ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2013/07(10)-0933-09
doi: 10.3778/j.issn.1673-9418.1305013

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

基于风险决策的文本特征选择方法*

赵世琛[1],王文剑[1,2+],郭虎升[1]
1. 山西大学 计算机与信息技术学院,太原 030006
2. 山西大学 计算智能与中文信息处理教育部重点实验室,太原 030006

Text Feature Selection Approach Based on Venture Decision*

ZHAO Shichen[1], WANG Wenjian[1,2+], GUO Husheng[1]
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China
+ Corresponding author: E-mail: wjwang@sxu.edu.cn

ZHAO Shichen, WANG Wenjian, GUO Husheng. Text feature selection approach based on venture decision. Journal of Frontiers of Computer Science and Technology, 2013, 7(10): 933-941.

Abstract: The selection of feature words would severely affect the accuracy of text categorization. In view of this situation, this paper proposes a novel text feature selection approach based on dynamic venture decision. This approach uses utility function to evaluate the utility value of each feature word in text categorization, then uses venture decision method to work out the loss of each feature word, finally selects some feature words with lower losses for reducing dimensions. The proposed approach is applied to the spam filtering and Web category in Chinese. The experimental results on several benchmark datasets show that the proposed feature selection approach can select those feature words which will influence the classification results greatly. In so doing, the accuracy of text classification can be improved significantly.

Key words: text categorization; feature selection; venture decision

摘 要: 在中文文本分类中,特征词的选择会严重影响文本分类的准确率。针对这一问题,提出了基于风险决策的文本特征选择方法,通过构造效用函数来评价文本中每个特征词对分类结果的效用值,再采用风险决策

* The National Natural Science Foundation of China under Grant Nos. 60975035, 61273291 (国家自然科学基金); the Foundation for Returnees of Shanxi Province under Grant No. 2012008 (山西省回国留学人员科研基金).
Received 2013-03, Accepted 2013-05.
CNKI网络优先出版:2013-05-15, http://www.cnki.net/kcms/detail/11.5602.TP.20130515.1608.001.html.

方法计算出每个特征词的损失期望,最终选择部分损失期望小的特征词以达到降维的目的。将该方法应用于中文垃圾邮件过滤与网页分类中,实验结果表明,该方法可以选取出对分类结果影响更大的特征词,使文本分类的各项指标明显提高。

关键词:文本分类;特征选择;风险决策

文献标志码:A **中图分类号**:TP18

1 引言

中国互联网络信息中心在《第30次中国互联网络发展状况调查统计报告》[1]中指出,截止至2012年6月,中国网站数量为250万,其中包含了海量的文本信息。这些基于Web的文本数据与传统文本数据有着很大不同:前者大部分是以动态网页的形式出现,即文本数据集是动态的,而后者大多为静态的文本数据。因此,如何对动态文本数据进行有效的分析成为近年来数据挖掘的一个重要研究方向。

文本分类是文本数据挖掘的一个重要分支。目前常用的文本分类方法主要有三类:基于统计的方法[2-4]、基于连接的方法[5]和基于规则的方法[6-8]。其中基于统计的方法通过统计分析训练集中特征词分布情况,更加全面、客观、准确地反应了原始文本中数据信息,故而这类方法应用最为广泛。目前常用的基于统计的方法有支持向量机(support vector machine, SVM)[9]、朴素贝叶斯(naive Bayes, NB)、类中心向量、回归模型和最大熵模型等,在这些方法中,由于SVM具有坚实的理论基础和良好的泛化能力,具有更好的分类效果。

空间向量模型(vector space model, VSM)是处理文本数据时的一种常用表示模型[10],VSM一般先从文本数据集中选取出合适的特征词组成特征词典,然后通过特征词典构造出每条文本的向量模型。不同的特征选择方法会对文本分类精度产生很大影响,目前在文本分类中常见的特征选择方法有互信息[11-12]、CHI统计[13-14]、信息增益[15]等。这些方法通过赋予每个特征词一定权值,从而选出合适的特征词以提高分类的精度。虽然这些方法可以间接反应出特征词对文本分类的贡献度,但是仍会遗漏一部分对分类有重要影响的特征词。此外,对于动态文本分类问题,这些方法并不能有效地将“动态”这一特性表达出来。

决策理论是运筹学和控制论中的重要组成部分,风险决策是研究不确定性决策问题的一种系统分析方法,其目的是改进决策过程,从一系列备选方案中找出一个能满足一定目标的合适方案,其日益广泛地用于商业、经济、实用统计、法律、政治等各方面[16-20]。

本文提出两种基于风险决策的文本特征选择方法,将特征词的选择看成一系列决策优化问题,从而选取出对分类结果影响大的特征词,以达到提高分类精度的目的。

2 基于动态风险决策的特征选择方法

因为无法准确判断那些在各类别文本中均出现的词是否对文本的分类起作用,也无法判断那些只在某类文本中出现的词是否能更好地表示该类文本的特征,所以将它们用做特征词构造特征词典时就会有一定的风险。为了构造更好的特征词典,本文采用风险决策方法进行特征词的选择。首先针对静态文本数据提出一种特征词的静态风险决策特征选择方法,在此基础上提出特征词动态风险决策特征选择方法。

2.1 基于静态风险决策的特征选择方法

这种方法首先统计文本中所有出现词的词频和概率分布,再采用效用函数评价每个特征词对于区分每类文本的效用值,最后用静态风险决策方法计算出每个特征词的期望损失值,选出期望损失值小的特征词组成特征词典。

这里记文本的类别为状态 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$,其中 n 为文本的类别数,将每类文本在训练集中出现的概率分别记为 $\pi(\theta_1),\pi(\theta_2),\cdots,\pi(\theta_n)$。每个特征词是否被选取看成是一次行动,由原始特征词典组成行动集 $A=\{a_1,a_2,\cdots,a_m\}$,其中 m 为特征词的个数。

函数 $U(\theta_i, a_j)$ 表示选择特征词 a_j 对产生状态 θ_i 的效用值,其定义为:

$$U(\theta_i, a_j)=\frac{tf(a_j)\times \ln(N_{\theta_i}/n_{\theta_{ij}}+0.01)}{\sqrt{\sum_{a\in A}[tf(a_j)\times \ln(N_{\theta_i}/n_{\theta_{ij}}+0.01)]^2}} \quad (1)$$

其中,$tf(a_j)$ 表示特征词 a_j 在它所在的文本中出现的频率;N_{θ_i} 为第 i 状态中文本的总数目;$n_{\theta_{ij}}$ 表示出现特征词 a_j 的文本在状态 θ_i 下出现的数目。分母进行归一化处理。

定义损失函数为 $L(\theta_i, a_j)=-U(\theta_i, a_j)$,表示采取特征词 a_j 对状态 θ_i 的损失值。将可行的行动集和可能出现的状态表示出来,由此可得到静态风险决策问题的决策表,如表1所示。

Table 1 The decision table of static risk

表1 静态风险决策表

	a_1	a_2	…	a_j	…	a_m
θ_1	L_{11}	L_{12}	…	L_{1j}	…	L_{1m}
θ_2	L_{21}	L_{22}	…	L_{2j}	…	L_{2m}
⋮	⋮	⋮		⋮		⋮
θ_i	L_{i1}	L_{i2}	…	L_{ij}	…	L_{im}
⋮	⋮	⋮		⋮		⋮
θ_n	L_{n1}	L_{n2}	…	L_{nj}	…	L_{nm}

针对自然状态集 Θ,选取特征词 a_j 时损失函数 $L(\Theta, a_j)$ 对特征词 a_j 的期望值称为风险函数,记做 $R(\Theta, A)$,即

$$R(\Theta, A)=E_{\Theta}^{A}[l(\Theta, A)]=\sum_{\theta\in\Theta} l(\theta, a)p(a|\theta) \quad (2)$$

当自然状态的先验概率为 $\pi(\Theta)$ 时,风险函数 $R(\Theta, A)$ 关于自然状态 Θ 的期望值即为贝叶斯风险,记做 $r(\pi(\Theta), A)$,即

$$r(\pi(\Theta), A)=E^{\pi}[R(\Theta, A)]=E^{\pi}[E_{\Theta}^{A}l(\theta, a)]=\sum_{\theta\in\Theta} l(\theta, a)p(a|\theta)\pi(\theta) \quad (3)$$

如果 $r(\pi(\Theta), a_i)<r(\pi(\Theta), a_j)$,则特征词 a_i 比特征词 a_j 更适合放在特征词典中。通过设定一个阈值,将贝叶斯风险小于阈值的特征词抽取出来组成新的特征词典,以达到降维的目的。

2.2 基于动态风险决策的特征选择方法

静态风险决策方法选取出的特征词可以更加准确地代表特征词所属的文本类别,但是对于动态文本数据,静态风险决策方法与传统方法一样不能将数据"动态"这一特性表示出来。因此在静态风险决策方法基础上提出了一种特征词的动态风险决策方法。这种方法在处理动态文本分类时,可以根据之前分类的结果对特征词的选取进行不断的修正,构造出可以更准确地描述当前文本数据集中特征词分布情况的动态特征词典。

为了构造动态特征词典,效用函数需要重新构造。记文本的分类状态为 $\lambda=\{\lambda_0, \lambda_1\}$,其中 λ_0、λ_1 分别表示文本分类正确和分类错误的状态,则效用函数 $U(\lambda_i, a_j)$ 定义如下:

$$U(\lambda_i, a_j)=\frac{tf(a_j)\times \ln(N_{\lambda_i}/n_{\lambda_{ij}}+0.01)}{\sqrt{\sum_{a\in A}[tf(a)\times \ln(N_{\lambda_i}/n_{\lambda_{ij}}+0.01)]^2}} \quad (4)$$

虽然式(4)与式(1)形式上大体相同,但是文本的分类状态完全不同,式(1)中 θ 表示的是文本的类别状态,而在式(4)中 λ 表示的是文本分类结果的状态。因此将采用式(1)的方法称为静态风险决策方法,将采用式(4)的方法称为动态风险决策方法。由式(4)得到每个词对当前状态的效用值,进而构造出动态风险决策方法的决策表,如表2所示。

Table 2 The decision table of dynamic risk

表2 动态风险决策表

	a_1	a_2	…	a_i	…	a_m
λ_1	L_{11}	L_{12}	…	L_{1i}	…	L_{1m}
λ_2	L_{21}	L_{22}	…	L_{2i}	…	L_{2m}

将决策表中的参数代入式(5)中,即可得到每个特征词的风险期望:

$$r(\pi(\lambda), A)=E^{\pi}[R(\lambda, A)]=E^{\pi}[E_{\lambda}^{A}l(\lambda, a)]=\sum_{\lambda} l(\lambda, a)p(a|\lambda)\pi(\lambda) \quad (5)$$

式(5)中,$l(\theta, \lambda)$ 体现了特征词 a 对分类状态 λ 的损失值,当状态 λ 为分类正确时,特征词 a 的损失值为负数,反之为正数;$p(a|\lambda)$ 为状态 λ 时特征词 a 出现的频数,

意味着当分类正确时，特征词出现的频数越高，则它的损失期望就会越小，当分类错误时，特征词出现的频数越高，则它的损失期望就越大；$\pi(\lambda)$ 为状态 λ 出现的先验概率，其中 $\pi(\lambda_1)>\pi(\lambda_2)$ 且 $\pi(\lambda_1)+\pi(\lambda_2)=1$，这样分类正确中损失值较大的特征词会得到更大的损失期望，而分类错误中损失值较小的特征词的损失期望会变得更小。最终通过以上三步得到更加合适客观的特征词权重。

3 实验结果和分析

将本文方法分别用于中文垃圾邮件过滤与网页分类中，以检验算法的有效性。并与互信息、信息增益和CHI统计三种常用特征选择方法进行实验比较。

3.1 中文垃圾邮件过滤

实验采用中国教育和科研计算机网紧急响应组(CCERT)2006年5月公布的共16 000封电子邮件数据集，其中包含8 000封正常邮件和8 000封垃圾邮件[21]。实验将邮件随机分为四组，每组由正常邮件和垃圾邮件各1 000封组成训练集，另取正常邮件和垃圾邮件各1 000封组成测试集。评价指标采用准确率(correct rate，CR)、正常邮件通过率(normal mail rate，NMR)和正确过滤率(correct rejection rate，CRR)[10]，分别定义如下：

$$CR=\frac{N_{H\to H}+N_{S\to S}}{N_{total}} \tag{6}$$

$$NMR=\frac{N_{H\to H}}{N_{H\to S}+N_{H\to H}} \tag{7}$$

$$CRR=\frac{N_{S\to S}}{N_{S\to H}+N_{S\to S}} \tag{8}$$

其中，$N_{H\to H}$ 表示将正常邮件判断为正常邮件的数目；$N_{H\to S}$ 表示将正常邮件判断为垃圾邮件的数目；$N_{S\to H}$ 表示将垃圾邮件判断为正常邮件的数目；$N_{S\to S}$ 表示将垃圾邮件判断为垃圾邮件的数目；N_{total} 表示邮件总数。

为验证本文方法的有效性，将其与CHI统计、互信息和信息增益三种方法进行了实验比较。由于降维后特征词典的大小也会严重影响到实验结果，因此实验通过对第一组数据集降维后选取不同的维度，以检验其值的选择对邮件过滤准确率的影响。实验结果如图1所示。通过观察可以发现，当特征词典维度低于600时，垃圾邮件过滤的准确率随着维度的增加而增加，而当维度大于等于600后，准确率趋于稳定，变化较小。因此在中文垃圾邮件过滤实验中，将降维后特征词典的大小设定为600维。

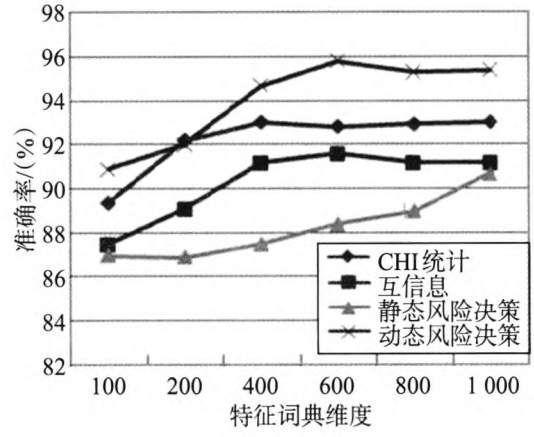

Fig.1 The dimension effect on accuracy in spam filtering

图1 垃圾邮件过滤中维度的选择对准确率的影响

将降维后的空间向量模型用SVM分类器进行学习，其中SVM采用粒度高斯核函数，文献[22-23]对文本分类中SVM参数的选取进行了详细的实验和说明，因此本文中正则参数 C 取100，核参数 σ 取1.0。由于信息增益方法得到的实验结果较差，因此表3只列出CHI统计、互信息、静态风险决策和动态风险决策的实验结果(↑表示值越大越好)。

从表3可以观察到，动态风险决策方法在整体准确率上是最好的。在重要指标 *NMR* 上，第一、三、四组数据集上，动态风险决策方法要比另三种方法高出1%~7%，而在第二组数据集上，动态风险决策方法略低于CHI统计0.5%。在指标 *CRR* 上，由于数据集的特殊性，除动态风险决策方法外，另三种方法在降维后特征词典中的大部分词取自非垃圾邮件，因此造成三种方法的 *CRR* 值均不是很高；而动态风险决策方法直接从当前分类结果出发，分析每个特征词对当前分类结果的损失期望，避免了上述问题，因此在指标 *CRR* 上要明显优于另三种特征选择方法。

Table 3 The accuracy of spam filtering

表3 垃圾邮件过滤精度 (%)

组别	方法	CR↑	NMR↑	CRR↑
第一组	CHI统计	86.8	95.4	78.2
	互信息	91.6	94.4	88.8
	静态风险决策	88.4	97.2	79.6
	动态风险决策	95.8	97.6	94.1
第二组	CHI统计	86.1	98.1	74.2
	互信息	87.5	96.1	79.0
	静态风险决策	85.9	96.2	75.6
	动态风险决策	95.4	97.6	93.2
第三组	CHI统计	95.7	98.1	73.4
	互信息	91.8	92.1	91.5
	静态风险决策	84.8	96.5	73.1
	动态风险决策	97.2	99.6	94.8
第四组	CHI统计	92.9	97.5	88.3
	互信息	91.7	90.7	92.7
	静态风险决策	83.4	96.1	70.7
	动态风险决策	97.8	97.6	98.1

图2是四种方法在四组邮件集上各指标平均值的比较。从图2可以明显看到,动态风险决策方法三个指标均优于其余三种方法。在指标整体准确率中,比另三种方法高出7%~10%;在重要指标正常邮件通过率中,比另三种方法高出1%~5%;在指标正确过滤率中,比另三种方法高出7%~20%。

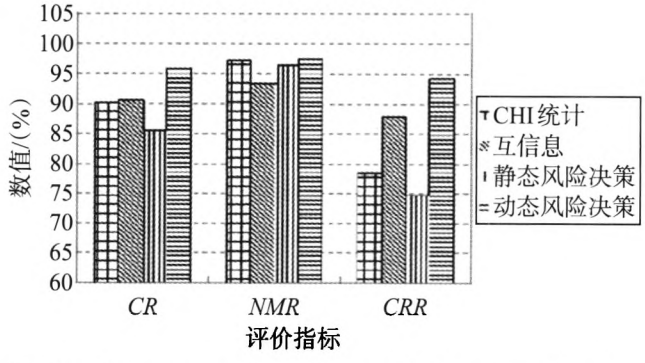

Fig.2 The average value of four methods in spam filtering

图2 垃圾邮件过滤中四种方法各指标的平均值

3.2 网页分类

3.2.1 中文网页分类

实验采用由复旦大学整理的中文网页语料库(http://ishare.iask.sina.com.cn/f/22774613.html),从中选取出环境、计算机、交通、教育、太空、体育、农业、艺术和政治10大类共2 600篇文本。实验共分为四组,每组随机抽取1 000篇作为训练集,另随机抽取400篇作为测试集,其中每类文本在训练集与测试集中所占比例相同。实验通过查准率(Precision)、召回率(Recall)和 $F\text{-}measure_1$ 评价算法的好坏[4],评价指标分别定义如下:

$$查准率=\frac{检索到相关文档数}{检索到文档总数} \tag{9}$$

$$召回率=\frac{检索到相关文档数}{所有相关文档总数} \tag{10}$$

$$F\text{-}measure_1=\frac{2\times查准率\times召回率}{查准率+召回率} \tag{11}$$

互信息方法在文本多分类问题中效果较差,因此选用信息增益和CHI统计两种方法进行对比实验。通过实验可以观察到,降维后特征词典的维度在大于等于1 200维后网页分类的 $F\text{-}measure_1$ 趋于平缓(见图3),因此在网页分类中,将降维后特征词典的大小设定为1 200维。

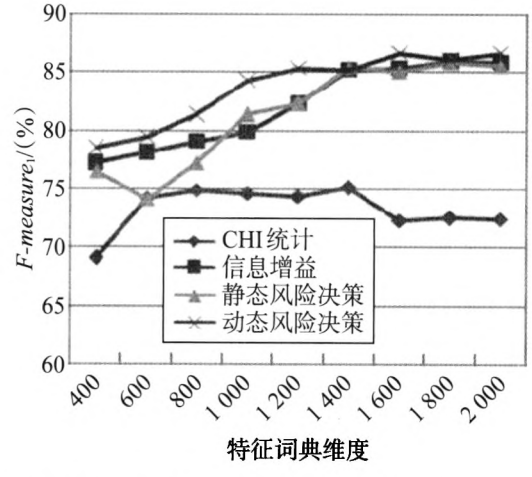

Fig.3 The dimension effect on $F\text{-}measure_1$ in Web page classification

图3 网页分类中降维后维度的选择对 $F\text{-}measure_1$ 值的影响

将降维后的空间向量模型用一对一SVM多分类器进行学习,其中SVM采用粒度高斯核函数,正则参数 C 取100,核参数 σ 取1.0[21]。实验结果如表4所示。

Table 4 The accuracy of Chinese Web page classification

表4 中文网页分类精度 (%)

组别	方法	$F\text{-}measure_1$↑	查准率↑	召回率↑
第一组	CHI统计	74.4	60.9	95.5
	信息增益	82.3	71.6	96.8
	静态风险决策	82.4	72.5	95.2
	动态风险决策	85.4	75.9	97.6
第二组	CHI统计	75.4	62.6	94.8
	信息增益	80.1	68.2	96.7
	静态风险决策	81.4	71.5	94.4
	动态风险决策	82.4	71.9	96.5
第三组	CHI统计	80.7	69.7	96.1
	信息增益	83.1	72.4	97.2
	静态风险决策	84.2	75.2	95.7
	动态风险决策	82.9	71.9	97.9
第四组	CHI统计	81.5	70.6	96.4
	信息增益	82.2	71.3	96.8
	静态风险决策	80.3	69.7	94.6
	动态风险决策	83.7	73.7	96.9

由表4可以观察到，在主要指标查准率上，动态风险决策方法在四组数据集中均优于另三种方法。在整体指标 $F\text{-}measure_1$ 上，动态风险决策方法在第一、二、四组数据集中均优于另三种方法，而在第三组数据集中，静态风险决策结果较好，动态风险决策方法略低于信息增益和静态风险决策。在指标召回率上，动态风险决策方法除了在第二组数据集中低于信息增益0.2%，在其余数据集上均高于另三种方法。

图4是四种方法在四组邮件集上各指标平均值的比较。由图4可以看出，动态风险决策方法在三个指标上实验结果均为最好。具体在指标 $F\text{-}measure_1$ 上，动态风险决策比另三种方法高出2%~5%，其次依次为静态风险决策、信息增益和CHI统计。在指标查准率上，动态风险决策比另三种方法高出1%~8%，其次依次为静态风险决策、信息增益和CHI统计。在指标召回率上，动态风险决策比另三种方法高出1%~3%，其次依次为信息增益、CHI统计和静态风险决策。

3.2.2 英文网页分类

为验证本文方法的鲁棒性，选取20-newsgroup数据集（http://download.csdn.net/download/miracletiger/2474955）中的alt.atheism、rec.autos、rec.sport.baseball、sci.crypt、sci.electronics、misc.forsale、comp.graphics、rec.sport.hockey、sci.med、rec.motorcycles、sci.space，共10个大类，2 600篇网页，以验证本文方法对英文数据集的有效性。实验分为四组，每组随机抽取1 000篇作为训练集，另随机抽取400篇作为测试集，其中每类文本在训练集与测试集中所占比例相同。本文依旧选择信息增益和CHI统计两种方法进行对比实验。通过设定阈值，将特征词典降低到1 200个特征词。将降维后的空间向量模型用一对一SVM多分类器进行学习，其中SVM采用粒度高斯核函数，正则参数 C 取100，核参数 σ 取1.0。实验结果如表5所示。

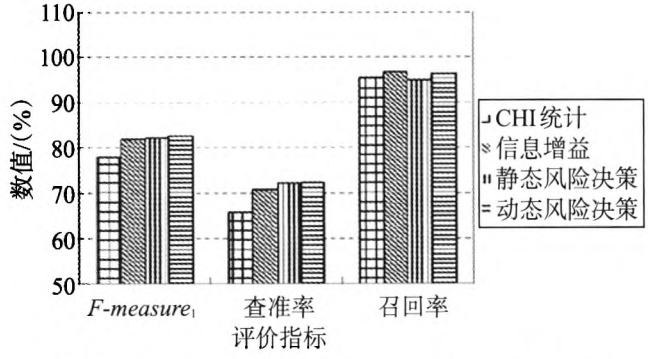

Fig.4 The average value of four methods in Chinese Web page classification

图4 中文网页分类中四种方法各指标的平均值

由表5可以观察到，在主要指标 $F\text{-}measure_1$ 和查准率上，动态风险决策方法在四组数据集上都要优于另三种方法；而在召回率上，四种特征选择方法在所有数据集上均有较高的召回率。

图5是四种方法在四组邮件集上各指标平均值的比较。由图5可以观察到，在综合指标 $F\text{-}measure_1$ 中，动态风险决策比CHI统计、静态风险决策、信息增益分别高出12%、6%和3%。在重要指标查准率上，动态风险决策比另三种方法依次高出15%、9%和4%。在指标召回率上，四种方法的评价召回率都在99%以上。

从以上实验可以看出，动态风险决策方法在平均分类精度上要优于另四种方法，并且对于不同的

Table 5 The accuracy of English Web page classification

表5 英文网页分类精度 (%)

组别	方法	$F\text{-}measure_1$↑	查准率↑	召回率↑
第一组	CHI统计	72.4	57.0	99.1
	信息增益	83.2	71.6	99.3
	静态风险决策	79.8	66.4	100
	动态风险决策	86.2	75.7	100
第二组	CHI统计	69.9	54.3	98.1
	信息增益	80.5	67.6	99.3
	静态风险决策	74.9	60.1	99.5
	动态风险决策	83.5	71.9	99.6
第三组	CHI统计	70.4	54.5	99.5
	信息增益	78.2	64.2	100
	静态风险决策	75.6	60.9	99.5
	动态风险决策	82.5	70.2	100
第四组	CHI统计	70.7	54.8	99.5
	信息增益	77.8	63.7	100
	静态风险决策	73.9	58.8	99.5
	动态风险决策	79.3	65.8	99.6

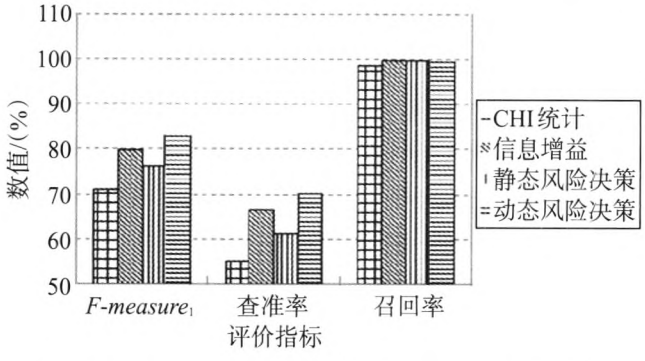

Fig.5 The average value of four methods in English Web page classification

图5 英文网页分类中四种方法各指标的平均值

文本数据集,风险决策方法始终能保持较高的准确率。动态风险决策方法之所以优于其他方法,主要原因是因为特征选择的侧重点不同:CHI统计、互信息、信息增益和静态风险决策方法属于传统的特征选择方法,通过判断特征词所含有的信息量大小来进行特征词的选择,特征词所含的信息量间接反映了特征词对文本分类的贡献程度。而动态风险决策特征选择方法则是从训练集的结果出发,通过实践直接去判断每个特征词对分类结果的影响程度。这种方法并没有将之前分类错误的文本中的特征词全部排除,而是去除那些在分类正确文本中出现频率低,而在分类错误文本中出现频率高的特征词,保留在分类正确文本中出现频率高,但在分类错误文本中出现频率一般的特征词。这种方法选出的特征词典更能直观地代表文本数据集中特征词的分布情况,因此分类效果要强于另四种方法。

4 结束语

本文针对中文文本分类问题提出了一种动态风险决策特征选择方法。这种方法将特征选择看成一个决策问题,用损失函数来评价每个特征词对文本分类的贡献,从而选择出合适的特征词以达到降维的目的,因而较传统方法有更高的准确率和稳定性。但对于某些实际应用如电子邮件的过滤,由于每日有上亿封电子邮件的收发,这样即使邮件过滤器达到99%的准确率,仍意味着百万封的邮件分类错误。在未来的工作中,将结合有效的模型选择方法进一步提高文本的分类能力。

References:

[1] China Internet Network Information Center. 30th China Internet network development state statistic report[DB/OL]. (2012)[2013-01]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201207/t20120723_32497.htm.

[2] Lewis D D. Naïve (Bayes) at forty: the independence assumption in information retrieval[C]//LNCS 1398: Proceedings of the 10th European Conference on Machine Learning (ECML '98) , Dortmund, Germany, 1998. Berlin, Heidelberg: Springer-Verlag, 1998: 4-15.

[3] Joachims T. Text categorization with support vector machines: learning with many relevant features[C]//LNCS 1398: Proceedings of the 10th European Conference on Machine Learning (ECML '98), Dortmund, Germany, 1998. Berlin, Heidelberg: Springer-Verlag, 1998: 137-142.

[4] Yang Yiming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, l(1): 76-88.

[5] Xu Yan. A formal study of feature selection in text categorization[J]. Journal of Communication and Computer, 2009,

6(4): 32-41.

[6] Wang Meihua, Zhang Hongbin, Ding Renshuang. Research of text categorization based on SVM[C]//Proceedings of the 2011 International Conference on Informatics, Cybernetics, and Computer Engineering (ICCE '11), Melbourne, Australia, 2011. Berlin, Heidelberg: Springer-Verlag, 2011: 69-77.

[7] Azam N, Yao Jingtao. Comparison of term frequency and document frequency based feature selection metrics in text categorization[J]. Expert Systems with Applications, 2012, 39(5): 4760-4768.

[8] Wiener E J, Pedersen J, Weigend A. A neural network approach to topic spotting[C]//Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR '95), Las Vegas, USA, 1995: 317-332.

[9] Apte C, Damerau F, Weiss S. Text mining with decision rules and decision trees[C]//Proceedings of the Conference on Automated Learning and Discovery, Pittsburgh, USA, 1998: 1-5.

[10] Lent B, Swami A, Widom J. Clustering association rules[C]// Proceedings of the 13th International Conference on Data Engineering (ICDE '97), Birmingham, UK, 1997. Washington, DC, USA: IEEE Computer Society, 1997: 220-231.

[11] Li Ronglu. Text classification and related technology research[D]. Shanghai: Fudan University, 2005: 4-5.

[12] Drucker H, Wu Donghui, Vapnik V N. Support vector machines for spam categorization[J]. IEEE Transactions on Neural Networks, 1999, 20(5): 1048-1054.

[13] Mesleh A M. Chi square feature extraction based SVMs Arabic language text categorization system[J]. Journal of Computer Science, 2007, 3(6): 430-435.

[14] Peng Hanchuan, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[15] Xu Yan, Li Jintao, Wang Bin. A study on constraints for feature selection in text categorization[J]. Journal of Computer Research and Development, 2008, 45(4): 596-602.

[16] Yang Yiming, Pederson J O. A comparative study on feature selection in text categorization[C]//Proceedings of the 14th International Conference on Machine Learning (ICML '97), Nashville, USA, 1997. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1997: 412-420.

[17] Liu Tao, Liu Shengping, Chen Zheng, et al. An evaluation on feature selection for text clustering[C]//Proceedings of the 20th International Conference on Machine Learning (ICML '03), Washington, USA, 2003: 488-495.

[18] Yue Chaoyuan. Decision theory and method[M]. Beijing: Science Press, 2003: 56-83.

[19] Hsu W-K, Tseng C-P, Chiang W-L, et al. Risk and uncertainty analysis in the planning stages of a risk decision making process[J]. Natural Hazards, 2012, 61(3): 1355-1365.

[20] Li Huaxiong, Zhou Xianzhong. Risk decision making based on decision-theoretic rough set: a three-way view decision model[J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 1-11.

[21] CCERT. Spam dataset[DB/OL]. (2005)[2013-01]. http://www.ccert.edu.cn/spam/sa/datasets.htm.

[22] Hou Yan, Wang Wenjian, A SVM Chinese email filtering approach based on dynamic feature dictionary[J]. Computer Science, 2008, 35(3): 49-51.

[23] Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection[C]//Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '97). Washington, DC, USA: IEEE Computer Society, 1997:130-136.

附中文参考文献:

[1] 中国互联网络信息中心. 第30次中国互联网络发展状况调查统计报告[DB/OL]. (2012)[2013-01]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201207/t20120723_32497.htm.

[11] 李荣陆. 文本分类及其相关技术研究[D]. 上海: 复旦大学, 2005: 4-5.

[15] 徐燕, 李锦涛, 王斌. 文本分类中特征选择的约束研究[J]. 计算机研究与发展, 2008, 45(4): 596-602.

[18] 岳超源. 决策理论与方法[M]. 北京: 科学出版社, 2003: 56-83.

[22] 侯岩, 王文剑. 一种基于动态特征词典的SVM中文电子邮件过滤方法[J]. 计算机科学, 2008, 35(3): 49-51.

ZHAO Shichen was born in 1989. He is a master candidate at Shanxi University. His research interests include machine learning and data mining.
赵世琛(1989—),男,山西太原人,山西大学硕士研究生,主要研究领域为机器学习,数据挖掘。

WANG Wenjian was born in 1968. She is a professor and Ph.D. supervisor at Shanxi University, and the senior member of CCF. Her research interests include machine learning and computing intelligence, etc.
王文剑(1968—),女,山西太原人,博士,山西大学教授、博士生导师,CCF高级会员,主要研究领域为机器学习,计算智能等。

GUO Husheng was born in 1986. He is a Ph.D. candidate at Shanxi University, and student member of CCF. His research interests include machine learning and data mining.
郭虎升(1986—),男,山西太谷人,山西大学博士研究生,CCF学生会员,主要研究领域为机器学习,数据挖掘。

基于风险决策的文本特征选择方法

作者： 赵世琛， 王文剑， 郭虎升， ZHAO Shichen， WANG Wenjian， GUO Husheng

作者单位： 赵世琛,郭虎升,ZHAO Shichen,GUO Husheng(山西大学计算机与信息技术学院,太原,030006)， 王文剑,WANG Wenjian(山西大学计算机与信息技术学院,太原030006;山西大学计算智能与中文信息处理教育部重点实验室,太原030006)

刊名： 计算机科学与探索

英文刊名： Journal of Frontiers of Computer Science & Technology

年，卷(期)： 2013,7(10)

参考文献(28条)

1. China Internet Network Information Center.30th China Internet network development state statistic report 2013

2. Lewis D D Na(ī)ve (Bayes) at forty:the independence assumption in information retrieval 1998

3. Joachims T Text categorizaton with support vector machines:learning with many relevant features 1998

4. Yang Yiming An evaluation of statistical approaches to text categorization 1999(01)

5. Xu Yan A formal study of feature selection in text categorization 2009(04)

6. Wang Meihua;Zhang Hongbin;Ding Renshuang Research of text categorization based on SVM 2011

7. Azam N;Yao Jingtao Comparison of term frequency and document frequency based feature selection metrics in text categorization 2012(05)

8. Wiener E J;Pedersen J;Weigend A A neural network approach to topic spotting 1995

9. Apte C;Damerau F;Weiss S Text mining with decision rules and decision trees 1998

10. Lent B;Swami A;Widom J Clustering association rules 1997

11. Li Ronglu Text classification and related technology research 2005

12. Drucker H;Wu Donghui;Vapnik V N Support vector machines for spam categorization[外文期刊] 1999(05)

13. Mesleh A M Chi square feature extraction based SVMs Arabic language text categorization system 2007(06)

14. Peng Hanchuan;Long F;Ding C Feature selection based on mutual information criteria of max-dependency,max-relevance,and min-redundancy[外文期刊] 2005(08)

15. Xu Yan;Li Jintao;Wang Bin A study on constraints for feature selection in text categorization[期刊论文]-Joumal of Computer Research and Development 2008(04)

16. Yang Yiming;Pederson J O A comparative study on feature selection in text categorization 1997

17. Liu Tao;Liu Shengping;Chen Zheng An evaluation on feature selection for text clustering 2003

18. Yue Chaoyuan Decision theory and method 2003

19. Hsu W-K;Tseng C-P;Chiang W-L Risk and uncertainty analysis in the planning stages of a risk decision making process 2012(03)

20. Li Huaxiong;Zhou Xianzhong Risk decision making based on decision-theoretic rough set:a three-way view decision model 2011(01)

21. CCERT Spam dataset 2013

22. Hou Yan;Wang Wenjian A SVM Chinese email filtering approach based on dynamic feature dictionary[期刊论文]-Computer Science 2008(03)

23. Osuna E;Freund R;Girosi F Training support vector machines:an application to face detection 1997

24. 中国互联网络信息中心 第30次中国互联网络发展状况调查统计报告 2013

25. 李荣陆 文本分类及其相关技术研究[学位论文] 2005

26. 徐燕;李锦涛;王斌 文本分类中特征选择的约束研究[期刊论文]-计算机研究与发展 2008(04)

27. 岳超源 决策理论与方法 2003

28. 侯岩;王文剑 一种基于动态特征词典的SVM中文电子邮件过滤方法[期刊论文]-计算机科学 2008(03)