基于支持向量机的邮件过滤

王清翔　广　凯　潘金贵
（南京大学计算机软件新技术国家重点实验室　南京 210093）

摘　要　随着万维网的兴起和电子邮件的快速发展，大量的垃圾电子邮件也随之在互联网上泛滥.电子邮件过滤就是要在大量邮件中过滤出垃圾邮件，帮助用户找到所需的邮件。本文讨论了基于机器学习方法实现垃圾邮件过滤的原理，提出一种改进的基于支持向量机的邮件过滤技术，该方法使用互信息度函数，结合 Z-测试进行特征选择，使用 SVM(支持向量机)构造分类超平面来进行文本分类。实验表明，提高了中文邮件过滤的准确性。

关键词　支持向量机，文本分类，邮件过滤，互信息 Z-测试

Classify E-mails by Support Vector Machine

WANG Qing-Xiang　GUANG Kai　PAN Jin-Gui
(State Key Lab for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract　With the rapid explosion of unsolicited bulk e-mail, also known as "spam", has generated a great need for reliable anti-spam e-mail filters. Mail filtering technique is to used to detect the spam and find what's useful. We discuss the theorem to implement anti-spam, and propose an improved anti-spam mechanism based upon Support Vector Machine, mutual information and Z-test. We demonstrate the performance and positive results of the algorithm on a challenging problem.

Keywords　SVM, Text classification, Spam filter, Mutual information, Z-test

电子邮件是一种半结构化的数据。它由非结构化数据和结构化数据两部分组成。非结构化的数据包括主题和正文等域，这些域允许各种形式的自然语言。对非结构化部分的处理要用到文本分类的技术。文本分类在自然语言处理(NPL)和信息获取(IR)中占有重要地位。文本分类的一种应用就是反垃圾邮件过滤，阻止那些不期而至的商业邮件。近年来，应用机器学习进行文本分类和垃圾邮件过滤的研究进展很快，其中包括基于规则分类[1]、朴素贝叶斯算法[2]、基于记忆的学习算法[4]、决策树[5]、支持向量机[6]和不同学习算法的结合[7]。本文介绍了 SVM 的原理，提出一种改进的基于 SVM 的邮件过滤技术，该方法使用互信息度函数，结合 Z-测试进行特征选择，然后使用 SVM 构造分类超平面来进行文本分类。

1　支持向量机

支持向量机是上世纪 90 年代最早由 V. Vanpik 提出的一种通用学习算法[8]，它通过构造最优超平面对向量进行分类，在解决小样本学习、非线性及高维模式识别问题中表现较好。

1.1　线性支持向量机

样本 x 为 d 维空间中的向量，y 为样本 x 的类标签。d 维空间中 N 个已经标记类别的训练向量记为 $(x_1, y_1), \cdots, (x_N, y_N)$。首先讨论两类的情况，即 y_i 标记为 $+1$ 或 -1。如果训练集是线性可分的，我们可以找到权重向量 w^*，使得 $||w^*||$ 最小，且满足：

$$\begin{aligned} w^* \cdot x_i - b &\geqslant 1 \quad \text{当 } y_i > 1 \\ w^* \cdot x_i - b &\leqslant -1 \quad \text{当 } y_i < 1 \end{aligned} \tag{1}$$

式(1)对应着两个平行的分类超平面，如图 1 所示。最优分类超平面满足边界(margin，最近的点到分类超平面的距离)最大。满足式(1)的向量，称为支持向量。

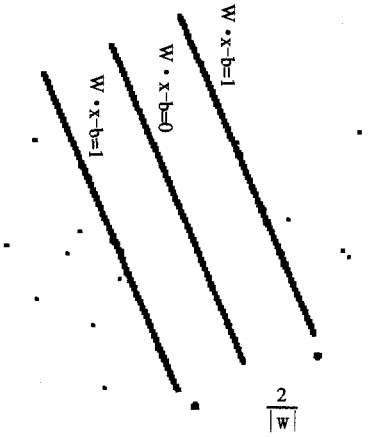

图 1　分类超平面

引入松弛变量 $\xi_i (i=1, \cdots, N)$，寻找最优分类超平面可表示为寻找式(2)的最小值：

$$F(w^*) = \| w^* \|^2 + C\sum_{i=1}^{N} \xi_i \tag{2}$$

需满足约束：

$$y_i(w^* \cdot x_i - b) \leqslant 1 - \xi_i \tag{3}$$

式(1)第一项用于最大化边界(margin)，第二项使得具有容错能力，C 为常数，对松弛变量进行折衷。这是一个二次优化问题，可以转换成对其对偶函数的求解：

王清翔　硕士研究生，主要研究领域为嵌入式系统，机器学习；**潘金贵**　教授，博士生导师，主要研究领域为多媒体信息处理、多媒体远程教育系统。

$$\min\left[\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j(x_i\cdot x_j)a_ia_j-\sum_{i=1}^{N}a_i\right] \tag{4}$$

满足约束：

$$\sum_{i=1}^{N}y_ia_i=0,\ 0\leqslant a_i\leqslant C \tag{5}$$

这是一个典型的二次规划问题，已经有高效的解决方法[12]。对于未知向量 x 的判别函数可表示为：

$$y=\mathrm{sgn}[\sum_{i=1}^{N}a_iy_i\sum_{j=1}^{N}(x_i\cdot x_j)(x_j\cdot x)+b] \tag{6}$$

1.2 非线性支持向量机

最初的超平面算法是线性的，1992 年 Vapkin 等提出了用核函数的方法解决非线性分类面的问题，此解决方法与线性支持向量机十分类似，只是所有点乘都被代之以非线性的核函数。即使原始的向量空间线性不可分，也可以通过转化函数 ϕ 转化到高维线性可分的空间，函数 $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$ 称为核函数。式(4)可变为：

$$\min\left[\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_jK(x_i\cdot x_j)a_ia_j-\sum_{i=1}^{N}a_i\right] \tag{7}$$

满足约束 $\sum_{i=1}^{N}y_ia_i=0,0\leqslant a_i\leqslant C$

相应的判断函数变为

$$y=\mathrm{sgn}[\sum_{i=1}^{N}a_iy_i\sum_{j=1}^{N}K(x_i\cdot x_j)K(x_j\cdot x)+b] \tag{8}$$

1.3 多标记支持向量机

上文所述算法限于对双标记向量的分类，要处理 k-标记支持向量机，通常的做法是 1-rest，对每类都构建一个超平面 $y=w_i^*\cdot x+b_i(i=1,\cdots,k)$，来区分向量 x 是否属于类 i。决策函数可以表示为：

$$C(x)=\max(w_i^*\cdot x+b_i)\quad i=1,\cdots,k \tag{9}$$

其几何含义是，超平面与 x 正向距离最远的类，被判定为向量 x 的类标签。训练方法可以参考文[11]。

2 邮件预处理以及分类实现

邮件处理通常由三部分构成：文本抽取、信息预处理，训练测试过程，分类过程。邮件分类模型如图 2 所示。

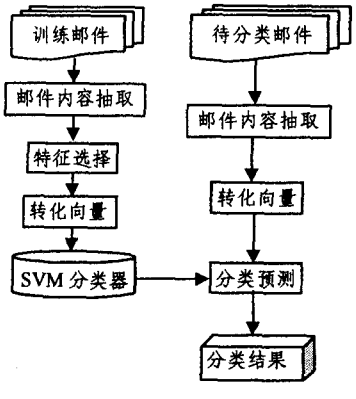

图 2 分类超平面

2.1 特征表示

将文档 d 转化成向量的过程中，一个词对应向量中的一维。设向量为 x。向量第 i 维的值为 x_i，对应的词为 w_i。特征表示的常用方法有以下几种[6]：

TF(Term Frequency)：x_i 为单词 w_i 出现的次数

TF-IDF(Term Frequency - Inverse Document Frequency)：$TF(w_i)$ 为单词 w_i 出现的次数，D 为训练集中所有文档的总数，D_w 为单词 w_i 曾在其中出现过的文档数量。a 为常数，在我们实验中取 0.01。

$$x_i=\frac{TF(w_i)\times\log(\frac{D}{D_{w_i}}+a)}{\sqrt{\sum_{k=1}^{n}TF^2(w_i)\times\log^2(\frac{D}{D_{w_i}}+a)}} \tag{10}$$

2.2 特征选择

特征表示完成以后，如果把所有单词均作为特征，则文本向量的维度相当大，而且有些维度对分类起很小作用或不起作用。我们采用评估互信息度的方式来进行特征选择。互信息是测量搭配强度的物理量，公式如下：

$$I(x_i,y_i)=P(x_i,y_i)\times\log\frac{P(x_i,y_i)}{P(x_i)\cdot P(y_i)}=\frac{freq(x_i,y_i)}{N}\log\frac{N\times freq(x_i,y_i)}{freq(x_i)\times freq(y_i)} \tag{11}$$

其中 N 为训练集文档个数，$freq(x_i)$ 为出现 w_i 的文档个数，$freq(y)$ 为类别为 y 的文档个数，$\mathrm{freq}(x_i,y)$ 为类别 y 且出现单词 w_i 的文档个数。

如果直接保留互信息大于某阈值的单词作为特征，则对于不同的互信息度分布，设定的阈值也不同，难以确定阈值。假设互信息度符合正态分布，我们通过 Z-测试[12,13]将互信息调整为标准正态分布，这样就可以确定同一的阈值求解。计算公式如下：

$$z_{ij}=\frac{I(x_i,y_i)-E_j}{\sqrt{\mu_j}} \tag{12}$$

其中 E_j 为互信息均值，u_j 为方差。

$$E_j=\frac{1}{n}\sum_{i=1}^{n}I(x_i,y_j) \tag{13}$$

$$\mu_j=\frac{1}{n}\sum_{i=1}^{n}(I(x_i,y_j)-E_j)^2 \tag{14}$$

转化为标准正态分布以后，根据统计理论，z_{ij} 绝大部分分布在[−3,+3]。可以在该范围设定一个比较大的阈值 $thrs$，然后每次递减 Δt，通过 10 倍交叉测试来寻找正确率最高的阈值。

3 实验

实验语料库采用 CERNET 共享的 2005 年 6 月份的电子邮件(http://www.ccert.edu.cn/spam/sa/datasets.htm)，已经人工分类。从中随机选出正常邮件 1852 封，垃圾邮件 1152 封进行实验。

首先采用 MIME parser 模块抽取邮件的文本，然后进行分词、特征选择、向量转化、最后训练得到模型。向量转化采用 TF-IDF 方法，评价采用 10 折交叉测试方法，从语料库中选 9 成做为训练集，1 成做测试集，如是重复 10 次，得到查准率，查全率。

表 2 SVM 使用 RBF 核函数

	垃圾邮件		正常邮件	
	查准率	查全率	查准率	查全率
不特征选择	94.64%	93.31%	94.00%	96.00%
MI + Z-Test	98.28%	94.12%	95.87%	98.81%

表 3 SVM 使用线性核函数

垃圾邮件	正常邮件			
	查准率	查全率	查准率	查全率
不特征选择	97.73%	96.01%	97.02%	98.31%
MI + Z-Test	98.28%	94.12%	95.87%	98.81%

(下转第 116 页)

$e_{ij}^{ba}=f_{ba}(\langle A,O,S_i^a,G_j^a,T^m,T^c,C\rangle)$ (6)

每个合计单元格的值由以下多元组决定：

$e_{ij}^{aa}=f_{ba}(\langle A,O,S_i^a,G_j^a,T^m,T^c,C\rangle)$ (7)

多维联机分析系统为每个分析方法提供界面一致的接口。系统通过该接口将用户的分析要求传递给每个方法的实现过程，并由每个实现过程根据以上各式及本身的实现逻辑计算附加单元格的值。

4.3 XOLDAS 内嵌的分析方法

多维联机分析系统的方法库中包含多种的分析方法分别为：基本情况分析、结构分析、总量分析、支持度分析、可信度分析、改善度分析、相关分析和差异显著性检验。在此由于篇幅限制就介绍二种主要分析方法的算法："总量分析"和"支持度分析"[3,4]。

4.3.1 总量分析

说明在所选范围内同时兼有主词项$_i$和宾词项$_j$性质的分析对象的指标值，分析表如表 2 所示。主词和宾词分别指定为业务空间中的一个维，附加宾词项为行合计，附加主词项为列合计，合计单元格的值为所有指标值的总计。

表 2 总量分析表的一般形式

宾词 \ 主词	宾词项$_1$	…	宾词项$_n$	行合计
主词项$_1$	指标值$_{11}$	…	指标值$_{1n}$	行合计项$_1$
…	…			…
主词项$_m$	指标值$_{m1}$		指标值$_{mn}$	行合计项$_m$
列合计	列合计项$_1$	…	列合计项$_n$	总计项

表中：$指标值_{ij}=f_s(\langle O,S_i^b,G_j^b,T^c,C\rangle)$

$行合计项_i=\sum_{j=1}^{n}指标值_{ij}$　　$列合计项_j=\sum_{i=1}^{m}指标值_{ij}$

$总计项=\sum_{i=1}^{m}\sum_{j=1}^{n}指标值_{ij}$

4.3.2 支持度分析

说明在所选范围内，同时兼有主词项$_i$和宾词项$_j$性质的分析对象的指标值，占全部指标值合计的比例。其分析表如表 3 所示。

表 3 支持度分析表的一般形式

宾词 \ 主词	宾词项$_1$	…	宾词项$_n$	行合计
主词项$_1$	指标值$_{11}$	…	指标值$_{1n}$	P_1
…	…			…
主词项$_m$	指标值$_{m1}$		指标值$_{mn}$	P_m
列合计	Q_1	…	Q_n	1.00

其中：$指标值_{ij}=\dfrac{f_s(\langle O,S_i^b,G_j^b,T_c,C\rangle)}{f_s(\langle O,T^c,C\rangle)}$

$P_i=\sum_{j=1}^{n}指标值_{ij}$　　$Q_j=\sum_{i=1}^{m}指标值_{ij}$

$总计项=\sum_{i=1}^{m}P_i=\sum_{j=1}^{n}Q_j=1.00$

结束语 数据分析算法在 XOLDAS 的方法库中起重要作用，业务空间和方法库的建立使用户在分析时只需将全部精力集中到所要分析的业务问题上，从业务空间中选择适当的维表达自己的分析要求，系统根据用户的分析要求自动完成数据库连接、数据抽取与计算、结果显示等操作，将分析的结果展示给最终用户。

参 考 文 献

1 Nguyen T B, Min Tjoa A, Wagner R R. An Object Oriented Multidimensional Data Model for OLAP. In: Proc. WAIM. Shanghai, China, June 2000. 130～132
2 肖昭媛. 多维联机数据分析模型和系统设计方法. 上海海运学院学报，2003，24(4)：46～48
3 袁卫，何晓群，金勇进，等. 新编统计学教程. 北京：经济科学出版社，1999
4 肖昭媛. 统计学原理与应用. 上海：上海交通大学出版社，2002
5 经霄. OLAP 系统平台—多维联机分析系统 XOLDAS 的分析与设计[D]. 上海：上海海事大学，2004

(上接第 94 页)

表 1 是 SVM 使用线性核函数的结果，表 2 是 SVM 使用 RBF 核函数的方法。可以看出使用 MI 和 Z 测试提高了查准率和查全率。由于 CERNET 语料库在文本向量空间相对集中，因此查准率和查全率比通用的英文语料库高。

结束语 本文讨论了基于机器学习方法实现垃圾邮件过滤的原理，提出一种改进的基于支持向量机的邮件过滤技术，该方法使用互信息度函数，结合 Z-测试进行特征选择，使用 SVM 构造分类超平面来进行文本分类。实验表明，提高了中文邮件过滤的准确性。下一步工作：针对海量语料库中的样本选择问题，实现算法找出最具代表性的训练集以及中文邮件的预处理研究，针对中文特点改进分词及向量转化以及。

参 考 文 献

1 Cohen W W. Learning rules that classify e-mail. In: the AAAI Spring Symposium on Machine Learning in Information Access, Standford, CA, AAAI Press, 1996. 18～25
2 Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the AAAI Workshop, Madison Wisconsin. AAAI Press: [Technical Report WS-98-05]. 1998. 55～62
3 Androutsopoulos, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos C D, Stamatopoulos P. Learning to filter spam e-mail: A comparision of a Naive Bayesian and a memory-based approach. In: H. Zaragoza, P. Gallinari, and M. Rajman, eds. Proc. Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD 2000), Lyon, France, 2000. 1～13
4 Rennie J D M. ifile: An application of machine learning to e-mail filtering. In: Proc. KDD-2000 Workshop on Text Mining, Boston. MA, 2000
5 Carreras X, Màrquez L. Boosting trees for anti-spam email filtering. In: Proc. International Conference on Recent Advances in Natural Language Processing(RANLP-01), Tzigov Chark, Bulgaria, 2001
6 Drucker H, Wu Donghui, Vapnik V N. Support vector machines for spam categorization. IEEE Trans. On Neural Networks, 1999, 10(5): 1048～1050
7 Sakkis G. Androutsopoulos I, et al. Stacking classifiers for anti-spam filtering of e-mail. In: L. Lee and D. Harman, eds. Proc. 6th Conference on Empirical Methods in Natural Language Processing(EMNLP 2001), Pittsburgh, PA, Carnegie Mellon University, 2001. 44～50
8 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer, 1995
9 Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers [A]. In: Haussler D, ed. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory [C]. ACM Press, 1992. 144～152
10 Platt J C. Sequential minimal optimization: A fast algorithm for training support vector machictor machines. In Advances in Kernel Method: Support Vector Learning In: Scholkopf, Burges, and Smola, eds. Cambridge, MA: MIT Press, 1998. 185～208
11 Weston J, Watkins C. Support vector machines for multi-class pattern recognition. In: Proceedings of the 6th European Symposium on Artificial Neural Networks (ESANN), 1999
12 Smadja F. Retrieving collocation from text: Xtract. Computational Linguistics, 1993, 19(1): 143～175
13 李涓子，等. 语言模型一种改进的最大熵算法及其应用. 软件学报，1999，10：258～263