改进的一分类支持向量机的邮件过滤研究

秦 谊[1],裴 峥[1],杨霁琳[2]
QIN Yi[1],PEI Zheng[1],YANG Ji-lin[2]

1.西华大学 数学与计算机学院,成都 610039
2.西南交通大学 数学学院,成都 610031
1.School of Mathematics & Computer Science,Xihua University,Chengdu 610039,China
2.School of Mathematics,Southwest Jiaotong University,Chengdu 610031,China
E-mail:pqyz@263.net

QIN Yi,PEI Zheng,YANG Ji-lin.Filtering e-mail based on improved One-class Support Vector Machine.Computer Engineering and Applications,2009,45(20):151-153.

Abstract: Because there are many users in server,and users have different understand or admitting degrees for the content of e-mails,uncertain information processing is dealt with in filtering e-mails.From the content of e-mails point of view,filtering e-mails always deals with privacy,this is disadvantage for largely collecting e-mails and evaluating them.Filtering e-mail based on improved one-class SVM is proposed,the advantages of the method are (1)users only give membership degrees for uncertain e-mails which will be dealt with;(2)classing e-mails model is constructed by a kind of e-mail samples;(3)membership degrees are discussed in one-class SVM,and membership degrees are also used to decide punish factors.Simulation shows that the method is effective.

Key words: One-class Support Vector Machine(1-SVM);e-mail filtering;membership degree;uncertainty;Ordered Weighted Averaging(OWA) operator

摘 要:服务器端存在多个用户,且人们对邮件内容的理解和认可程度不同,因此邮件过滤中涉及到不确定信息的处理。就邮件内容来看,邮件过滤通常涉及到隐私,不利于大量收集样本并评价打分。因此提出了一种基于改进的一分类支持向量机的邮件过滤方法。该方法优点在于:(1)用户只需为不确定性很强的待区分邮件给出隶属度;(2)只需收集和训练一类邮件样本,便可以建立邮件分类模型;(3)把隶属度首次引入到 1-SVM 中,并且由隶属度的值的大小来确定惩罚因子的值。通过仿真实验验证了该方法的有效性。

关键词:一分类支持向量机;邮件过滤;隶属度;不确定性;有序加权平均算子
DOI:10.3778/j.issn.1002-8331.2009.20.045 **文章编号:**1002-8331(2009)20-0151-03 **文献标识码:**A **中图分类号:**TP18

1 引言

随着互联网的普及和发展,电子邮件以其方便、快捷、高效的优势成为最受欢迎的网络功能之一,同时也成为人们日常生活中信息交流的重要手段之一。但随之而来的垃圾邮件已成为互联网上迫在眉睫需要解决的重大问题。从 20 世纪 90 年代以来,许多专家和研究者都提出了各种垃圾邮件的过滤方法,其中也包括多种机器学习方法[1-4]。支持向量机是一种基于统计学习理论的较新的机器学习方法,已被成功运用在许多分类领域,包括垃圾邮件过滤问题[5-9]。

在实际生活中,一封邮件是垃圾邮件还是合法邮件,不同的用户往往有不同的认识,而且还有程度的问题,因此对邮件过滤的处理被视为不确定信息处理问题。有些合法邮件和垃圾邮件对广大用户来说是很明确的,其用户的意见应该是一致的,即它们的不确定性几乎为零;而另一些邮件,则会让用户产生较大的分歧意见,其不确定性很大。将邮件的不确定性形式化时,通常是给每封邮件样本赋予一个隶属度,而该隶属度是通过 OWA 算子聚合用户们对邮件样本的评价信息而得[8]。但在现实生活中,大量的合法邮件通常都涉及到用户的个人隐私,要将其公布并进行打分评价,无疑是一件比较困难的事情。当然想要大量地收集这一部分合法邮件也并不是件容易的事情。

针对以上的情况,提出了一种基于改进的一分类支持向量机的垃圾邮件过滤方法。首先,根据邮件样本的不确定性程度,把邮件分为:明确合法邮件、待区分邮件(不确定性较强的邮件)和明确垃圾邮件;接着把垃圾邮件作为邮件样本集上的模糊概念,利用 OWA 算子聚合出待区分邮件属于垃圾邮件的隶

基金项目:四川省重大科技专项项目(No.2008GZ0118);四川省杰出青年基金(No.06ZQ026-037)。
作者简介:秦谊(1958-),女,实验师,主研方向:智能信息处理,数据挖掘;裴峥(1968-),男,教授,主研方向:语言值处理,智能信息处理;杨霁琳(1981-),女,博士生,主研方向:智能信息处理,粗糙集理论与应用。
收稿日期:2009-03-24 **修回日期:**2009-05-07

属度,同时定义明确垃圾邮件的隶属度;然后为了能更好地提高合法邮件的检测率和降低垃圾邮件的错分率,通过邮件样本不同的隶属度的值来定义 SVM 中惩罚因子的值。最后以垃圾邮件作为训练样本,用改进的一分类支持向量机作为分类器去检测合法邮件。

2 一分类支持向量机

1-SVM(One-class Support Vector Machines)[10-11]是 SVM 的一种扩展,是 Bernhard Schlkopf 等人于 1999 年提出的,用于解决一类问题。其基本思想是,在选定了一个核函数以后,把空间中的坐标原点视为另一类中惟一的点,并且引入松弛变量,进而可以应用传统的二类支持向量机。文中将通过构造球体的方法来实现一分类支持向量机[11],其方法:给定一个正类样本点集$\{x_i,i=1,\cdots,l\}$,$x_i\in R^d$,设法找到一个以 a 为中心,以 R 为半径的能够包含所有样本点的最小球体。如果直接进行优化处理,所得到的优化区域就是一个超球体。为了使优化区域更紧致,这里采用支持向量机中的核映射思想。首先用一个非线性映射 ϕ 将样本点映射到高维特征空间,然后在特征空间中求包含所有样本点的最小超球体,这样能获得原空间中更紧致的优化区域。同时用核函数来代替高维空间中的内积运算,即找一个核函数 $K(x,y)$,使得 $K(x,y)=\langle\phi(x),\phi(y)\rangle$,从而简化了计算,也避免了明确知道映射 ϕ。于是优化问题为:

$$\min\ F(R,a,\xi_i)=R^2+C\sum_{i=1}^{l}\xi_i \tag{1}$$

约束为:

$$(\phi(x_i)-a)(\phi(x_i)-a)^{\mathrm{T}}\leqslant R^2+\xi_i$$
$$\xi_i\geqslant 0,i=1,\cdots,l \tag{2}$$

其中,参数 C 为大于零的常数。其优化问题的对偶形式为:

$$\max\sum_{i=1}^{l}\alpha_iK(x_i,x_i)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j) \tag{3}$$

约束为:

$$\sum_{i=1}^{l}\alpha_i=1\quad 0\leqslant\alpha\leqslant C,i=1,\cdots,l \tag{4}$$

求解该二次优化问题可以得到 α 的值,通常大部分 α_i 将为零,不为零的 α_i 所对应的样本同样被称为支持向量(SV)。按照 Kuhn-Tucker 定理,并采用核函数代替高维空间中的内积运算,则有

$$\alpha_i\left(R^2+\xi_i-K(x_i,x_i)+2\sum_{j=1}^{l}\alpha_jK(x_j,x_i)-a^2\right)=0\quad i=1,\cdots,l$$
$$\beta_i\xi_i=0,i=1,\cdots,l \tag{5}$$

并可得:$(C-\alpha_i)\xi_i=0,i=1,\cdots,l$。对应于 $0<\alpha_i<C$ 的样本满足

$$R^2-\left[K(x_i,x_i)-2\sum\alpha_jK(x_j,x_i)+a^2\right]=0 \tag{6}$$

因此,用任意一个支持向量可以由公式(7)求出 R 的值。对于测试样本 z,设

$$f(z)=(\phi(z)-a)(\phi(z)-a)^{\mathrm{T}}=$$
$$K(z,z)-2\sum_{i=1}^{l}\alpha_iK(z,x_i)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j) \tag{7}$$

若 $f(z)\leqslant R^2$,则 z 为正常点,否则 z 为异常点。

3 邮件过滤的不确定性模型建立

3.1 邮件过滤的不确定性

邮件过滤中不确定性的形式化,通常是给每个邮件样本赋予一个隶属度 $s_i(0<s_i\leqslant 1)$,作为邮件样本的一个特性,即这封邮件属于合法邮件或是垃圾邮件的程度。但某部分邮件的不确定性很低,即对于用户来说,它们是百分之百的合法邮件或垃圾邮件;而另一部分邮件的不确定性却较高,它根据用户的不同理解,划分成合法邮件或是垃圾邮件的可能性都比较大。这部分邮件称为待区分邮件。因此,根据邮件的不确定性程度把所有邮件概括为三类:明确合法邮件、待区分邮件、明确垃圾邮件。并假设这三部分邮件覆盖了所有的邮件且相互独立。

(1)明确合法邮件:①私人信件;②商务信件;③订制信件(用户在网上自行订制的一些邮件);④信息反馈信件(用户在网上注册、登录等一些系统回复信息)。这一部分邮件,对于用户来说,是非常重要的合法邮件,不允许被错分和丢失的。同时这部分邮件隐私度和保密性也都较高。

(2)明确垃圾邮件:①病毒性信件;②涉及黄色、赌博、反动思想等不健康信件。对于这部分邮件,用户们是绝对不愿意在自己邮箱里见到并花时间处理的。

(3)待区分邮件:①某些网站的强制信息;②网络服务和网络功能的推销;③商品广告;④业务推销等。这部分邮件具有很强的多样性和不确定性,每个用户根据自己的理解给出的判断都可能不同。

从上面的分类可以看出,对于客户服务器端的多个用户,邮件过滤的不确定性主要针对待区分邮件。

3.2 邮件样本隶属度的计算

用隶属度表示邮件过滤中存在的不确定性。邮件根据其不确定性程度被分为三类,明确合法邮件由于涉及到许多隐私,不但收集起来较难,把它们全部公开让所有用户评价就更难。而邮件的不确定性主要针对的是待区分邮件,因此只收集明确的垃圾邮件和待区分邮件作为训练样本集,且把垃圾邮件看成是训练样本集上的一个模糊概念,定义待区分邮件的隶属度为属于垃圾邮件的程度。

直观地,在邮件服务器端有 m 个用户,即:$U=\{u_1,u_2,\cdots,u_n\}$,$(m\geqslant 2)$,以及 n 个邮件样本$\{(x_1,y_1),\cdots,(x_n,y_n)\}$,其中 u_k 表示第 k 个用户。用户 u_k 针对邮件训练样本集中每封邮件给出的评价向量为 $\boldsymbol{s}_k=(s_1^k,s_2^k,\cdots,s_n^k)^{\mathrm{T}}$,其中 s_j^k 表示用户 u_k 对邮件 x_j 属于垃圾邮件程度的评价信息,$s_j^k\in(0,1]$。当 s_j^k 趋于 0 时,表示用户 u_k 认为样本点 x_j 是垃圾邮件的可能性很小;相反,当 s_j^k 趋于 1 时,表示用户 u_k 认为样本点 x_j 是垃圾邮件的可能性很大。于是,n 个用户对第 j 个邮件 x_j 的评价向量为 $\boldsymbol{B}^j=(s_j^1,s_j^2,\cdots,s_j^m)^{\mathrm{T}}$。用 OWA 算子[12]对每封邮件 x_j 的评价信息进行群集结,即得到邮件 x_j 的群体评价值:

$$F_j=F(s_j^1,s_j^2,\cdots,s_j^m)=H^j(B^j)^{\mathrm{T}}=\sum_{k=1}^{m}w_j^k s_j^{\sigma(k)},j=1,2,\cdots,n \tag{8}$$

其中 $\boldsymbol{B}^j=F(s_j^{\sigma(1)},s_j^{\sigma(2)},\cdots,s_j^{\sigma(m)})^{\mathrm{T}}$ 中的元素是有序的,对任意 $l\geqslant k$,有 $s_j^{\sigma(l)}\leqslant s_j^{\sigma(k)}$,且 $s_j^{\sigma(k)}\in(0,1]$。$s_j^{\sigma(k)}$ 是$(s_j^{\sigma(1)},s_j^{\sigma(2)},\cdots,s_j^{\sigma(m)})$中按从大到小顺序排在第 k 位的元素。

$$w^i=Q\left[\frac{i}{m}\right]-Q\left[\frac{i-1}{m}\right] \tag{9}$$

OWA 算子中,模糊量词 Q 表示为:

$$Q(r)=\begin{cases}0 & 0\leqslant r<\alpha\\ \dfrac{r-\alpha}{\beta-\alpha} & \alpha\leqslant r\leqslant\beta\\ 1 & \beta<r\leqslant 1\end{cases} \tag{10}$$

确定模糊量词 Q 中参数(α,β)的值就能得到权重向量 $\boldsymbol{w}^i$,从而得到邮件样本的最终评价值。选取“大多数”的群集结原则,即参数$(\alpha,\beta)=(0.3,0.8)$。而邮件样本中的明确垃圾邮件的综合评价值即隶属度,统一为 $s_j=1$。

4　基于一分类支持向量机的邮件过滤

本文邮件过滤被当作是一个一分类问题。所以采用一分类支持向量机作为分类器,只需训练垃圾邮件样本,合法邮件作为检测。

4.1　改进的一分类支持向量机

在本文邮件过滤方法中,每封邮件样本 x_i 多了一个属性,即垃圾邮件的隶属度 s_i。邮件样本集被表示为:$D=\{(x_1,y_1,s_1),\cdots,(x_n,y_n,s_n)\}$,但一分类支持向量机的训练样本是:$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$。当隶属度同 1-SVM 相结合后,其一分类支持向量机的优化问题就被改写成为:

$$\min F(R,a,s_i,\xi_i)=R^2+C\sum_{i=1}^{l}s_i\xi_i \tag{11}$$

约束为:

$$(\phi(x_i)-a)(\phi(x_i)-a)^{\mathrm{T}}\leqslant R^2+\xi_i$$
$$\xi_i\geqslant 0,i=1,\cdots,l \tag{12}$$

其中,$s_i\in[\sigma,1]$表示第 i 个邮件样本属于垃圾邮件类的隶属度,$\sigma>0$ 为足够小的数。参数 C 为大于零的常数,通常被称为惩罚因子。在 1-SVM 中,惩罚因子 C 决定了在超球面最大半径和最小训练错误之间的折衷度。引入拉格朗日函数

$$L(R,a,\alpha_i,\beta_i,\xi_i)=R^2+C\sum_{i=1}^{l}s_i\xi_i-$$
$$\sum_{i=1}^{l}\alpha_i\{R^2+\xi_i[\phi(x_i)^2-2a\phi(x_i)+a^2]\}-\sum_{i=1}^{l}\beta_i\xi_i \tag{13}$$

其中,$\alpha_i\geqslant 0,\beta_i\geqslant 0,i=1,\cdots,l$。

函数 L 的极值应满足条件

$$\frac{\partial L}{\partial R}=0,\frac{\partial L}{\partial a}=0,\frac{\partial L}{\partial\xi_i}=0 \tag{14}$$

从而得

$$\sum_{i=1}^{l}\alpha_i=1\quad a=\sum_{i=1}^{l}\alpha_i\phi(x_i)\quad s_iC-\alpha_i-\beta_i=0\quad i=1,\cdots,l \tag{15}$$

并将此优化问题转换成其对偶形式为:

$$\max\sum_{i=1}^{l}\alpha_iK(x_i,x_i)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j) \tag{16}$$

约束为:

$$\sum_{i=1}^{l}\alpha_i=1\quad 0\leqslant\alpha\leqslant s_iC\quad i=1,\cdots,l$$

按照 Kuhn-Tucker 定理,并采用核函数代替高维空间中的内积运算,则有

$$\alpha_i\left(R^2+\zeta_i-K(x_i,x_i)+2\sum_{j=1}^{l}\alpha_jK(x_j,x_i)-a^2\right)=0\quad i=1,\cdots,l$$
$$\beta_i\xi_i=0,i=1,\cdots,l \tag{17}$$

根据式(15)可得:$(s_iC-\alpha_i)\xi_i=0,i=1,\cdots,l$。$K(x,y)$为核函数,使得 $K(x,y)=\langle\phi(x),\phi(y)\rangle$;对应于 $0<\alpha_i<s_iC$ 的样本 x_i 为支持向量样本;$\alpha_i=0$ 的样本 x_i 为能够被正确分类的样本;而对应于 $\alpha_i=s_iC$ 的样本 x_i 为不能够被正确分类的样本。对应于 $0<\alpha_i<s_iC$ 的样本满足

$$R^2-\left[K(x_i,x_i)-2\sum\alpha_jK(x_j,x_i)+a^2\right]=0 \tag{18}$$

因此,用任意一个支持向量可以由公式(14)求出 R 的值。对于测试样本 z,设

$$f(z)=(\phi(z)-a)(\phi(z)-a)^{\mathrm{T}}=$$
$$K(z,z)-2\sum_{i=1}^{l}\alpha_iK(z,x_i)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j) \tag{19}$$

若 $f(z)\leqslant R^2$,则 z 为正常点,否则 z 为异常点。

由上面的公式看出:因为惩罚因子 C 为常数,当 s_i 越大,样本 x_i 被错分的可能性越小,其分类超球面 R 的值越大;当 s_i 越小,样本 x_i 被错分的可能性越大,其 R 的值越小。

根据邮件样本的隶属度 s_i 的定义,用 1-SVM 作为分类器能保证合法邮件具有较高检测率,同时减少垃圾邮件的错分率,从而整体提高分类模型的有效性。

4.2　仿真实验

从实际生活和语料集中一共收集邮件 496 封,其中明确合法邮件 97 封、待区分邮件 186 封、明确垃圾邮件 213 封。待区分邮件和明确垃圾邮件作为训练样本,合法邮件样本用来检测。首先对邮件样本进行预处理,特征选择采用信息增益(IG)法:将训练集中的所有词按照信息增益计算值的大小排序,选取排在前面约 80%的词作为特征集。同时预处理还包括了去停用词和词汇还原。最终邮件样本被表示成二维向量。

在实验中,假设 10 个用户对每个邮件样本做了评价,即,给出训练样本集中每封邮件的隶属度 s_i。采用合法邮件的准确率(LP)和查全率(LR)以及垃圾邮件的准确率(SP)和查全率(SR)对所述方法进行评价:

$$\text{准确率}(precision)=\frac{\text{分类正确的邮件数}}{\text{该类实际所得邮件数}}$$

$$\text{查全率}(recall)=\frac{\text{分类正确的邮件数}}{\text{该类应有邮件数}}$$

实验时,把整个邮件样本分成 4 份进行交叉验证,取 4 次的平均值,以得到更为可靠的准确率和查全率。其结果如表 1:

表 1　实验结果

序	LP/(%)	LR/(%)	SP/(%)	SR/(%)
1	90.21	93.52	90.47	87.55
2	87.53	91.50	89.36	92.85
3	90.10	92.37	90.54	91.89
4	88.52	90.61	91.02	92.13
Avg	89.09	92.00	90.35	90.86

然后运用相同的样本集,让 1-SVM同 SVM 和 FSVM 进行比较实验,以验证该方法的有效性。

(下转 168 页)