第33卷 第20期 计 算 机 工 程 2007年10月
Vol.33 No.20 Computer Engineering October 2007
·软件技术与数据库· 文章编号：1000—3428(2007)20—0061—03 文献标识码：A 中图分类号：TP311

基于OWA算子与FSVM的邮件过滤

杨霁琳，彭 宏，邓 爽，赵毓高
（西华大学数学与计算机学院，成都 610039）

摘 要：不同用户对邮件的合法性有着不同的认识，因此对邮件过滤的研究应视为不确定信息处理问题，该文提出了一种融合有序加权平均(OWA)算子与模糊支持向量机(FSVM)的邮件过滤方法。其主要思想是利用OWA算子来产生每封邮件的一个合理的综合评价值作为其隶属度，采用FSVM对邮件进行分类。仿真实验结果验证了该方法的有效性。

关键词：OWA算子；模糊支持向量机；邮件过滤；分类器

E-mail Filtering Based on OWA Operator and FSVM

YANG Ji-lin, PENG Hong, DENG Shuang, ZHAO Yu-gao
(School of Mathematics & Computer Science, Xihua University, Chengdu 610039)

【Abstract】 Whether an e-mail is a legal one or not, different users hold different opinions. As a result, research of e-mail filtering should be considered as dealing with the uncertainties. This paper proposes an integrated method of ordered weighted averaging(OWA) operator and fuzzy support vector machines(FSVM). Its main idea is to obtain a reasonable value of comprehensive evaluation as each mail's membership degree by using the OWA operator method, then classify each mail though FSVM. Simulative experiments are conducted to verify the effectiveness of the method.

【Key words】 OWA operator; fuzzy support vector machines; e-mail filtering; classifier

电子邮件逐渐成为人们日常生活中信息交流的重要手段之一。但近些年来，垃圾邮件泛滥，不仅耗费网络资源，而且对企业正常运作和用户的正常工作造成严重干扰。为了防范垃圾邮件，人们提出了各种垃圾邮件过滤方法[1]和多种机器学习方法[2~5]。

支持向量机（SVM）是由Vapnik等人提出的一种基于结构风险最小化原理的新颖机器学习方法[6]，它具有小样本、良好的推广性能、全局最优等特点，已被成功地运用于许多分类问题的研究。在垃圾邮件过滤中运用SVM方法的研究已引起了一些研究者的兴趣[3~5]。已有机器学习方法在进行垃圾邮件过滤时，一封邮件要么被确定地分类为垃圾邮件，要么被确定地分类为合法邮件。但实际上，一封邮件是垃圾邮件还是合法邮件，不同的用户有不同的认识，而且还有程度的问题，因此，对邮件过滤的处理应被视为不确定信息处理问题。由于邮件服务器端存在众多的邮件用户，并且不同的用户视每封邮件为垃圾邮件或合法邮件的程度可能是不同的，因此使用机器学习方法对邮件进行分类时，需要一种方法来产生每封邮件属于垃圾邮件或合法邮件程度的综合评价值（即模糊隶属度）。因此，本文引入有序加权平均(OWA)[7]算子来得到模糊隶属度。

针对以上问题，本文提出了融合OWA算子与SVM的邮件过滤方法。首先采用OWA算子来产生每封邮件的一个合理的综合评价值作为其隶属度；然后采用SVM方法作为分类器对邮件进行分类。由于邮件的处理被视为模糊信息的处理，本文采用文献[8]提出的模糊支持向量机(FSVM)作为邮件分类器。

1 模糊支持向量机

在邮件分类使用支持向量方法的已有研究中[3~5]，均采用标准的支持向量机。在这些研究中，邮件样本被明确地标记为一类（合法邮件或垃圾邮件）。本文的研究中，每个邮件训练样本被赋予一个模糊隶属度，为此采用FSVM作为邮件分类器。

假设有如下的邮件训练样本集：

$$D=\{(x_1,y_1,s_1),\cdots,(x_n,y_n,s_n)\} \tag{1}$$

其中，$x_i\in R^N$表示第i个邮件样本；$y_i\in\{-1,1\}$代表邮件类（合法邮件或垃圾邮件）；$s_i\in[\sigma,1]$表示第i个邮件样本属于一类的隶属度，$\sigma>0$ 为足够小的数。支持向量机方法寻找最优超平面即为求解如下的二次规划问题：

$$\begin{aligned}&\text{Min}\quad \frac{1}{2}w\cdot w+C\sum_{i=1}^{n}s_i\xi_i\\&\text{s.t.}\quad y_i(w\cdot z_i+b)\geq 1-\xi_i,\quad i=1,\cdots,n\\&\qquad \xi_i\geq 0,\quad i=1,\cdots,n\end{aligned} \tag{2}$$

其中，C 为正则化参数。构造 Lagrangian 函数，并通过求Lagrangian函数鞍点的方法，可将式(2)转化为其对偶问题：

$$\begin{aligned}&\text{Min}\quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{n}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{n}y_i\alpha_i=0\\&\qquad 0\leq\alpha_i\leq s_iC,\quad i=1,\cdots,n\end{aligned} \tag{3}$$

于是，分类器的决策函数为

$$f(x)=sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b\right) \tag{4}$$

称对应于$\alpha_i>0$ 的样本x_i为支持向量。在式(3)与式(4)中，

基金项目：四川省教育厅基金资助重点项目(2005A117)
作者简介：杨霁琳(1981－)，女，硕士研究生，主研方向：数据挖掘，支持向量机；彭 宏，教授；邓 爽、赵毓高，硕士研究生
收稿日期：2006-10-30 **E-mail：**yjl524@163.com

$K(x_i,x_j)$为支持向量核，表示某特征空间Z的内积，即：$z_i \cdot z_j=\phi(x_i)\cdot\phi(x_j)=K(x_i,x_j)$。

2 OWA 算子

有序加权平均(OWA)算子由Yager于 1988 年提出，用于有效地融合多组模糊的和不确定的信息[7]。下面给出OWA算子的描述：

假设F: $R^m \to R$，有一与F相关联的m维加权向量，$\boldsymbol{w}=(w^1,w^2,\cdots,w^m)$，$w^i\in[0,1]$，$1\le i\le m$，且$\sum w^i= w^1+ w^2+\cdots+ w^m=1$ 使得

$$F(a^1,a^2,\cdots,a^m)=\sum_{i=1}^{m} w^i b^i \tag{5}$$

其中，b^i是$(a^1,a^2,\cdots,a^m)$中第i个最大元素，则称F为m维OWA算子（有序加权平均算子）。

有序加权向量$\boldsymbol{w}=(w^1,w^2,\cdots,w^m)$可由下列公式确定：

$$w^i=Q\left[\frac{i}{m}\right]-Q\left[\frac{i-1}{m}\right],\quad i=1,2,\cdots,m \tag{6}$$

其中，Q 为模糊量词。Q 由下式给出：

$$Q(r)=\begin{cases}0 & r<\alpha\\ \dfrac{r-\alpha}{\beta-\alpha} & \alpha\le r\le\beta\\ 1 & r>\beta\end{cases} \tag{7}$$

其中，$\alpha,\beta,r\in[0,1]$。模糊量词$Q(r)$对应的参数(α,β)可有多种取值方式，习惯上在“大多数”,“至少一半”和“尽可能多”这 3 种基本原则下，参数(α,β)分别为(0.3, 0.8)，(0, 0.5)，(0.5, 1)[7~9]。

3 融合 OWA 算子与 FSVM 的邮件过滤方法

同一封邮件，有的用户可能将其视为垃圾邮件，其他的用户可能因为其中包含自己感兴趣的内容而将其视为合法邮件，因此，一封邮件是垃圾邮件还是合法邮件，不同的用户可能有不同的认识。进一步地认为这还存在一个程度问题，即一封邮件被视为垃圾邮件或合法邮件，不同的用户可能有不同的程度值。因此，为每个邮件样本赋予一个模糊隶属度s_i $(0<s_i\le1)$，把邮件过滤问题当作不确定信息处理问题。模糊隶属度s_i 被当作邮件样本的一个特性。在为有多个邮件用户的场合设计过滤器时，如邮件服务器端、分类器应具有不确定信息处理能力。同时，由于邮件服务器端存在众多邮件用户，不同的用户视每封邮件是垃圾邮件或合法邮件的程度可能是不同的，因此在对邮件进行分类时，需要一种方法来产生每封邮件属于垃圾邮件或合法邮件程度的综合评价值（即模糊隶属度）。在本文的邮件过滤方法中采用融合OWA算子与FSVM的技术。其主要思想包括如下两个方面：

(1)OWA 算子聚合邮件样本的模糊隶属度

假定在邮件服务器端有m个用户，以及n个邮件样本$\{(x_1,y_1),\cdots,(x_n,y_n)\}$。对待同一邮件，不同的用户认为它是垃圾邮件的程度可能不一样，或者说对每个样本的评价是不一样的。设用户集为$U=\{u_1,u_2,\cdots,u_m\}$ $(m\ge2)$，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$x_j$所给出其属于垃圾邮件程度的评价信息，$s_j^k\in(0,1]$。当$s_j^k$趋向于 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大；相反，当$s_j^k$趋向于 1 时，表示用户$u_k$认为样本点$x_j$是合法邮件的可能性很大。于是，这$m$个用户对第$j$个邮件$x_j$的评价向量为$\boldsymbol{B^j}=(s_j^1,s_j^2,\cdots,s_j^m)^{\mathrm{T}}$。接下来，利用OWA算子对于每封邮件$x_j$的评价信息进行群集结，从而给出关于邮件$x_j$的群体评价值：

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

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$l\ge k$，有$s_j^{\sigma(l)}\le s_j^{\sigma(k)}$，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。

在垃圾邮件过滤中，由于错分一封合法邮件要比错分一封垃圾邮件的后果要严重得多，因此要保证较多的合法邮件不被错分成垃圾邮件。为此在用OWA算子计算每封邮件的综合评价值时，对于权重向量$\boldsymbol{H^j}$的确定，本文采用“至少一半”的群集结原则，即参数$(\alpha,\beta)=(0,0.5)$。

根据式(6)和式(7)可得到需集结的样本点的权重向量$\boldsymbol{H^j}=\{w_1^j,w_2^j,\cdots,w_m^j\}$，进而由式(8)计算出邮件样本$x_j$的综合评价值$s_j$。注意到，由于在$\boldsymbol{B^j}$中是按各元素的值从大到小排列的，因此采用“至少一半”的群集结原则，即参数$(\alpha,\beta)=(0,0.5)$，这个群集结原则能聚合“至少一半”以上的用户评价信息，关键是它使得权重向量$\boldsymbol{H^j}=\{w_1^j,w_2^j,\cdots,w_m^j\}$里的值，全部集中在该向量前面，即对于一封邮件样本，它在评价信息里的从大到小的排序中越是靠前，它的权重就越比后面的大。这就保证了当得到$F^j=\boldsymbol{H^j}(\boldsymbol{B^j})^{\mathrm{T}}$值时，使尽可能多的合法邮件获得相对比较高的评价值，以保证尽可能多的合法邮件不被错分。同时从前面的计算可以知道：在训练样本点中，邮件属于合法邮件的可能性越大，它的F^j值就越接近 1，甚至等于 1；属于垃圾邮件的可能性越大，则它的F^j值就越靠近 0。

(2)FSVM 邮件分类器

通过前面的计算得到了邮件样本的综合评价s_j，该评价值指示了邮件样本属于合法邮件或者是垃圾邮件的程度。这里将s_j作为第j个样本点x_j的隶属度，并视为邮件样本的一个特性。这样就构成了形如式(1)训练样本集D。因此，邮件过滤问题就变为不确定信息处理问题，采用FSVM作为邮件分类器。在实验中，选用线性核作为FSVM的核函数。

4 实验与结果

4.1 实验方法

目前，Internet 邮件一般都采用 MIME 编码格式。邮件的主题和邮件内容通常采用 Base64 和 Quoted Printable 编码，识别前需要根据邮件主题和邮件内容的编码方式对其进行解码。对邮件去掉了 HTML 格式的 tag、附件等，保留了邮件正文的纯文本内容。在对邮件样本进行预处理时，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。

经上述预处理步骤后，在使用如前所述的FSVM处理邮件过滤的方法来训练分类器时，考虑m个用户分别对每个邮件样本进行评价，得到每个邮件样本的评价信息s_j^k，于是这m个用户对第j个邮件x_j的评价向量为$\boldsymbol{B^j}$，接下来就利用OWA算子的“至少一半”的群集结原则，得到最终邮件x_j群体评价值F_j，这里仍然约定$s_j^k\in(0,1]$。当s_j^k趋向于 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大；相反，当s_j^k趋向于 1 时，表示用户u_k认为样本点x_j是合法邮件的可能性很大。F_j也就是FSVM中的模糊因子s_j，即每个邮件样本的隶属度。

4.2 实验结果

因为在实际生活中收集垃圾邮件远比收集合法邮件要容易得多，为此笔者收集到 1 148 封垃圾邮件和 573 封合法邮件，并假设有 10 个用户对每个邮件样本做了评价。

本文采用合法邮件的准确率和查全率对本文所述方法进行评价。准确率为某类邮件中判断正确的邮件数与该类实验所得邮件数的比率；查全率为某类邮件中判断正确的邮件数与人工分类时该类应有的邮件数的比率。准确率和查全率反映了分类质量的两个不同方面，二者必须综合考虑，不可偏废。因此，本文还采取了一种新的评估指标 $F1$ 测试值来综合评估，其数学公式如下：

$$F1\text{测试值}=\frac{\text{准确率}\times\text{查全率}\times 2}{\text{准确率}+\text{查全率}}$$

实验时将总共 1 721 封邮件分为 6 份，每份大约 286 封，每次取其中的 5 份作为训练集，另一份为测试集，如此循环进行 6 次交叉验证。同时运用相同的样本集，让 FSVM 同标准的 SVM 和贝叶斯进行比较实验，以验证该方法的有效性。

在实验中，虽然该方法在训练分类器时，用到 OWA 算子稍微增加了训练时间，但最后对比实验结果表明，虽然本方法的合法邮件准确率略高于标准 SVM，但在合法邮件的查全率上则明显高于其他两种方法，以及相对较高的 $F1$ 测试值。这说明本文方法的性能要优于标准的 SVM 和贝叶斯。

表 1　实验结果

序号	合法邮件正确率/(%)	合法邮件查全率/(%)
1	92.59	97.77
2	91.07	96.23
3	93.45	94.58
4	90.96	95.02
5	90.14	96.89
6	92.77	96.47
Avg	91.83	96.16

表 2　几种方法的对比结果

方法	合法邮件		垃圾邮件		合法邮件标准率/(%)	合法邮件查全率/(%)	F1 测试值/(%)
	式别数	误识数	式别数	误识数			
贝页斯	527	46	1 036	112	82.47	91.97	86.96
标准 SVM	536	37	1 089	59	90.08	93.54	91.78
FSVM	551	22	1 099	49	91.83	96.16	93.95

5　结论

本文所提出的融合 OWA 算子与模糊支持向量机的邮件过滤方法，把邮件的过滤视为不确定信息处理问题。该方法利用 OWA 算子来聚合每个用户的评价信息来产生每封邮件的隶属度，并作为 FSVM 的模糊因子。仿真试验表明，这种融合 OWA 算子与 FSVM 的方法其性能优于标准的 SVM 和贝叶斯。

参考文献

1 Cohen W. Learning Rules That Classify E-mail[C]//Proc. of AAAI Spring Symposium on Machine Learning in Information Access. 1996.

2 Saham M, Dumais S, Heckerman D, et al. A Bayesian Approach to Filtering Junk E-mail[C]//Proceedings of AAAI Workshop on Learning for Text Categorization. 1998: 55-62.

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

4 Kolcz A, Alspector J. SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs[C]//Proceedings of the TextDM'01 Workshop on Text Mining-held. 2001: 309-347.

5 Rios G, Zha H. Exploring Support Vector Machines and Random Forests for Spam Detection[C]//Proceedings of the 1st Conference on Email and Anti-Spam. 2004: 284-292.

6 Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

7 Yager R. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1988, 18(1): 183-190.

8 Lin C, Wang S. Fuzzy Support Vector Machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

9 Ronald R. Families of OWA Operators[J]. Fuzzy Sets and Systems, 1993, 59(1): 125-148.

(上接第 60 页)

复杂度更高。过程仿真和过程评估是在相对封闭的边界条件下对资源进行纯数学化的建模，这方面的研究还有待进一步的探讨。

参考文献

1 Rosemann M, Muehlen Z M. Evaluation of Workflow Management Systems — — A Meta Model Approach[J]. Australian Journal of Information Systems, 1998, 6(1): 103-116.

2 Muehlen M. Organizational Management in Workflow Applications: Issues and Perspectives[J]. Information Technology and Management Journal, 2004, 5(3): 271-291.

3 Kumar A, Verbeek H M W. Organizational Modeling in UML and XML in the Context of Workflow Systems[C]//Proceedings of the 18th Annual ACM Symposium on Applied Computing. 2003: 603-608.

4 Muehlen Z M. Resource Modeling in Workflow Applications. Workflow Management Conference[D]. Jesuiten-kolleg: University of Münster, 137-153.

5 Russell N, Edmond D. Workflow Resource Patterns[D]. Eindhoven: Eindhoven University of Technology, 2004.

6 Du W, Davis J, Huang Y N, et al. Enterprise Workflow Resource Management[Z]. Palo Alto, CA, USA: Hewlett Packard Laboratories, 1999.

7 Bussler C. Policy Resolution in Workflow Management Systems[J]. Digital Technical Journal, 1995, 6(4): 26-49.

8 Momotko M, Subieta K. Dynamic Changes in Workflow Participant Assignment[C]//Proceedings of the 6th East-european Conference on Advances in Databases and Information Systems. 2002.

9 Netjes M, Reijers H A. Analysis of Resource-constrained Processes with Colored Petri Nets[C]//Proceedings of the 6th Workshop on the Practical Use of Coloured Petri Nets and CPN Tools. 2005.