基于一类SVM的不良信息过滤算法改进

丁霄云 刘功申 孟 魁
(上海交通大学信息安全工程学院 上海 200240)

摘 要 互联网的高速发展使得通过网络传输的文件监控和过滤成为一个热门课题。使用传统的基于字符串匹配的算法显然无法满足呈几何爆炸级别的信息增长的监管需求。而使用SVM确实可以提高分类效率,但依然存在维数过大导致存储资源和计算能力浪费的现象。为了有效减少SVM的维数,提出通过使用特征简约对向量机的维数进行约束的一个一类SVM算法改进。实验表明:在选用相同数量的特征词的前提下,改进算法使得不良信息分类和过滤的正确率有明显提高。

关键词 特征简约,一类SVM算法,分类
中图法分类号 TP391 **文献标识码** A

Research and Improvement of Filter Algorithm of Malicious Information Based on One-class SVM

DING Xiao-yun LIU Gong-shen MENG Kui
(School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract The research of monitoring and filtering of the files transporting through internet is getting hotter and hotter now. The traditional algorithm based on string-matched is not able to meet the need of the huge increase of information. Although SVM model can surely improve the efficiency of the classification, the problem that SVM's too large dimension will affect the speed of examine still exists. It also causes a waste of storage space and compute ability. One algorithm was raised by first reducing the dimension by some specific algorithm before classification. The analysis result shows that after the improvement, we can get a more accurate result.

Keywords Feature reduce, One-class SVM, Classification

1 引言

互联网在近几十年迅猛发展,使得网络已经成为信息化的重要组成部分,然而随之而来的却是参差不齐的信息充斥着互联网空间。最为典型的就是,许多敏感的或者包含不良信息的文件在互联网上流传,严重影响了网络内容的安全性。及时发现通过互联网传播的危害公共安全的文件,监控实时信息系统中敏感文件有没有被恶意转发,是目前一个有待深入研究的课题。

那么如何在海量的互联网传输内容中监控过滤有无不良信息呢?普通的基于字符串匹配的算法显然无法满足呈几何爆炸级别的信息增长的监控需求。为了满足需求,研究人员使用分类这一概念来解决这个海量数据挖掘的问题。作为一种优秀的分类算法,一类支持向量机就是在这样的背景下被提出的。

支持向量机(SVM, Support Vector Machine)[1-3]是一种监督式的学习方法。它是建立在统计学习理论的VC维理论和结构风险最小原理基础上的,根据有限的样本信息在模型中的复杂性,在特定训练样本的学习精度和无错误地识别任意样本的能力之间寻求最佳折衷,以求获得最好的推广能力。它在解决小样本、非线性及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中。文献[4]的综合实验,证明了SVM在召回率和准确率上相对于朴素贝叶斯算法有明显优势,并且在两类中分出一类,优势更为明显。文献[5]更是综合比较了决策树算法、朴素贝叶斯算法和K-最近邻算法,论证了SVM算法具有理论完备、适应性强、全局最优等优点。而一类支持向量机则是支持向量机的一个特例,这一算法模型意味着在将数据分到两个不同类的训练过程中,训练样本绝大部分属于其中的一个类,仅有少量的训练数据属于另外一个类。这正符合目前网络中绝大部分是正常数据,仅有少量的可能是敏感或者非法数据的特征。

然而目前使用的SVM在实际应用时却存在一些不足。其中比较突出的一点就是维数爆炸。这是因为文本内包含的词语量非常大,商务印书馆出版的《现代汉语词典》第5版(2005年5月出版)中收录了65000个词,使用如此高维度对存储资源和计算能力是一种严重浪费。国内外研究人员也在针对这个缺点不断改进SVM算法。文献[6]将KNN分类算法结合到SVM中,提高了SVM在临界面的分类准确性,却损失了时间上的性能。文献[7]考虑从核函数的角度来解决这个问题,然而其局限在对于不同的训练样本需要选择不同的核函数,应用面不广。文献[8]提出分层训练(Hierarchical Training)的概念,效果明显,但是会占用更多的计算机内存。

本文受973计划(2013CB329603),国家自然科学基金项目(61272441,61171173)资助。
丁霄云(1988—),女,硕士生,主要研究方向为内容安全;刘功申 男,副教授,硕士生导师,E-mail:lgshen@sjtu.edu.cn(通信作者)。

文献[9]提出的垃圾邮件过滤算法中未考虑到实际系统，正常信息量明显大于不良信息，训练样本不够均匀，仿真效果有进一步提高的空间。

本文将使用特征简约来对 SVM 算法进行维数控制。特征简约[10]是一个十分复杂的问题，目前文本特征选择的算法有基于文档频率(Document Frequency)、信息增益(Information Gain，IG)、开方拟和检验方法(CHI 统计)、互信息(mutual Information)、潜在语义分析 LSA、期望值交叉算熵、文本证据权、term strength(TS)、GSS Coefficient、odds ratio[9]等。

本文将针对如何有效减少 SVM 的维数，提出通过使用特征简约对向量机的维数进行约束的一个一类 SVM 算法改进，分别使用了基于文档频率的特征简约、信息增益以及开方拟和检验方法这 3 种方法与 SVM 结合。然后本文设计实现一个系统，对不良信息进行分类处理，判断输入文本是否属于正常的文本信息。最后文本通过设计若干个对比实验，证明该算法在提高其效率的同时，也保持 3 分类的准确性。实验表明这套系统在运用到针对内容的不良信息的分类中有良好的效果。最后，本文将对 SVM 的算法改进做一下小结和展望。

2 算法思路改进

支持向量机的原理是将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面，分隔超平面使两个平行超平面的距离最大化。两个平行超平面间的距离或差距越大，分类器的总误差越小。虽然传统意义上，SVM 可以接受任何维度的数据进行训练，然而其效率和准确性会受到一定影响。引入特征简约可以解决这一矛盾。

2.1 特征简约

使用特征简约的原因有 3 个。第一是把特征词的个数限制在一个合理的范围内，即降低 SVM 的向量维度，减少计算量。需要进行特征选择的另外一个原因是特征词的选取要求。一般来说，特征词的选择在最初很难保证其准确和全集同分布性，而这是提高分类精准性的基础。因此有效地进行特征筛选是必要的。进行特征选择的第 3 个原因是各个类型的权重不同，例如根据系统应用的不同，在安全级别较高的系统中，宁可将安全的流量判断为不安全的一类，也不能将不安全的内容判断为安全的。

基于文档频率(Document Frequency，DF)、信息增益 (Information Gain，IG)和开方拟和检验方法 (CHI 统计)是常用的 3 种特征简约的方法，各有不同的特点。

2.1.1 基于文档频率的特征简约

在基于文档频率(DF)的方法中，使用特征词在一个类别中出现的文档数量来表示这个特征词与该类别的相关度。在某个类别中越多的文档中出现特征词被保留的可能性越大。显然，文档频率方法的实现最简单、算法复杂度最低。

然而仅仅从词语出现的频率这个角度考虑分类并不能达到理想的效果，有些词虽然出现很多，但是并不能作为区分文本归类的依据，例如，在本次数据库文本的测试数据中，“人”这个名词出现的频率很高，达到 2261 次，但是“人”并不能作为区分正常与非正常样本的依据。这就用到了停词的概念。

停词也叫停用词，是指在自然语言处理中往往出现频率很高，但实际意义又不大的词(例如“人”，“的”，“非常”等等)，去除这些停用词也可以有效地使最具代表性和相关度的特征词被计算到 SVM 的维度中，从而提高算法的准确率。

基于文档频率的特征简约的计算公式如下所示：

$$DF(x_i)=\frac{x_i}{\sum_{i=1}^{n}x_i}$$

其中，x_i 为从训练数据中提取的关键名词中第 i 个词语出现的次数。

实验表明在某些情况下 DF 方法与其他几种方法的分类性能比较接近。

2.1.2 信息增益

信息增益 (IG) 是公认较好的特征选择方法，它刻画了一个词语在文本中出现与否对文本分类的影响。在信息增益中，衡量标准是看特征能够为分类系统带来多少信息，带来的信息越多，该特征越重要。对一个特征而言，我们通过计算系统引入该特征和未引入该特征的前后信息量的差值来定义这个特征给系统带来的信息量，也就是一个词语在文本中出现前后的信息熵之差是否足够大。某个词语的信息增益值越大，说明它对分类的贡献就越大。

对于一个文本系统，它的信息熵可以用以下公式表示：

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log p(x_i)$$

式中，x_i 为第 i 个特征，$p(x_i)$表示 x_i 出现的频率。

因此，对于一类 SVM 中的每个分类 c_i，当包含特征 x_i 时，系统的信息熵为

$$H(c)=-\sum_{i=1}^{n}p(c_i)\log p(c_i)$$

其中，$p(c_i)$为类别 c_i 出现的概率。

而为了计算剔除特征 x_i 的情况下系统的信息熵，则必须使用条件熵，即计算特征 x_i 在被固定下来后系统的信息熵。对于一个文本系统，它的条件熵可以用以下公式表示：

$$H(c|x)=-\left(\sum_{j=1}^{n}p(c_i|x_j)\log p(c_i|x_j)+p(x_i')\sum_{j=1}^{n}p(c_i|x_j')\log p(c_i|x_j')\right)$$

其中，$p(c_i|x_j)$表示特征 x_j 出现时类别 c_i 出现的概率，$p(c_i|x_j')$表示特征 x_j 不出现时类别 c_i 出现的概率。

定义第 j 个特征 x_i 带给系统的信息增益为系统的信息熵减去当特征 x_i 固定时的信息熵，即：

$$IG(x_i)=H(c)-H(c|x)=H(c)+p(x_i)\sum_{j=1}^{n}p(c_i|x_j)\log p(c_i|x_j)+p(x_i')\sum_{j=1}^{n}p(c_i|x_j')\log p(c_i|x_j')$$

式中，$p(x_i)$表示 x_i 出现的频率，$p(x_i')$表示 x_i 没有出现的频率，$H(c)$表示两个类别的信息熵，$p(c_i|x_i)$表示特征 x_i 出现时类别 c_i 出现的概率，$p(c_i|x_j')$表示特征 x_j 不出现时类别 c_i 出现的概率。

2.1.3 开方拟和检验方法

开方检验其实是数理统计中一种常用的检验两个变量独立性的方法。开方检验最基本的思想就是通过观察实际值与理论值的偏差来确定理论正确与否。具体操作的时候常常先假设两个变量确实是独立的，即原假设，然后观察实际值，也可以叫做观察值与理论值。这个理论值是指“如果两者确实独立”的情况下应该有的值的偏差程度，如果偏差足够小，我们就认为误差是很自然的样本误差，是测量手段不够精确导致或者偶然发生的，两者确确实实是独立的，此时就接受原假

设；如果偏差大到一定程度，这样的误差不太可能是偶然产生或者测量不精确所致，我们就认为两者实际上是相关的，即否定原假设，而接受备择假设。那么用什么来衡量偏差程度呢？

假设理论值为 E，实际值为 x。但是如果仅仅使用所有样本的观察值与理论值的差值之和

$$\sum_{i}^{n}(x-E)$$

来衡量当有多个观察值 x_1, x_2, x_3 的时候，很可能 x_1-E, x_2-E, x_3-E 的值有正有负，因而互相抵消，使得最终的结果看上去好像偏差为 0，但实际上每个都有偏差。此时很直接的想法便是使用方差代替均值，这样就解决了正负抵消的问题，即使用

$$\sum_{i}^{n}(x-E)^2$$

为了让均值的大小不影响到方差的大小，进一步修改公式为

$$\sum_{i}^{n}\frac{(x-E)^2}{E}$$

在文本分类问题的特征选择阶段，我们主要关心一个词 t 与一个类别 c 之间是否相互独立。如果独立，就可以说词 t 对类别 c 完全没有表征作用，即我们无法根据 t 出现与否来判断一篇文档是否属于 c 这个分类。但本文所使用的与最普通的开方检验不同普通的开方检验需要一个阈值，当小于这个阈值时，我们认定比较的两者具有相关关系。我们无法设定阈值，因为很难说词 t 和类别 c 关联到什么程度才算是有表征作用，我们只想借用这个方法来选出一些最最相关的即可。因此在本文中，我们使用“词 t 与类别 c 不相关”来做原假设，选择的过程也变成为每个词计算它与类别 c 的开方值，对其从大到小排序，此时开方值越大越相关，取前 k 个就可以了。

定义 4 个变量，它们的关系如图 1 所示。

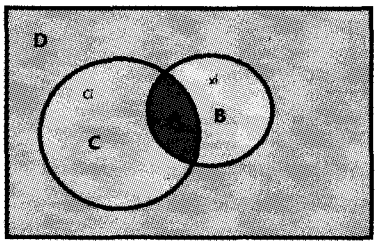

图 1　4 个变量之间的关系

这样，定义开方分布

$$\chi^2(x_i,c)=\frac{(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)}$$

3　系统实现

3.1　系统流程图

在第 2 节的算法理论推导和分析的基础上，本文设计了一套系统来对网络中传送的流量进行基于 SVM 改进算法的分类和过滤，第 4 节中会深入分析改进后的一类 SVM 算法性能与正确率可以获得多大幅度的提升。

本文的系统的输入为从网络流量中实时捕获的文本文件，所有的文本预先已经被标签了是否为不良信息，被用作为训练数据。系统首先对文本进行预处理后，然后分词。在本系统中，只抓取名词和动词作为 SVM 可能使用的特征向量，这是因为对于一篇完整的文章，名词和动词包含更多的信息量，更能体现出一篇文章的表述和类别。然后系统分别使用第 2 节提出的 3 种不同的特征词选择算法选取 SVM 的向量。在得到这些信息后，系统将对所有的文章进行一类 SVM 训练，可以得到各篇文章是否为不良信息的分类。最终系统检查所有文章的判定结果是否正确，计算出分类的正确率与关键词数量以及特征值分类算法的关系。本系统还保存了分类过程中的特征词，用以分析不同特征词选择算法本身的优劣。

系统流程图如图 2 所示。

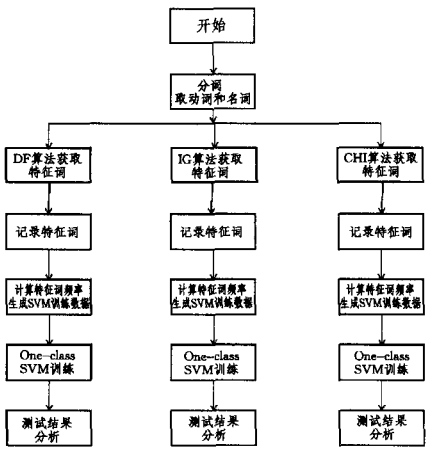

图 2　系统流程图

3.2　系统具体实现

系统的数据采集部分基于 wireshark 的网络协议分析模块，可以将网络中传输的文本保存下来。系统使用的分词系统是中国科学院计算技术研究所研发的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)，它功能强大，主要功能包括中文分词；词性标注；同时支持用户词典。在使用该系统分词后，保存所有的名词和动词作为 SVM 向量的备选词。

系统共使用了 3 种分类算法，下面将一一介绍它们的实现。

3.2.1　DF 算法的实现

基于文档频率的特征简约算法相对来说比较简单，它需要统计出所有数据库中的非停词的词语出现频率，然后根据出现频率进行排序，选择出现最多的若干个作为特征词供 SVM 算法进行维度映射，具体选取的数量根据系统需要的精确率可以选取 500 个、1000 个等，根据系统所能承受的计算量设定。实际操作中，首先对数据库中的每一篇文章进行分词，只保留名词和动词作为特征词的备选词。然后对每一个不存在于停词表中的备选词进行数量统计，并记录到频率表中。以上的这些步骤也是 IG 和 CHI 算法的预处理，在下文的介绍中就不再重复描述。最后，将频率表中出现的备选词根据出现数量的多少进行排序，选择前 n 个作为 DF 算法得到的特征词，算法结束。

3.2.2　IG 算法的实现

IG 算法的本质在于将特征的重要程度量化之后再进行选择，特征能够为分类系统带来多少信息熵，带来的信息越多，该特征越重要。根据第 2 节中的推导，需要对每一个预处理得到的备选词进行信息熵和条件熵的值的计算。此处需要

注意的是本文中由于需要将每个特征给系统带去的信息熵的计算应用到一类 SVM 算法中，而各个类别的出现频率差异很大，因此计算 $P(c_i)$ 时需要考虑不同类别中文档的数量。在对每个备选词引入熵值的计算值后，根据这个值从大到小来排序，选择前 n 个作为 IG 算法得到的特征词，算法结束。

3.2.3 CHI 算法的实现

CHI 算法的本质是观察备选词，以它的观察实际值与理论值的偏差来确定我们的假设“该词出现与该文章属于该类”的正确与否。它使用了分类（在本文中为正常和不良信息两个类别）和是否存在该关键词的 4 种可能出现的关系，对于每种关系出现的数量以及这 4 种关系，通过加权运算后的大小来确定该备选词是否可以作为区分不同类别的依据。具体实现时，计算每个备选词出现在正常类中的篇数、未出现的正常类中的篇数、出现在非正常类中的篇数，以及未出现在非正常类中的篇数。最后通过公式计算得到权值。根据这个值从大到小来排序，选择前 n 个作为 CHI 算法得到的特征词，算法结束。

图 3 展示了 3 种不同的特征词选择算法的流程图。

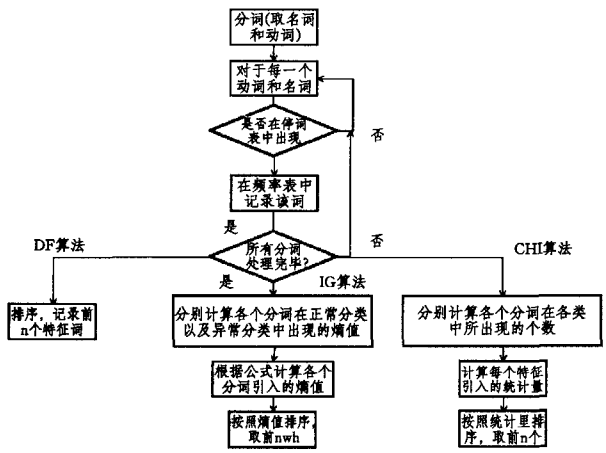

图 3 不同的特征词选择算法

4 实验及分析

为了分析结合特征简约选择对于一类 SVM 算法的优化，本文利用第 3 节描述的系统，设计了如下的一系列实验，对已有数据库的文本信息进行训练，然后进行结果测试。数据库数据包含一些非正常的反动言论、成人笑话等不良信息，此外数据库中还含有正常的文本，其中，正常的文本信息占不良信息文本的 10%。数据库中的文本信息总量大约在 1000 篇。

4.1 特征词个数对 SVM 的影响

针对本文使用的数据库，对于不使用任何特征简约算法的 SVM，其特征值在 13000 个左右。为了在同水平下与 DF、IG、CHI 算法比较准确率，对原始 SVM 的特征值进行随机挑选。而 DF、IG、CHI 算法则使用第 2 节中叙述的算法进行维度的控制，最后获得表 1 显示的数据。

从表 1 中我们可以得出，原始一类 SVM 在只随机使用极少特征量时，效果很差，随着特征量选择数目的增加，效果逐渐变好，而引入了特征值选择算法后则表现出非常良好的效果，在只使用少量特征量时，依然可以获得较好的分类效果。

表 1 特征词个数与不同特征词选择算法的正确率关系

特征词个数	100	500	1000	5000
原始 SVM	56.2%	64.6%	75.7%	75.1%
DF	90.1%	92.1%	93.2%	94.5%
IG	79.1%	82.6%	81.6%	80.1%
CHI	78.2%	80.1%	82.6%	89.3%

最终我们得出的结论是，特征词选择算法的预处理是有效的，经过特征词选择算法，选取的特征词能够使分类准确率得到提高；然而并不是特证词维数越高，分类准确率越高。

4.2 特征词选取的比较

表 2 列出了使用上述 3 种算法前 20 个特征词的情况，通过对特征词的甄选，也可以看出算法的好坏。

表 2 不同特征词选择算法得出的特征词比较

	DF	IG	CHI
各种不同算法的选择排名前 20 个特征词的情况	法	危害	西班牙
	功	适用	诉
	中国	拥有	江
	迫害	侵略	外交部
	大法	签发	签发
	中共	面临	司法
	学员	成立	力
	修	作出	网
	美国	引渡	国家
	弟子	指控	法庭
	看	引用	群体
	修炼	惩罚	罪
	表示	参与	酷刑
	想	流亡	法
	师父	加以	功
	真相	允许	适用
	国家	犯	江泽民
	讲	投机	罗干
	纪元	出现	贾庆林
	经济	判决	薄熙来

通过查看特征词文件可以进一步论证我们的结论：IG 和 CHI 甄选出的特征词集合交集中名词比较多，而 DF 算法中动词比较多，所以 IG 和 CHI 这两种特征词选择算法对文本分类性能的影响也类似，更具有代表性。而通过观察比较，IG 法、CHI 法按类别选取的特征词词汇均含有“适用”这个词。其实这个词是一个噪声词。由此我们可以得出这样的结论：IG 法、CHI 法虽然有抑制高频词噪声和低频词噪声的能力，但是归根结底，这两种方法都是基于频率的统计推断，不能有效抑制全部高频词噪声，如果要提高特征词集合抑制高频词噪声的能力，可能需要借助于基于贝叶斯的统计算法进行推断。

4.3 不同特征简约算法效果的比较

正如第 2 节所描述的，DF 算法实现简单，算法复杂度低，比较通用。IG 算法是公认最好的特征选择算法。而 CHI 算法运用了统计学的原理，在低频词汇里可能出现选取不准的问题。

下面设计了实验，在对数据库文本选取不同数量（分别是 50，100，500，1000，2000，5000，10000）的特征词时，分别对算法进行数据分析。

图 4 显示在选取不同特征词个数的情况下，随机选择特征词以及 3 种特征简约算法的正确率比较。

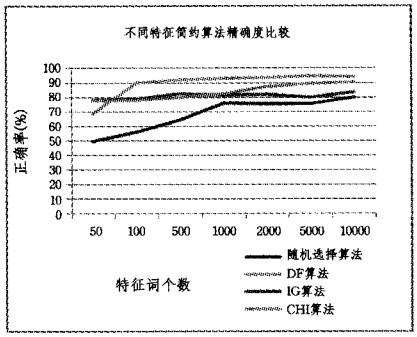

图 4　不同特征简约算法精确度比较

4.3.1　随机选择特征词算法分析

从随机选择特征词算法的特征词个数与正确率关系曲线图中可以看到，在选择的特征词数量较少时，算法几乎无法使用，然后随着选取的特征词的数量增加，正确率也逐渐提升，但是必须达到 10000 这样一个维度才可以有一定的正确率，算法性能不高。

4.3.2　DF 算法数据分析

从 DF 算法的特征词个数与正确率关系曲线图中可以看到，在选择的特征词数量较少时，算法的正确率不尽理想，然后随着选取的特征词的数量增加，正确率也逐渐提升，但是达到 92%左右后就不再提升。由此可见，DF 算法虽然实现简单，算法复杂度低，但是对于要求太高的系统，并不适用。

4.3.3　IG 算法数据分析

从 IG 算法的特征词个数与正确率关系曲线图中可以看到，IG 算法在特征词数量较少时已经表现出了不错的正确率，这正是因为 IG 算法计算了词语带给系统的信息熵，选择对于文本信息量最大的特征词使得系统在只有较少关键词时也能有很好的表现。

信息增益最大的问题还在于它只能考察特征对整个系统的贡献，而不能具体到某个类别上，这就使得它只适合用来做所谓全局的特征选择，也就是指所有的类都使用相同的特征集合，而无法做单独某个类的特征选择。每个类别有自己的特征集合，因为有的词对这个类别很有区分度，对另一个类别则无足轻重。这些词可能将无法被选为特征词。

4.3.4　CHI 算法数据分析

从 CHI 算法的特征词个数与正确率关系曲线图中可以看到，CHI 算法的正确率非常稳定，有较强的稳定性，同时，由于本身算法的特点，其选择特征词的基于统计的特性使得每个关键词都对文本所属分类的判断有很大的贡献。

但 CHI 算法也并非就十全十美了。例如，算法只统计文档中是否出现词 t，却不管 t 在该文档中出现了几次，这会使得它对低频词有所偏袒，因为它夸大了低频词的作用。甚至会出现一些情况：一个词在一类文章的每篇文档中都只出现了一次，其开方值却大过了该类文章在 99%的文档中出现了 10 次的词，其实后面的词才是更具代表性的，但只因为它出现的文档数比前面的词少了“1”，特征选择的时候就可能筛掉后面的词而保留了前者。这就是开方检验著名的“低频词缺陷“。因此开方检验也经常同其他因素如词频综合考虑来扬长避短。如果需要进一步的改进，则可以考虑如果能预先排除一些低频词的出现，绕开 CHI 的低频词带来的不精确性，相信精度会有进一步的提高。

4.4　不良信息过滤效果

本文选取了反动信息和成人小说两种类别的不良信息，分别选取了 500、1000、2000、5000 个特征词使用系统对其进行过滤判断，得到了如图 5、图 6 所示的识别率。

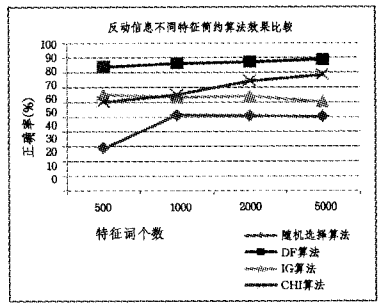

图 5　反动信息不同特征简约算法效果比较

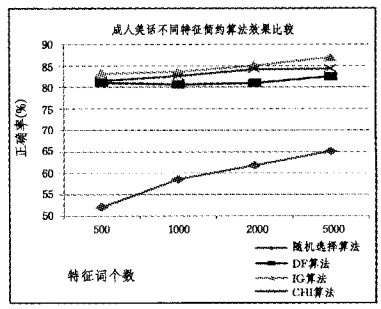

图 6　成人笑话不同特征简约算法效果比较

从图中可以清楚地看到，在使用随机选择特征词时，成人笑话的识别率异常低。这是因为在有些种类的不良信息文本中，词语的选择与正常文本相近，如果随机地取舍这些关键词作为 SVM 的向量则很难将它们与正常文本区分开来。这时，只有借助于特征简约算法来合理地分析备选词，才能够达到一个良好的效果。这也是我们可以在图中得出的结论：DF 算法、IG 算法、CHI 算法的正确率都远高于随机选择算法。

实验表明，在只选取部分特征词作为 SVM 算法的向量时，必须使用特征简约算法预先作处理，否则可能得到很不理想的结果。

结束语　SVM 算法是一种优秀的分类算法。本文在一类 SVM 算法的基础上引入了特征简约的预处理步骤，使得在选用相同数量的特征词的前提下，算法的正确率大大提高。实验表明，IG 算法、CHI 算法的鲁棒性更强，而 DF 算法更简洁明了。相较于普通一类的 SVM 算法，本文的改进使得系统在对不良信息的分类和过滤中的正确性有大幅的提高。

参 考 文 献

[1]　冯长远，普杰信. Web 文本特征选择算法的研究[J]. 计算机应用研究，2005，22(7)

[2]　杨凯峰，张毅坤. 基于文档频率的特征选择方法[J]. 计算机工程，2010，36(17)

[3]　http://tech.ddvip.com/2009-03/1237883850112130_4.html

[4]　詹毅. 朴素贝叶斯算法和 SVM 算法在 Web 文本分类中的效率分析[J]. 成都大学学报，2013，32(1)

[5]　陈燃燃. 基于 SVM 算法的 Web 分类研究与实现[M]. 北京：北京邮电大学，2010

[6]　曹建芳，王鸿斌. 一种新的基于 SVM-KNN 的 Web 文本分类算法[J]. 计算机与数字工程，2010

（下转第 114 页）

型针对 UML 状态机图、类图和 OCL 表达式自动生成测试用例。他提出了将基于边界的覆盖准则与基于数据流、控制流或者基于迁移的覆盖准则结合起来，并且开发了配套的工具 ParTeG[14]，生成的测试用例将满足基于数据流、控制流的覆盖准则和边界覆盖准则。本文的思想与之不谋而合。两者最主要的区别是他将状态图迁移上的守卫条件(OCL 表达式)转换为带输入参数的条件，然后使用这些条件，基于边界覆盖准则生成测试用例。文[12]提出了一种新的准则 MDCC，它将条件覆盖准则(CC)与多维边界覆盖准则(MD)结合。边界值的选择是针对输入参数值的范围采用面向分区的方法(Partition-Oriented testing)。本文提出的准则与他们的思想类似，但是本文主要针对逻辑表达式和边界覆盖准则，边界值的选择运用了基于边界域的 ON-OFF 方法，适用范围更广。

赵瑞莲[8]提出了一种基于谓词切片的字符串测试数据生成方法，即对选定路径上给定的字符串谓词，以相应的动态切片标准为准则，形成关于输入变量的谓词切片；对谓词中变量的每一个字符生成给定字符串谓词边界的 ON-OFF 测试点。本文与其不同，主要针对逻辑表达式，重点是逻辑边界覆盖测试准则的提出。

T-VEC Tester[16]是一个基于模型的自动化软件测试工具。T-VEC 也注重于系统的边界测试，倡导要覆盖软件的所有边界(We Cover All Boundaries)。这也说明了边界测试的自动化越来越引起人们的重视。

结束语 针对边界值分析很少应用于自动化测试的情况，本文将逻辑覆盖准则与边界值覆盖准则结合，提出了一系列基于模型的逻辑边界覆盖准则。基于这些准则可产生具体的包括测试数据和预期输出的测试用例。生成的测试用例既能满足相应的逻辑覆盖准则又能检测到系统的边界情况，并能控制从形式化模型中生成测试用例的数量。对应的工具原型已经实现。下一步的研究主要是对本文提出的测试准则进行度量与评估，并希望将测试工具原型集成到一些主流的测试工具中。

参 考 文 献

[1] Amman P, Offutt J. Coverage criteria for logical expressions[C]// Stephanie K, ed. Proc. of the 14th Int'l Symp. on Software Reliability Engineering. Washington: IEEE Computer Society Press, 2003: 99-107

[2] Kosmatov N, Legeard B, Peureux F, et al. Boundary coverage criteria for test generation from formal models[C]// Software Reliability Engineering, International Symposium. 2004: 139-150

[3] 刘玲，缪淮扣. 对逻辑覆盖软件测试准则的公理化评估[J]. 软件学报，2004，15(9)：1301-1310

[4] 钱忠胜，缪淮扣. 基于规约的若干逻辑覆盖测试准则[J]. 软件学报，2010，21(7)：1536-1549

[5] Bouquet F, Legeard B, Vacelet N, et al. Faster Analysis of Formal Specifications[C]// Proceedings of the 6-International Conference on Formal Engineering Methods (ICFEM'04), LNCS. Seattle, USA: Springer Verlag, November 2004

[6] Legeard B, Peureux F, Utting M. Automated Boundary Testing from Z and B[C]// Proceedings of the International Conference on Formal Methods Europe (FME'02), volume 2391 of LNCS. Copenhagen, Denmark: Springer Verlag, 2002: 21-40

[7] 赵瑞莲. 软件测试方法研究[D]. 北京：中国科学院，2001

[8] Jorgensen P C. 软件测试[M]. 韩柯，杜旭涛，译. 北京：机械工业出版社，2003：70-75

[9] 刘畅，王轶辰，刘斌，等. 软件边界组合测试的典型案例分析[J]. 计算机工程与应用，2009，45(20)：74-77

[10] Bouquet F, Legeard B. Reification of executable test scripts in formal specification-based test generation: the Java Card transaction mechanism case study[C]// Proceedings of the International Conference on Formal Methods Europe (FME'03), volume 2805 of LNCS. Pisa, Italy: Springer Verlag, 2003: 778-795

[11] Caritey N, Gaspari L, Legeard B, et al. Specification-based testing-Application on algorithms of Metro and RER tickets (confidential)[R]. Technical Report TR-03/01. LIFC-University of Franche-Comt'e and Schlumberger Besancçon, 2001

[12] Weißleder S, Sokenou D. Automatic Test Case Generation from UML Models and OCL Expressions[C]// Testing of Software-From Research To Practice (associated with Software Engineering 2008). February 2008

[13] Weißleder S, Schlingloff B-H. Quality of Automatically Generated Test Cases based on OCL Expressions[C]// ICST. April 2008

[14] Weißleder S. ParTeG (Partition Test Generator)[OL]. http://parteg. sourceforge. net. Oct 2009

[15] Weißleder S. Test models and coverage criteria for automatic model-based test generation with uml state machines[D]. Präsident der Humboldt-Universität zu Berlin. 2010

[16] T-Vec: We Cover All Boundaries [OL]. http://www. t-vec. com/

(上接第 90 页)

[7] Maji S. Efficient Classification for Additive Kernel SVMs[J]. Pattern Analysis and Machine Intelligence, 2013, 35(1)

[8] Erdmann M, Nguyen D D. Hierarchical Training of Multiple SVMs for Personalized Web Filtering[C]// PRICAI 2012: Trends in Artificial Intelligence. 2012

[9] Maldonado S, L'Huillier G. SVM-Based Feature Selection and Classification for Email Filtering[M]. Pattern Recognition-Applications and Methods, 2013

[10] 许高建. 基于 Web 的文本挖掘技术研究[J]. 计算机技术与发展，2007，17(6)

[11] 申红，吕宝粮. 文本分类的特征提取方法比较与改进[J]. 计算机仿真，2006，23(3)

[12] 闭乐鹏，徐伟，宋瀚涛. 基于一类 SVM 的贝叶斯分类算法[J]. 北京理工大学学报，2006，26(2)

[13] 刘文，吴陈. 一种新的中文文本分类算法-One ClassSVM—KNN 算法[J]. 计算机技术与发展，2012

[14] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[M]. Machine Learning-International, 1997

[15] Manevitz L M. One-class svms for document classification[J]. The Journal of Machine Learning Research, 2002, 2: 139-154

[16] Chang C-C, Lin C-J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3)

[17] Li Wen-kai. A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2011, 40(2): 717-725