邮件过滤中特征选择方法的性能评价与分析 *

赵 静[a,b], 刘培玉[a,b], 许明英[a,b]
(山东师范大学 a. 信息科学与工程学院; b. 山东省分布式计算机软件新技术重点实验室, 济南 250014)

摘 要: 基于内容的邮件过滤本质是二值文本分类问题。特征选择在分类之前约简特征空间以减少分类器在计算和存储上的开销,同时过滤部分噪声以提高分类的准确性,是影响邮件过滤准确性和时效性的重要因素。但各特征选择算法在同一评价环境中性能不同,且对分类器和数据集分布特征具有依赖性。结合邮件过滤自身特点,从分类器适应性、数据集依赖性及时间复杂度三个方面评价与分析各特征选择算法在邮件过滤领域的性能。实验结果表明,优势率和文档频数用于邮件过滤时垃圾邮件识别的准确率较高,运算时间较少。

关键词: 邮件过滤; 特征选择; 朴素贝叶斯; 支持向量机
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0693-05
doi:10.3969/j.issn.1001-3695.2012.02.078

Evaluation and analysis of feature selection methods for e-mail filtering

ZHAO Jing[a,b], LIU Pei-yu[a,b], XU Ming-ying[a,b]
(*a. School of Information Science & Engineering, b. Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan* 250014, *China*)

Abstract: The nature of content-based e-mail filtering is a binary text classification problem. Feature selection methods reduced the feature dimension before classifying e-mails in order to reduce the cost of computing and storage, while filtering some noise features to improve the classification accuracy. Feature selection was an important factor which decided the accuracy and timeliness of e-mail filtering. However, every feature selection algorithm had different performance in the same environment, and was affected by classifiers and data distribution. Combining characteristics of e-mail filtering, this paper evaluated and analized the following aspects of feature selection methods which used to filter e-mails: classifier adaptability, data set dependence, time complexity. Experimental results show that odds ratio and document frequency have higher accuracy and less computing time when they are used to filter emails.

Key words: e-mail filtering; feature selection; Naïve Bayes; SVM

0 引言

随着互联网的发展,电子邮件成为人们交流和通信的主要方式。同时,网络中存在的大量垃圾邮件浪费了用户的时间,降低了用户使用邮箱的兴趣,损害了 ISP 的利益,对用户和互联网造成巨大危害。Symantec Intelligence Report 显示,2011 年 7 月垃圾邮件占全球邮件总量的 77.8%,较 6 月份上升 4.9%[1]。中国互联网协会反垃圾邮件中心最新发布的 2011 年第一季度中国反垃圾邮件状况调查报告显示:中国网民平均每周收到垃圾邮件的数量为 13.8 封,垃圾邮件占比 40.1%;超过 66% 的用户收到了"欺诈类"类型的垃圾邮件;每周处理垃圾邮件的时间为 7.9 min[2]。由此可见,垃圾邮件已经成为世界各国共同面临的棘手问题。

为有效防止垃圾邮件,国内外众多学者及研究机构提出多种反垃圾邮件技术,如基于黑白名单的过滤技术、基于规则的过滤技术、基于统计的内容过滤技术、身份认证技术、基于行为模式的垃圾邮件识别技术等。其中,基于统计的内容过滤技术因过滤效果好、能够及时捕捉垃圾邮件特征的变化、人工干预少,在反垃圾邮件过程中发挥了重要作用。因此,研究内容过滤的关键技术以提高垃圾邮件识别的准确率和召回率具有重要的现实意义。

1 相关研究

基于统计的邮件内容过滤技术将邮件分为合法邮件和垃圾邮件,是文本分类的应用领域之一。特征选择作为文本分类的重要环节,解决了特征项集维数过高或存在较多噪音特征词,从而增加分类运算时间和空间复杂度、降低分类准确率的问题。由于各方法的性能差异使得其在同一应用环境中有不同分类结果,且同一特征选择算法在不同应用领域中也表现出不同特征。邮件过滤与一般的文本分类不同,一个性能良好的特征选择算法应符合邮件过滤对实时性、准确性的要求。

1.1 常用特征选择方法

目前,文本分类常用的特征选择方法有文档频数、信息增益、期望交叉熵、互信息、文本证据权、优势率等。

1)文档频数(document frequency,DF)

DF 表示训练集合中包含特征项的文本数目,是最简单的

收稿日期: 2011-08-12; **修回日期**: 2011-09-13 **基金项目**: 国家自然科学基金资助项目(60873247);山东省高新自主创新专项工程资助项目(2008ZZ28);山东省自然科学基金重点资助项目(ZR2009GZ007)

作者简介:赵静(1987-),女,山东聊城人,硕士研究生,主要研究方向为网络信息过滤(sdzhjing1987@163.com);刘培玉(1960-),男,教授,博导,主要研究方向为网络信息安全、网络系统规划、网络信息资源管理;许明英(1987-),女,硕士研究生,主要研究方向为网络信息过滤.

评估函数。该方法假设噪声词或所含信息量少的稀有单词对分类影响小,可以删去,其优点是计算复杂度低,在实际应用中效果好,能适应于大规模数据集。但在实际应用中,稀有单词可能包含重要的判断信息,简单舍弃将影响分类器的精度。

2)信息增益(information gain,IG)

信息增益是一种基于熵的评估方法,定义为特征 t 在文本中出现前后的信息熵之差,计算采用式(1)。

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

IG 衡量某个特征是否存在对类别预测的影响,同时考虑特征出现和未出现两种情况,倾向于选择在某一类别中出现频率高而在其他类别中出现频率低的特征。由于不在文本中出现的特征词对分类的贡献往往小于其带来的干扰,考虑其未出现情况反而降低了信息增益的效果,尤其是在样本分布和特征分布不均匀的情况下较为明显[3]。

3)CHI 统计(Chi-squared,CHI)

CHI 统计假设特征与类别间服从 χ^2 分布,可度量特征和类别之间的独立性。其值越高,特征项与类别之间的独立性越小,相关性越大,计算采用式(2)。

$$CHI(t) = \sum_{i=1}^{m} p(c_i)\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (2)$$

其中:A 表示特征 t 与类别 c_i 文档同现的次数,B 表示特征 t 出现而 c_i 类文档不出现的次数,C 表示 c_i 类文档出现而特征 t 不出现的次数,D 表示 c_i 类文档与特征 t 都不出现的次数,N 为总文档数。

CHI 方法用于特征选择时,倾向于选择在指定类别文本中出现频率高的特征词和在其他类别文本中出现频率比较高的词,在实际应用中其可靠性较好,无须因训练集的改变而人为调节特征阈值。但当特征与类别不符合 χ^2 分布时,其倾向于选择低频特征词。

4)互信息(mutual information,MI)

MI 度量特征与各类别的相关度,常用平均互信息来度量特征对类别的区分性,计算采用式(3)。

$$MI(t) = \sum_{i=1}^{m} p(c_i)\log\frac{p(t \mid c_i)}{p(t)} \quad (3)$$

互信息方法倾向于选择在某一类别中出现频率高但在其他类别中出现频率低的特征词。由于其计算过程中未考虑到特征出现的频率,使其删掉很多高频词,这使得互信息在特征选择中效果较差[4]。

5)期望交叉熵(expected cross entropy,ECS)

交叉熵反映了类别概率与给定特征词下类别概率之间的距离,交叉熵越大,类别区分性越好,其与信息增益相似,但没有考虑特征词未发生的情况,用于特征选择时采用式(4)计算。

$$ECE(t) = p(t)\sum_{i=1}^{m} p(c_i \mid t)\log\frac{p(c_i \mid t)}{p(c_i)} \quad (4)$$

6)文本证据权(weight of evidence for text,WET)

WET 衡量类别概率与给定特征条件下类别概率之间的差别[5]。与期望交叉熵相似,但其倾向于选择与类别强相关的特征,增强了特征之间的区分度,采用式(5)计算其值。

$$WET(t) = p(t)\sum_{i=1}^{m} p(c_i)\left|\log\frac{p(c_i)(1-p(c_i))}{p(c_i)(1-p(c_i \mid t))}\right| \quad (5)$$

7)优势率(odds ratio,OR)

优势率只关心目标类别,体现了特征项作为目标类别类内文本特征的优势[6]。当 pos 表示目标类,neg 表示非目标类时,采用式(6)计算。

$$OR(t) = \log\frac{p(t \mid c_{pos})(1-p(t \mid c_{neg}))}{(1-p(t \mid c_{pos}))p(t \mid c_{neg})} \quad (6)$$

1.2 算法的性能分析

本文 1.1 节分析了各算法用于特征空间降维时的计算公式及其优缺点,其性能的不同使得一方面在同一应用环境中性能不同。如文献[7]针对英文语料 Ruters-22173 比较了 DF/IG/MI/CHI/TS 五种特征选择方法的性能,实验结果显示:IG 和 CHI 最有效,DF 与 IG、CHI 的性能相近,在需要降低计算复杂度、节省开销时可用于代替 IG 和 CHI;MI 因偏向选择稀有词性能最差。另一方面,同一特征选择算法用于不同应用领域时性能也不同。如文献[8]针对中文网页比较 DF/CHI/IG/MI 四种方法,得到的结论与文献[1]基本相同,且 CHI/IG/DF 能够过滤掉 85% 以上的特征;文献[9]面向旅游领域的文本分类比较了 TF-IDF/ECE/IG/WET/MI 五种特征选择方法,ECE 因既考虑词频又考虑词的出现与类别的关系使得分类效果最好,IG 因考虑单词未出现情况使得性能最差。此外,当训练数据集分布不平衡时,多数特征选择方法倾向于选择高频词,这样对包含样本数较少的类别很不利。

特征选择选取有类别代表性的特征用于分类器分类,其性能还受分类器的影响。文献[7~9]均采用 KNN 分类器为平台进行研究,但文献[10]采用 Naïve Bayes 分类器比较 IG/ECE/WET/OR 等特征选择方法时发现优势率性能最优。文献[11]以人民网新闻语料、Naïve Bayes、文本相似度方法为平台比较各特征选择方法,发现其提出的类别区分词和多类优势率效果最好,IG 和 ECE 其次,DF 效果最差。

综上所述,特征选择算法性能受数据集的文本语种、数据分布信息、分类算法等因素的影响,各个算法各有利弊,不存在某种算法在所有应用领域都是最优的。因此,在本文对中文邮件过滤的特征选择方法的研究中,邮件过滤结果的“二值”性及邮件文本的特殊性,必然使各特征选择算法在邮件过滤中的性能规律与现有文本分类的研究有所不同。

然而,现阶段国内学者专注于研究某一特征选择算法在邮件过滤中的应用及其改进,如文献[12]改进互信息、文献[13]提出基于 Bayes 推理的特征选择方法以提高过滤效果,而对现有各特征选择算法在邮件过滤领域的比较与分析尚未见诸刊出。因此,寻找适用于邮件样本集与分类模型的特征选择方法对提高垃圾邮件过滤具有重要意义。

2 邮件过滤中特征选择方法的评价

2.1 评价平台

2.1.1 评价语料

实验采用 CCERT(中国教育和科研计算机网紧急响应组)公开的中文邮件样本集,其语料保留了邮件原文,邮件格式比较规范,是近年来中文邮件过滤研究的常用语料。该样本集包

含 2005 年 6 月收集的合法邮件 9 272 篇、垃圾邮件 25 088 篇,7 月收集的合法邮件 9 024 篇,垃圾邮件 20 308 篇。

2.1.2　评价工具

实验采用 Visual Studio 2010(C#)开发的邮件过滤平台,该平台的搭建借助于 Weka 工具和 SVM^{light} 工具。

邮件过滤是一个二值分类问题,涉及分词、文本表示、特征选择、分类算法等关键技术。在本平台中实现的关键技术如下:切词部分采用中国科学院的 SharpICTCLAS 切词系统,其理论准确率高达 97.58%,应用广泛[14];文本表示采用向量空间模型,特征选择为 1.1 节提及的七种特征选择算法,以切词后所有词语的 10% 为间隔选取特征用于分类;分类算法采用邮件过滤中效果好、被广泛采用的朴素贝叶斯和支持向量机。

实验配置:Pentium® Dual-Core CPU E5500 @ 2.8 GHz,2 GB 内存,320 GB 硬盘。

2.1.3　评测指标

垃圾邮件过滤的性能评价通常借用文本分类相关指标,即召回率、准确率及 F1 值。考虑到邮件对合法邮件误判的代价敏感性,本文引入合法邮件误判率。

召回率(SR)为

$$SR = \frac{n_{spam-spam}}{n_{spam}} \tag{7}$$

准确率(SP)为

$$SP = \frac{n_{spam-spam}}{(n_{spam-spam} + n_{ham-spam})} \tag{8}$$

F1 值为

$$F1 = \frac{2SR \times SP}{(SR + SP)} \tag{9}$$

合法邮件误判率(HM)为

$$HM = \frac{n_{ham-spam}}{n_{ham}} \tag{10}$$

其中:$n_{spam-spam}$ 为正确识别出的垃圾邮件数,$n_{ham-spam}$ 为合法邮件被误识别为垃圾邮件的数目,n_{spam} 为样本邮件中垃圾邮件的总数。

召回率反映了过滤器识别垃圾邮件的能力,召回率越高,未识别的垃圾邮件数目越少;准确率反映了合法邮件被误判为垃圾邮件的可能性,准确率越高,说明过滤器将合法邮件误判为垃圾邮件的可能性就越小;F1 值为召回率和准确率的调和平均;合法邮件误判率越低性能越好。

2.2　评价方案

特征选择通过降维减少分类器在计算和存储的开销,同时也通过过滤部分噪声以提高分类的准确性,是介于分类器与文档数据之间的一个重要环节,其性能与分类器和文档数据有着直接的关系。特征选择算法从文档数据中选择最有区分能力的特征项,一方面在分类学习过程中与数据分布有着直接的关系,即分类中实际数据的分布与样本数据中的假设是否相符、样本数据分布是否平衡,都对特征选择能否选出适合某分类器作出正确标记的特征集合具有重要影响;另一方面,最有区分能力的特征词的组合,未必能使分类器对文档作出最准确的标记,即一个良好的特征选择算法要服务于分类,要能从样本集中选择出一个适合于分类器作正确标记的特征集合。

邮件过滤作为文本分类的一个应用领域,在考虑到特征选择算法对分类器和数据集适应性的同时,也要体现邮件自身的特点:a)用户可以容忍接收到 10 封垃圾邮件,但却不能容许一封正常邮件被误判为垃圾邮件而被丢弃;b)邮件文本短小,讨论主题多样,相比普通文本数据稀疏性要高;c)无论对邮件服务器还是用户客户端,邮件过滤都对实时性要求比较高。这就要求特征选择算法除需要考虑到其对分类器、数据集的依赖性,还需考虑到邮件过滤自身的特点。

因此,邮件过滤中一个性能良好的特征选择方法应该具备如下特性:a)完全性,特征词语能够确实表示目标内容;b)区分性,根据特征向量,能将目标同其他的文本相区分,且能有力判别合法邮件;c)精练性,特征项集的维数应该尽可能地小;d)对数据集分布、分类器具有较高的适应性;e)时间复杂度低,过滤时间是用户所能接受的。由此,本文对邮件过滤中各特征选择方法的性能评价主要看其是否能够满足上述五个特征。

3　结果分析

性能评价实验采用 5 次交叉验证的方式进行,每次实验的测试集从训练集中随机抽取;性能分析采用 F1 值与合法邮件误判率 5 次交叉验证的均值衡量。为避免评价过程中经过分类器评价的偏差以及训练样本与测试样本数据分布不一致造成的第二次偏差,本文通过训练样本对特征选择函数进行评价。

3.1　分类器的适应性分析

朴素贝叶斯、支持向量机是基于内容的邮件过滤中最常用的效果较好的分类算法,但不同的特征选择算法在其上的分类性能有所不同。这是因为所有分类器都建立在某种假设之上,当分类器模型与服务于分类的特征选择算法选择特征的分布信息相一致时,分类器偏差小,分类的精度就高。因此,通过本实验来找出适合某分类器作出正确标记的特征集合。

实验从 2005 年 6 月的语料中,随机选取合法邮件、垃圾邮件各 2 000 篇作为训练集,并从训练语料中随机抽取 500 篇用于评测。分类器分别采用 Naïve Bayes、SVM,分类的 F1 值与合法邮件误判率 5 次交叉验证的均值如图 1 ~4 所示。

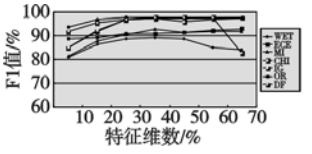

图1　朴素贝叶斯分类器 F1值比较

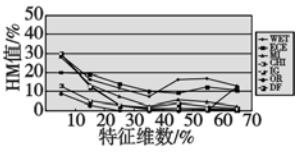

图2　朴素贝叶斯分类器 HM值比较

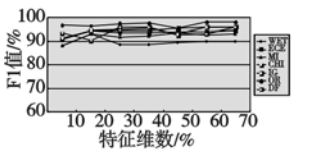

图3　支持向量机分类器 F1值比较

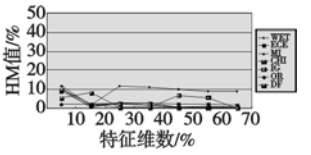

图4　支持向量机分类器 HM值比较

综合考察图 1 ~4 可以得出如下结论:

a)不同分类器、特征选择算法、特征维数下,分类后的垃圾邮件召回率维持在 90% 以上,合法邮件误判率上下波动,其 F1 值维持在 80% 以上。

b)各特征选择算法基本上符合随着特征维数的增多,F1 值越来越高、分类性能越来越好、当达到某种程度时趋于稳定,被误判的合法邮件越来越少的变化趋势。

c)各特征选择算法在选取特征集合总数的 30% ~40% 服务于分类器时,F1 值达到高点,此时垃圾邮件的召回率接近 100% 且合法邮件误判率相对较低。

d)OR、CHI、IG、DF 的性能较好,WET、MI 性能较差。该结论与文献[7,10]的实验结果大体一致,因为邮件过滤本质是一个文本分类问题,但其自身的特殊性又使得其与普通文本分类有所不同,这必然使得其结果相似但又有明显差异。

e)各特征选择算法,在朴素贝叶斯分类器上性能差异较大;而在 SVM 上 F1 值集中在 87.92% ~98%,其间差异较小。

f)在朴素贝叶斯分类器上,OR、DF 能较好地选择适合其分类的特征集合,该结论与文献[10]中相一致。

g)在支持向量机分类器上,OR、CHI 能够较好地提供用于其作出正确标记的特征集合。

针对上述实验结论,结合已有研究成果,本文给出如下分析:

a)各特征选择算法之间在特征项选取时有大量重合特征,如 IG 算法与 ECE 算法在特征项之间有较高相似性。文献[15]实验验证 IG、CHI、WET 等除 MI 之外的特征选择算法所选取的特征集合中相同特征项占 60% 以上。因此,将这些特征选择算法用于分类时,其性能必然是相似的,即在同一分类器上,各特征选算法下分类性能的变化将集中在某个区间内。

b)朴素贝叶斯基于两个假设[15]:(a)各特征之间相互独立,当分类数据集的特征依赖性较强时,分类质量比较差,反之亦然;(b)每个类中的文本都是基于某个混合模型产生的,即混合模型与类别之间存在一一对应关系。但邮件过滤问题并不遵守这个假设,正常邮件与合法邮件通常包含广告、欺诈信息等众多子类别。此外,混合模型与类别之间的一一对应关系很难建立,分类器存在较大偏差。基于统计的特征选择算法在特征独立性假设上与朴素贝叶斯保持一致,但邮件样本集中每个类别下邮件的主题具有多样性,这使得各算法在选取特征时不能很好地照顾到每个主题,即类内的特征分布不平衡性使得各算法性能差异较大。

c)支持向量机是基于统计学习理论的主动学习算法,其基于结构风险最小化原则,通过构造最优超平面对向量进行分类,得到较好的分类准确率[16]。其学习能力独立于特征空间的维数,在解决小样本、非线性问题中效果较好。因此,面对小样本训练的邮件过滤实验,在各特征维数上,支持向量机都表现出较高的 F1 值。正是因为各特征选择函数所选特征项的差异,使得其构造的最优超平面有所差异,最终使得分类效果得以区分。

3.2　数据集依赖性分析

邮件过滤中,数据集关于类别的分布往往是不均衡的,即垃圾邮件的数目往往要远多于合法邮件,CCERT 收集的语料中 2005 年 6 月份的垃圾邮件数目是合法邮件的 2.7 倍、7 月份是 2.25 倍。而在互联网环境中也存在不均衡现象,如 2010 年第 3 季度的平均垃圾邮件量占全球发送邮件的 88%,垃圾邮件是合法邮件数目的 4.8 倍,2010 年 4 季度占 83%,比率为 4.8∶1[17]。由此可见,邮件过滤中数据分布具有不平衡性,且垃圾邮件与合法邮件之间的数目比例是不断变化的。因此,应找到一个性能良好的特征选择算法使其对邮件的不平衡性有一定的适应能力。

本文实验从 2005 年 6 月的语料中,随机选取 1 000 篇合法邮件,依次以 2 000、3 000、4 000、5 000、6 000、7 000、8 000 随机选取垃圾邮件用于分类器学习,并从训练语料中各随机抽取 500 篇用于评测,各比例实验进行 5 次取均值。且为减少分类器对数据分布的影响,本文采用 SVM 作为分类器,其原因在于文献[18]对比分析了 SVM、NB、KNN 等方法在样本分布受控情况下的健壮性及分类效果与数据分布之间关系,得出如下结论:SVM 和 KNN 对样本分布的健壮性要好于 NB 等方法。图 5、6 是随机选取合法邮件 1 000 篇、垃圾邮件 5 000 篇 5 次实验的平均值。

对比图 5 与图 3、图 6 与图 4 可发现:各特征选择算法的分类性能不同程度下降,其中,在图 5 中性能比较好的 OR 和 DF 分类的 F1 值下降幅度较小、合法邮件误判率相对较低。在 30% 特征维数上,OR 的 F1 值仅下降 1.35%,DF 上升 0.88%;40% 特征维数上,OR 下降 0.68%,DF 下降 0.7%。而在合法邮件误判率上,优势率和 DF 较图 3 在大多维数中并没有显著提高。由此,优势率 OR 和 DF 对数据分布不平衡性的适应能力较其他算法更好。

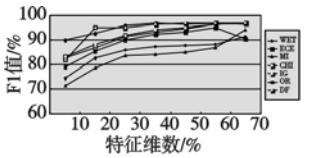

图5　不平衡数据集 F1值比较

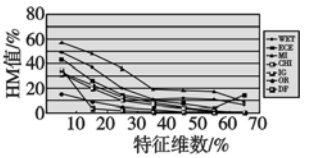

图6　不平衡数据集 HM值比较

针对上述实验结论,论文给出如下分析:DF 根据特征项在类别中出现文档数的百分比进行排序,是典型的类间不相关评估函数,而垃圾邮件文本短小、数据稀疏的特点使得在这样的数据分布条件下特征项的文档频率并没有失去平衡,从而使得 DF 算法下降幅度很小,且在某些维度上性能有所提升。而优势率将合法邮件集定义为目标类,是类间相关函数,但其性能在平衡分布中性能最佳且计算只与词条在类别中的文档频率有关,并不考虑特征项出现的词频,使得其在此情况下性能较好。

3.3　时间复杂度

据 2010 年第四季度中国反垃圾邮件状况调查报告显示,78.7% 的用户认为“收发邮件成功”“及时性”是邮箱性能的重要指标[17]。特征选择作为邮件过滤的重要环节,时间复杂度应尽量低。图 7 是 3.1 节实验中各特征选择算法在选取 50% 特征维数上 5 次实验运行时间的平均值。

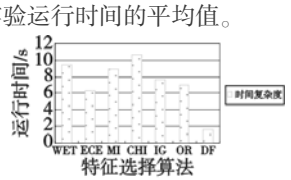

图7　特征选择算法运行时间

图 7 中,DF 的运算复杂度最低,OR、ECE 次之,CHI、IG、WET 计算复杂度相对高,这与文献[10]中关于时间复杂度的描述相一致。其运算时间与 1.1 节中各特征选择算法计算公式中加、减、乘、除的运算有关。

4　结束语

邮件过滤利用特征选择函数从邮件样本集中提取类别区

分能力强的特征项集,再利用特征项集与分类模型对待过滤邮件进行分类。特征选择介于分类器与邮件集合之间,是影响邮件过滤性能的重要因素。其通过降低特征空间维数减少分类器在计算和存储上的开销,同时过滤部分噪声以提高分类的准确性。

本文研究并总结了文本分类中常用的特征选择算法,分析了其优缺点,并重点研究其应用于邮件过滤时的性能差异。本文在朴素贝叶斯、支持向量机两个分类器上分析各特征选择算法对分类器的适应性,得到各算法在两个分类器上的召回率、准确率维持在 80% 以上。其中,OR、CHI、IG、DF 的性能较好,WET、MI 性能较差。并在 SVM 分类器上分析了各算法对数据集分布的依赖性,各算法在邮件样本集分布不平衡时性能均有所下降,OR、DF 下降幅度较小,受不平衡数据分布影响小;在各算法运行时间上,DF 计算最简单,CHI、IG 运行时间较长。综合考虑各因素,OR 在邮件过滤中性能较好。

参考文献:

[1] Symantec Intelligence Report[EB/OL]. (2011-07)[2011-07-22]. http://www. symantec. com/content/en/us/enterprise/other _ resources/b-intelligence_report_07-2011. en-us. pdf.

[2] 中国互联网协会反垃圾邮件中心. 2011 年第一季度中国反垃圾邮件状况调查报告[EB/OL]. (2011-05-24)[2011-07-22]. http://www. anti-spam. cn/ShowArticle. php? id = 11129.

[3] 靖红芳,王斌,杨雅辉,等. 基于类别分布的特征选择框架[J]. 计算机研究与发展, 2009, 46(9):1586-1593.

[4] SEBASTIANI F. Machine learning in auomated text categorization [J]. ACM Computing Surveys,2002,34(1):1-47.

[5] 陆玉昌,鲁明羽,李凡,等. 向量空间法中单词权重函数的分析与构造[J]. 计算机研究与发展, 2002,39(10):1205-1210.

[6] 刘海峰,张学仁,姚泽清,等. 基于类别选择的改进 KNN 文本分类[J]. 计算机科学, 2009,36(11):213-216.

[7] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning (ICML-97). San Francisco: Morgan Kaufmann,1997:412-420.

[8] 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, 2003,29(22):146-148.

[9] 单丽莉,刘秉权,孙承杰. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报, 2011, 43(1):319-324.

[10] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and Naïve Bayes[C]//Proc of ICML'99. San Francisco: Morgan Kaufmann,1999:258-267.

[11] 周茜,赵明生,扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004,18(3):17-23.

[12] 卢杨竹,张新有,祁玉. 邮件过滤中特征选择算法的研究及改进[J]. 计算机应用, 2009,29(10):2812-2815.

[13] 闫鹏,郑雪峰,李明祥,等. 二值分类中基于 Bayes 推理的特征选择方法[J]. 计算机科学,2008,35(7):173-176.

[14] 计算所汉语词法分析系统 ICTCLAS[EB/OL]. (2002-08-16)[2011-07-01]. http://www. nlp. org. cn/project/project. php? proj_id = 6.

[15] 杨创新. 基于机器学习的高性能中文文本分类研究[D]. 广州:华南理工大学,2009.

[16] 王清翔,广凯,潘金贵. 基于支持向量机的邮件过滤[J]. 计算机科学,2007,34(9):93-94.

[17] 中国互联网协会反垃圾邮件中心. 2010 年第四季度中国反垃圾邮件状况调查报告[EB/OL]. (2011-03-25)[2011-07-22]. http://www. anti - spam. cn/ShowArticle. php? id = 11010.

[18] YANG Y, LIU X. A re-examination of text categorization methods [C]//Proc of the 22nd ACM International Conference on Research and Development in Information Retrieval. Berkeley: ACM Press, 1999:42-49.

(上接第 692 页)

参考文献:

[1] MBAlib. Credit rating[EB/OL]. (2011-02-28)[2011-03-20]. http://wiki. mbalib. com/wiki/Credit_Rating.

[2] 佟雨兵,张其善,祁云山. 基于 PSNR 与 SSIM 联合的图像质量评价模型[J]. 中国图象图形学报,2006,11(12):1758-1763.

[3] GENG Guang-gang, JIN Xiao-bo, ZHANG De-xian. Evaluating Web content quality via multi-scale features[C]//Proc of ECML/PKDD 2010 Discovery Challenge. 2010.

[4] ERDELYI M, GARZO A, BENCZUR A A. Web spam classification: a few features worth more[C]//Proc of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality. [S. l.]:ACM Press, 2011.

[5] DAVIS R H, EDELMAN D B, GAMMERMAN A J. Machine-learning algorithm for credit-card applications[J]. IMA J Management Math,1992,4(1):43-51.

[6] VIEDMA H E, PASI G, PORCEL C, *et al*. Evaluating the information quality of Web sites: a methodology based on fuzzy computing with words[J]. Journal of the American Society for Information Science and Technology,2006,57(4):538-549.

[7] ECML/PKDD. Rules: ECML/PKDD 2010 discovery challenge[EB/OL]. (2010-06-25)[2011-03-28]. http://datamining. sztaki. hu/? q = en/DiscoveryChallenge/rules.

[8] PAGE L, BRIN S, MOTWANI R, *et al*. The PageRank citation ranking: bring order to the Web, Technical Report SIDL-WP-1999-0120[R]. [S. l.]:Stanford InfoLab,Stanford University,1999.

[9] QUINLAN J R. C4.5: programs for machine learning[M]. San Mateo: Morgan Kaufmann Publishers, 1993.

[10] BREIMAN L. Bagging predictors[J]. Machine Learning,1996,24(2):123-140.

[11] KEARNS M, VALIANT L G. Learning boolean formulae or factoring, Technical Report TR-1488[R]. Cambridge, MA:Aiken Computation Laboratory,Harvard University,1988.

[12] KEARNS M, VALIANT L G. Cryptographic limitation on learning boolean formulae and finite automata[J]. Journal of ACM,1994,41(1):433-444.

[13] JAERVELN K, KEKAELAEINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Trans on Information Systems, 2002,20(4):422-446.

[14] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proc of International Joint Conference on Artificial Intelligence. Montréal, Québec: Morgan Kaufmann Publishers,1995.