一种改进的支持向量机邮件分类器*)

熊忠阳　杜圣东　张玉芳
(重庆大学计算机学院　重庆 400044)

摘　要　在实际的邮件过滤应用中，由于垃圾邮件本身的一些因素，像传统的支持向量机分类模型把一个邮件样本明确地归为某一类就很容易出错，而以一定概率的输出判断是否属于某一类则较为合理。根据这种思想，本文在传统支持向量机邮件分类器基础上，提出了一种分类器优化思想，通过对分类输出进行概率计算，并对概率的阈值进行判断，从而确定邮件所属类别。实验证明这种方法是有效可行的。

关键词　支持向量机，文本分类，邮件过滤

An Improved E-mail Classifier Based on Support Vector Machine

XIONG Zhong-Yang　DU Sheng-Dong　ZHANG Yu-Fang
(Dept of Computer Science, Chongqing University, Chongqing 400044)

Abstract　In the real spam-filtering environment, because of the complicated factor of spam itself. It's easy to make mistakes just as the traditional support vector machine classifiers model doing-assigning an e-mail example to a class specifically. However, assigning an e-mail example to a class according to its probability output is a reasonable solution to the problem. According to the theory, we put forward an optimising idea based on the traditional SVM classification model. By computing the probability of output class, and judging the threshold of the probability, we can make sure which class the input email example belongs to. The experiment has proved that this method is efficient and feasible.

Keywords　Support vector machines, Text classification, E-mail filtering

1　引言

随着 Internet 技术的高速发展，网络已经作为一种新兴传播媒介流行起来，电子邮件更是成了人们工作生活必不可少的交流工具。但是，垃圾邮件问题日益严重，近来的调查显示，93%的被调查者都对他们接收到的大量垃圾邮件非常不满。垃圾邮件不仅耗费网络带宽和计算机时空开销，而且造成了很严重的安全问题。因此，反垃圾邮件工作就显得十分必要。迄今为止，垃圾邮件过滤做为一种主要的反垃圾邮件技术，还算不上成熟。由于垃圾邮件本身的一些因素，不同环境下能高效地检测出垃圾邮件是很难的。垃圾邮件的判定有很强的主观性，不同用户对同一邮件的判断结果可能会存在差异。基于以上原因，对垃圾邮件过滤技术进行研究具有很重要的现实意义。

统计学习理论[1]是一种专门研究小样本情况下机器学习规律的理论，支持向量机(Support Vector Machine，以下简称 SVM)作为一种新的数据挖掘技术，是在统计学习理论的基础上发展起来的一种新的机器学习算法，由于其基于结构风险最小化原则，能有效地解决过学习问题，具有良好的推广性能，即是由有限的训练样本集得到小的误差，能够确保对独立的测试样本集仍保持较小的误差。这些优良特性使 SVM 成为了继人工神经网络(ANN)、模式识别之后的又一研究热点。最有代表性的是美国邮政手写数字库识别研究成功地应用了 SVM，在其它应用领域比如人脸检测、语音识别、模式识别、图像处理、文本分类、邮件过滤等方面也取得了大量的研究成果。

本文主要分析和探讨了支持向量机在邮件分类过滤中的应用。首先对垃圾邮件过滤技术涉及到的主要方法和环节做了阐述，重点放在基于分类方法的邮件过滤；接下来对支持向量机用于邮件分类遇到的问题进行了分析：怎样构造 SVM 邮件分类器，并在传统分类器的基础上提出了自己的改进方法；在本文实验中，对改进后的 SVM 邮件分类器和传统的 SVM，NaiveBayes 用于邮件分类的实验效果进行了比较；最后是本文的结论陈述。

2　垃圾邮件过滤技术

邮件过滤是目前反垃圾邮件用到的主要技术[2]。垃圾邮件的自动过滤主要有基于规则和基于概率统计的两种方法(这两种方法都是基于邮件内容的过滤[3])。基于规则的方法主要有决策树、Boosting、粗糙集三种，这类方法一般是利用包含了各种约束条件的规则集做决策，即根据电子邮件是否匹配预先定义的规则来决定是否过滤邮件，如常用的关键词过滤；基于概率统计的邮件自动过滤方法主要有 Bayes、kNN、SVM、Rocchio、Winnow 几种，这类方法的研究已成为一种主要的趋势，以朴素贝叶斯算法为例，它具有方法简单、运算速度快、分类精确度高等优点，被广泛应用于邮件过滤领域。

不管是基于规则的方法还是基于概率统计的方法，垃圾

*)本文获得重庆市科委自然科学基金(基金号：CSTC2006BB2021)的资助。**熊忠阳**　博士生导师，主要研究领域为数据挖掘、数据库、并行计算、网络信息处理；**杜圣东**　硕士研究生，主要研究领域为数据挖掘；**张玉芳**　副教授，主要研究领域为数据挖掘、数据仓库、网络信息处理和远程教育。

邮件过滤问题的本质是一个文本分类[4]问题(如图 1 虚线框),即根据一定的分类算法和预定义的类别标号来确定待分类文本的类别。邮件文本分类主要有如下三个环节:邮件文本的表示、特征选择、分类器训练。邮件文本的表示和特征选择在文[5]中有详细分析和研究,这里不做过多描述,本文重点放在如何构造比较优化的 SVM 邮件分类器这一环节。

3 基于 SVM 的邮件分类方法

3.1 支持向量机

V. Vapnik 提出的支持向量机理论[6]最基本的思想之一是结构化风险最小化原则(Structural Risk Minimization, SRM),这要优于传统的经验风险最小化原则(Empirical Risk Minimization, ERM)。传统的支持向量机是通过构造一个最优超平面,对两类问题进行分割。所谓最优分类面就是要求分类面不但能将两类正确分开(保证经验风险最小),而且使分类间隔最大。下面以对 m 个训练样本:(x_1,y_1),(x_2,y_2),…,(x_m,y_m)求解最优分类超平面为例。主要是求解系数 w 和 b,使超平面 $(w\cdot x)+b=0$ 达到分类误差小、推广能力强的要求,必须满足最优分类超平面的条件:

$$y_i[(w\cdot x_i)]+b]\geqslant 1,(i=1,2,\cdots,m)$$

$$\min_{w}\phi(w)=\|w\|^2$$

根据最优化理论,利用 Lagrange 函数将上面问题转化为求解标准型二次规划问题:

$$\max W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$\text{s. t.}\sum_{i=1}^{m}\alpha_i y_i=0,\alpha_i\geqslant 0,(i=1,2,\cdots,m)$$

求解上式,得最优分类决策函数为

$$f(x)=\text{sign}\{\sum_{\alpha_i>0}\alpha_i y_i K(x_i,x)-b_0\}$$

b_0 可由约束条件 $\alpha_i[y_i(w^T x_i+b)-1]=0$ 求解,α_i 不为零的样本即为支持向量。

对于非线性二元分类,则通过某种事先选择的非线性映射(即核函数),将输入向量 x 映射到一个高维特征空间中,然后在这个高维空间中构造最优分类超平面。这种方法通过核函数做升维处理,避免了在高维特征空间中进行复杂的运算。

3.2 SVM 邮件分类器构造及优化

用 SVM 进行邮件分类时主要分两个阶段(训练和分类):

(1)训练阶段

Step1:建立邮件训练样本集 $(x_1,y_1),\cdots,(x_l,y_l)$,$x_i\in R^n$,$y_i\in\{-1,+1\}$。

Step2:选择合适的核函数及核参数,作为高维特征空间在低维输入空间的一个等效形式。

Step3:输入邮件训练样本规范化,将输入数据限定在核函数要求范围之内。

Step4:构造核矩阵 $H(n,n)$。

Step5:最大化下式,以求解拉格朗日系数 a

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j),0\leqslant a_i\leqslant C,$$

$$\sum_i a_i y_i=0$$

Step6:找出支持向量 SV,求解分类超平面系数 b。

Step7:建立训练邮件集的最优决策超平面,完成训练。

(2)分类阶段

Step1:装入 SVM 训练阶段的邮件样本数据,包括训练数据系数 a 和 b,还有得到的支持向量 SV。

Step2:根据

$$f(x)=\text{sgn}[(w^*)^T\phi(x)+b^*]$$

$$=\text{sgn}(\sum_{i=1}^{n}a_i^* y_i K(x_i,x)+b^*)$$

计算新输入测试邮件样本的相应决策输出值。

Step3:利用指示函数将 $f(x)$ 归为{−1,+1},作出分类决策:+1 为垃圾邮件,−1 为合法邮件。

邮件分类过滤面临一个关键的挑战,那就是对很多用户来说,宁愿接收垃圾邮件也不愿意收不到合法邮件。然而过滤技术可能阻止合法邮件的接收,也会漏过一些垃圾邮件。这说明对于邮件过滤技术来说,准确性是至关重要的。本文针对这一问题,对 SVM 邮件分类器进行了优化(如图 1)。

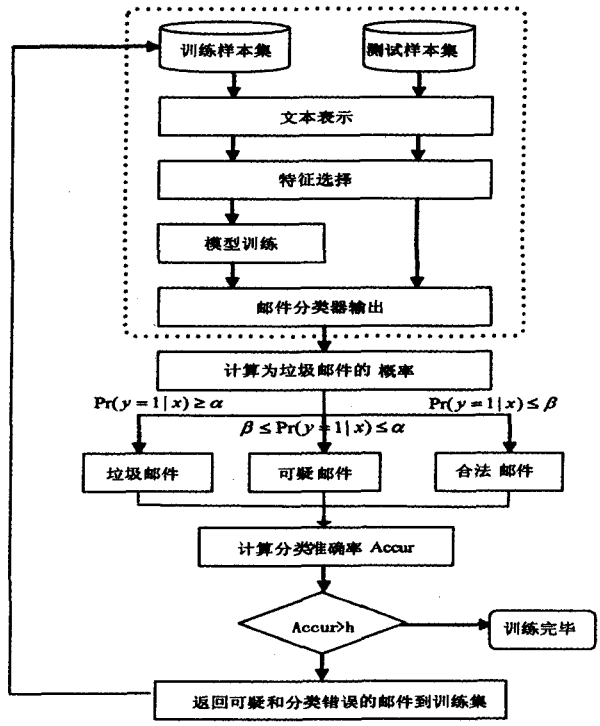

图 1 改进后 SVM 邮件分类器图

由于垃圾邮件本身的一些伪装方法比较好或者合法邮件的一些关键字和符合垃圾邮件的关键字匹配,每个邮件样本对划分的影响是不同的。一个邮件样本不能明确地归为某一类,而以一定概率或一定隶属度属于某一类,则会提高准确率,因此仅用邮件分类输出 $y\in\{-1,+1\}$ 表示类别信息并不恰当。为了使 SVM 更适合邮件过滤的分类,本文对其输出进行了如下处理[7]:

对一个输入邮件样本的分类输出 y,不直接根据+1 或−1 进行类别判断,而是输出属于某一类的概率。测试邮件样本 x 属于垃圾邮件的概率计算如下:

$$\Pr(y=1|x)\approx P_{A,B}(x)=\frac{1}{1+\exp(Af(x)+B)},y=f(x)$$

$\Pr(y=1|x)$ 为输入邮件样本 x 属于+1 类即为垃圾邮件的概率输出;A、B 两个参数的最佳取值为:对带标号的样本集,$\{(x_i,y_i)\}_{i=1}^{l}$,求解 $F(z)$,使之最小时 A、B 的计算值

$$\min_{z=(A,B)}F(z)=-\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i)),$$

$$\text{for } p_i=P_{A,B}(x_i),\text{and } t_i=\begin{cases}\dfrac{N_++1}{N_++2} & y_i=+1\\ \dfrac{1}{N_-+2} & y_i=-1\end{cases},i=1,\cdots,l$$

$F(z)$为 Platt 极小化负对数似然函数[7]，N_+为垃圾邮件样本数，N_-为合法邮件样本数。

由于输出概率 $\Pr(y=1|x)$越大，则为垃圾邮件的可能性越大；否则为合法邮件的可能性越大。所以，本文设置的两个参数α应该尽量大，β应尽量小；当输出概率大于α时，判定为垃圾邮件；小于β时，为合法邮件；界于β和α之间时，为可疑邮件。

根据上面的思想和后面图 2 参数调整效果的比较，两个参数取值范围在 $0.7<\alpha<1.0$ 和 $0<\beta<0.3$ 时的实验结果是比较合理的。分类输出判定完毕后，计算准确率指标 Accur。当准确率小于一定期望值 h 时，分类错误的邮件和可疑的邮件返回到训练邮件样本集，进行下一次训练，直到准确率满足该期望值时，则得到最优 SVM 邮件分类器。

4 实验结果分析

垃圾邮件过滤中的分类实际上属于文本分类，评价体系也借用文本分类中的一些指标（文[2]中有详细描述）。主要有查全率(Recall)、查准率(Precision)、准确率(Accuracy)和 F 值。下面只对准确率指标做描述（垃圾邮件为+1，合法邮件为−1）：

表 1 邮件分类评价指标表（单位为邮件数）

	专家判定为垃圾邮件	专家判定为合法邮件
分类器输出为垃圾邮件	A	B
分类器输出为合法邮件	C	D

$N=A+B+C+D$ 为邮件测试样本集总数，分类准确率为 $\text{Accur}=\frac{A+D}{N}$，这个指标反映了对所有垃圾邮件和合法邮件的判对率。

本文实验中主要使用了可以公开下载的 Spam Base 语料[8]。语料共有大约 4601 篇邮件，其中有 1813 篇垃圾邮件和 2788 篇合法邮件。实验中只对邮件标题和邮件正文进行训练和分类。在 Weka[9] 的基础上实现了优化的 SVM 分类器（Weka 中的 SVM 和 NaiveBayes 算法可直接进行文本挖掘）。Weka 是一个开源的数据挖掘算法集，支持很多流行的数据挖掘算法，如 SVM、NaiveBayes、神经网络、kNN 聚类等。针对本文提出的改进算法（SVM-PO），在实验过程中，根据 (α,β) 参数取值计算得到相应的 Accur 值。参数调整效果如图 2。

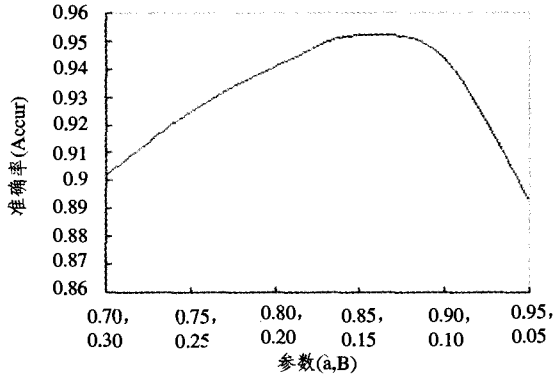

图 2 参数调整效果图

从图 2 可以看出，参数对取值选择在靠近(0.85,0.15)时，分类准确率最高；而当参数值取在(0.70,0.30)或(1.00,0.00)，即参数阈值边界周围时，准确率相对较低。在准确率期值最好 $h=0.952$，即(α,β)为(0.85,0.15)时，本文改进的 SVM-PO 邮件分类器和传统的 SVM 邮件分类器[10]、naive-Bayes 邮件分类器[11]的实验数据效果比较如表 2。

表 2 邮件分类实验结果比较

算法	Precision	Recall	Accuracy
NaiveBayes	0.898	0.832	0.895
SVM	0.927	0.867	0.934
SVM-PO	0.954	0.896	0.952

从表 2 中的数据可以看出，本文提出的基于 SVM 的改进邮件分类器 SVM-PO 相比传统的 SVM、NaiveBayes 邮件分类器，查全率、查准率和准确率都有一定提高，具有更好的分类效果。

结束语 SVM 算法在垃圾邮件过滤领域还未得到广泛应用，现在比较流行的垃圾邮件过滤技术主要是 NaiveBayes 算法。在实际的邮件过滤应用中，每个邮件样本对类别划分的影响不同，如果明确地归为某一类，则分类误差较大。以一定概率或一定隶属度进行分类，则会提高分类准确率。根据这种思想，本文在传统 SVM 邮件分类器的基础上，提出了改进的 SVM-PO 分类器，通过对分类输出进行概率计算，根据概率值所落在的参数阈值区间来判断邮件所属类别。相比传统的 SVM 邮件分类器和 Naive Bayes 邮件分类器，实验证明这种改进的邮件分类器是有效可行的，分类准确率有一定提高。下一步工作将会继续关注 SVM 邮件分类器的稳定性，同时解决一些该模型应用在实时邮件分类环境过程中所遇到的问题。

参 考 文 献

1 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995

2 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京：中国科学院计算技术研究所，2004

3 Kolcz A, Alspector J. SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs [A]. In: Proc. ICDM22001 Workshop on Text Mining, 2001

4 Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, 1998. 137～142

5 Yang Y, Pedersen J. A comparative study on feature selection in text categorization. In: International Conference on Machine Learning (ICML), 1997

6 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge U K: Cambridge University Press, 2000

7 Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola A, Bartlett P, Scholkopf B, et al, eds. Advances in Large Margin Classifiers. Cambridge, MA, 2000

8 Spambase. http://www.ics.uci.edu/～mlearn/MLRepository.html/, (2006-6-22)

9 Weka. http://www.cs.waikato.ac.nz/ml/weka/

10 Drucker H, Wu D, Vapnik V N. Support Vector Machines for Spam Categorization [J]. IEEE Transactions on Neural Networks, 1999, (20)5: 1048 ～ 1054

11 Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering. In: Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, 2000. 9～17