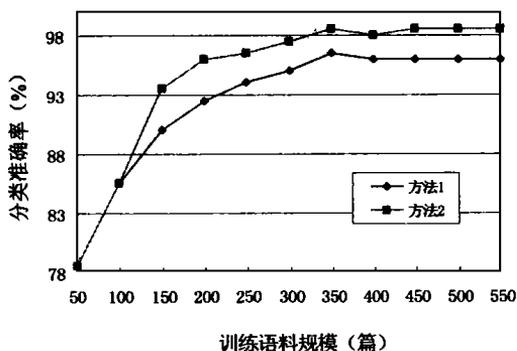


图1 两种方法分类准确率比较(封闭测试)

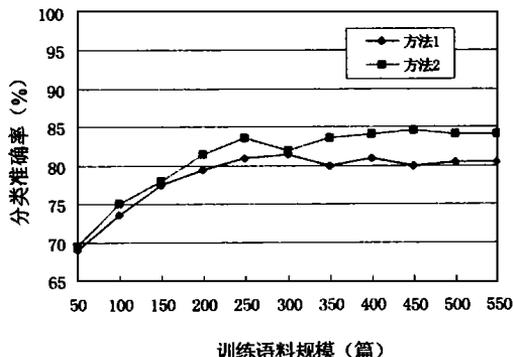


图2 两种方法分类准确率比较(开放测试)

## 五、进一步的工作

本文着重讨论文本分类研究中的特征词的选择问题。实验表明,用对数似然比测试将多种统计量结合起来进行特征词的选择是文本分类研究中一种较好的特征词选择方法。作为一种模式识别的方法,这种技术还可以广泛应用于语言信息处理的很多领域。下一步,作者将尝试将这种方法应用于语料库中新词的抽取以及领域相关术语的抽取等研究中。

## 参 考 文 献

- [1] Apte C *et al.* Automated learning of decision rules for text categorization. ACM Transaction on Information Systems July 1994 ,12(3)
- [2] Burtle C. Statistics in Linguistics. Basil Blackwell World Publishing Corp. 1985
- [3] Duda R O *et al.* Pattern Classification and Scene Analysis. John Wiley & Sons , NY , USA , 1973
- [4] Lewis D D *et al.* Evaluating and optimizing autonomous text classification systems. In: Proceedings of the 18<sup>th</sup> SIGIR Conference , 1995
- [5] Yang Y. Noise reduction in a statistical approach to text categorization. In: Proceedings of the 18<sup>th</sup> SIGIR Conference , 1995
- [6] Young S. The HTK Book. Cambridge University,1997
- [7] 吴军. 汉语语料的自动分类. 中文信息学报,1995(4)
- [8] 杨允信. 中文文件自动分类之研究. 见:台湾第六届计算语言学研讨会论文集,1993
- [9] 蔡元龙. 模式识别. 西安:西北电讯工程学院出版社,1986
- [10] 丁均彦. 文本分类系统的研究与实现[硕士学位论文]. 北京:清华大学,1998