基于用户兴趣集的在线垃圾邮件快速识别新方法

王友卫,刘元宁,凤丽洲,朱晓冬
(吉林大学计算机科学与技术系,吉林长春 130012)

摘 要: 为在不显著降低垃圾邮件识别精度的同时有效提高邮件识别速度,提出了一种在线垃圾邮件快速识别新方法.首先引入用户正、负兴趣集的概念,结合用户兴趣集及支持向量机对邮件进行分类;然后根据主动学习理论,结合训练集样本密度及改进角度差异方法寻找分类最不确定的样本并推荐给用户进行类别标注;最后将标注后样本及分类最确定性样本加入训练集,并使用样本价值评价新函数淘汰冗余样本以生成新的训练集.实验表明,本文方法的用户标注负担小,垃圾邮件识别精度高、速度快,具有较高的在线应用价值.

关键词: 垃圾邮件; 用户兴趣集; 支持向量机; 主动学习; 在线应用
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)10-1963-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.10.013

A Novel Quick Online Spam Identification Method Based on User Interest Set

WANG You-wei, LIU Yuan-ning, FENG Li-zhou, ZHU Xiao-dong
(*College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*)

Abstract: In order to improve the spam identification speed without sacrificing the accuracy seriously, a novel quick online spam identication method is proposed. Firstly, the conceptions of user positive interest set and user negative interest set are introduced, and emails are classified by combining user interest sets and support vector machine. Secondly, based on the active learning theory, the sample densities of different categories and the improved angle diversity method are used to select the most uncertainly classified samples, and the selected samples are recommended to users for labeling. Finally, the labeled and the classified samples with greatest possiblities are put into the training set, and a novel sample value evaluating function is proposed to filter the redundant samples for generating a new training set. Experimental results show that, the sample labeling burden of the proposed method is small, the spam identification accuracy is high, and the spam identification speed is fast, the high value of the proposed method on online application is proved.

Key words: spam; user interest set; support vector machine; active learning; online application

1 引言

在线垃圾邮件识别可以被看作是一种典型的文本分类问题[1].不同于传统文本分类,在线垃圾邮件识别面临以下几个重要问题:①难以在短时间内针对某一用户获得大量的标注样本;②算法针对分类响应实时性要求比较高;③用户个人喜好对邮件识别效果影响较大.因此,有效解决以上问题成为垃圾邮件在线识别的首要任务.

近年来,增量学习与主动学习理论已被广泛应用于文本分类[2,3].Syed 等提出了一种结合支持向量机(Support Vector Machine,SVM)的增量学习方法[4].该方法将规模较大的训练集等分成若干子集,通过将每个子集加入到先前子集对应的支持向量集中实现 SVM 增量学习.该方法验证了 SVM 支持向量在训练样本集表示方面的有效性,为后续增量学习方法奠定了基础,但其针对训练集中冗余样本处理过于简单,导致识别精度不高.为此,Wu 等人在选择待加入训练集和移除冗余样本过程中使用 KKT (Karush-Kuhn-Tucker)条件[5].算法样本训练及分类速度较快,但用户并未参与样本分类过程,因此所得结果往往与用户实际判断不符,导致在线分类精度不高.Amayri 等人将传统结合 SVM 邮件识别

收稿日期:2014-03-19;修回日期:2014-08-15;责任编辑:李勇锋
基金项目:国家科技成果转化项目(财建[2011]329,财建[2012]258)

方法与结合主动学习的 SVM 的邮件学习方法进行比较,指出主动学习能明显改善垃圾邮件的识别效果[6].Tong 等人结合 SVM 提出了边界采样(Margin Sampling,MS)主动学习方法.该方法选择距离分类边界最近的样本作为最不确定性样本推荐给用户[7].Hu 等人结合 MS 和样本间角度差异(Angle Diversity,AD)方法提出了一种新颖的不确定样本标注方法[8].该方法考虑了样本预测类别的不确定性对于样本推荐结果的影响,不足之处在于每次待标注样本选择过程都需实施多次 SVM 训练.Leng 等结合 MS 选择分类边界中分类最不确定样本,通过统计待标注样本标签改变率获得不同类别中心样本并将其加入训练集中[9].该方法在保证样本标注准确率的同时有效降低了人工标注负担;Chen 等人通过基于最优标号和次优标号主动学习方法去挖掘那些对当前分类器模型最有价值的样本进行人工标注,并借助带约束条件的自学习进一步利用样本集中大量的未标注样本,使得在花费较小标注代价情况下,能够获得良好的分类性能[10];Liu 等人将每封邮件分为多个域,通过结合主动学习并集成每个域对应的分类器所得结果来预测邮件最终类别[1].算法精度较高,但是不同域分类器间权重难于确定,且邮件不同域训练、分类过程独立进行,导致算法耗时较大.

上述垃圾邮件识别方法普遍面临下面问题:①样本识别过程直接使用 SVM 等分类器[4,6,8~10],识别响应速度受训练样本集规模大小、特征向量空间维数等因素影响较大;②MS 主动学习方法通过获得待标注样本与 SVM 分类边界的位置关系来决定样本的分类确定性程度,忽略了不同样本集样本分布特性的影响;③缺乏有效的冗余样本过滤机制,导致训练集规模无限制增长,增加了邮件在线训练负担;④训练集中早期样本易对新样本识别效果产生影响.针对上述问题,本文引入用户正、负兴趣集的概念,实现了一种结合用户兴趣集的在线垃圾邮件快速识别新方法.

2 本文方法

给定训练集 A_0 及增量邮件集合 $S_i(i=1,..,n_s,n_s$ 为增量集数目),图 1 给出了算法执行流程.图中具体步骤可描述如下:

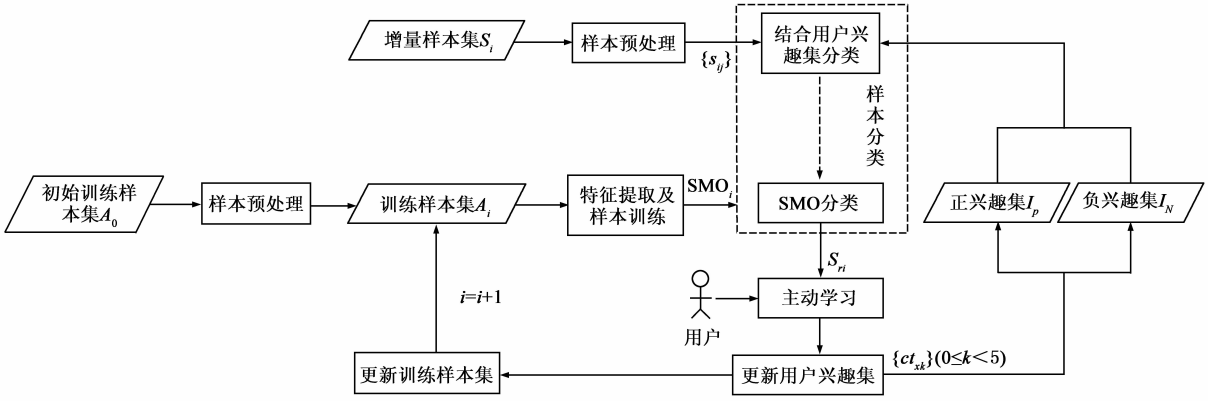


图1 本文方法流程图

2.1 样本预处理

先去除 A_0 及 S_i 中每封邮件中附件、标签、邮件头、停用词等信息,接着使用 Porter Stemming 算法进行词形还原[11].

2.2 特征提取及样本训练

传统方法单纯将单词作为特征用于文档表示,但是某些特殊符号(如:"!!!!!"、"://")等也是垃圾邮件区别于合法邮件的典型特征.本文同时将单词及文本中特殊符号作为待选特征,在特征提取过程中,使用类间-类内综合测量特征选择方法(feature selection based on comprehensive measurement both in inter-category and intra-category,CMFS)[12]从训练集 $A_i(i\geq 0)$ 中选择 n_f 个特征词构成特征词集合 S_f.鉴于序列最小优化算法(Sequential Minimal Optimization,SMO)能快速解决 SVM 分类过程面临的二次规划问题[13],本文使用 SMO 对 A_i 进行样本训练得到训练后分类器 SMO_i.

2.3 样本分类

针对增量集 S_i 中样本 s_{ij},先统计 s_{ij} 中每个单词 t_k 在 s_{ij} 中出现的频率 $p(t_k)$,接着将 s_{ij} 中所有单词集合 $\{t_k\}$ 按照 $p(t_k)$ 从大到小排序,选择前 m 个单词(本文取 $m=5$)并称它们为样本 s_{ij} 的内容主题词,记为 $\{ct_{ij0},ct_{ij1},\cdots,ct_{ijm-1}\}$,满足:若 $0\leq p\neq q\leq m-1$,则 $ct_{ijp}\neq ct_{ijq}$.

接着,本文给出以下定义:

(1)正兴趣:若用户认为 s_{ij} 为合法邮件,则 s_{ij} 中所有内容主题词称为该用户的正兴趣;

(2)负兴趣:若用户认为 s_{ij} 为垃圾邮件,则 s_{ij} 中所

有内容主题词称为该用户的负兴趣;

(3)正兴趣集:针对某用户生成的所有正兴趣称为该用户的正兴趣集;

(4)负兴趣集:针对某用户生成的所有负兴趣称为该用户的负兴趣集;

(5)用户兴趣集:针对某用户生成的正兴趣集、负兴趣集构成该用户的用户兴趣集.

假设当前正兴趣集为 I_P,负兴趣集为 I_N,阈值 $\Delta>0$,则本文样本分类过程可描述如下:

Step1:结合用户兴趣集分类:

如果: $\sum_{k=0}^{m-1} f_N(ct_{ijk})\times\alpha_k - \sum_{k=0}^{m-1} f_P(ct_{ijk})\times\alpha_k \geq \Delta$

则: s_{ij}为垃圾邮件;

如果: $\sum_{k=0}^{m-1} f_P(ct_{ijk})\times\alpha_k - \sum_{k=0}^{m-1} f_N(ct_{ijk})\times\alpha_k \geq \Delta$

则: s_{ij}为合法邮件;

否则:转 Step2;

上述过程中:

$$\left.\begin{aligned} f_P(ct_{ijk}) &= \begin{cases}1, & \text{如果 } ct_{ijk}\text{在 } I_P \text{ 中出现}\\ 0, & \text{否则}\end{cases}\\ f_N(ct_{ijk}) &= \begin{cases}1, & \text{如果 } ct_{ijk}\text{在 } I_N \text{ 中出现}\\ 0, & \text{否则}\end{cases}\\ \alpha_0 &= 0.3 > \alpha_1 = 0.2 = \alpha_2 = 0.2 = \alpha_3 = 0.2 = \alpha_4 = 0.1\\ \Delta &= \alpha_4 = 0.1\end{aligned}\right\} \quad (1)$$

Step2:使用分类器 SMO_i 对 s_{ij}进行分类.

可见,当用户兴趣集中缺乏用于识别样本的内容主题词或者 Step1 过程所得结果不可信时,样本的类别完全由 SMO_i 确定,此时,I_P、I_N 对于分类结果是没有影响的.

2.4　主动学习

为提高分类器的泛化能力,本文使用主动学习选择增量集合中分类最不确定的样本进行类别标注.如图 2 所示,由于待标注样本 x 处于分类超平面 $f(x)=0$ 的右侧,故由 MS 方法知,x 属于样本集 C_2的可能性更大.但是,观察发现 x 更符合样本集 C_1分布较离散的特点,故将其归结为 C_1类更加合理.进一步地,若将 x 推荐给用户标注并加入训练集,此时分类超平面将变为 $f'(x)=0$.由于样本 x'与 x 位置相近,引入 x'不仅不会明显改变新训练集分类超平面的位置,反而给用户带来了额外标注负担.

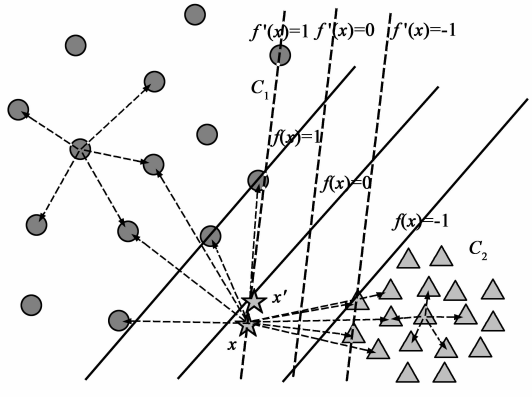


图2　样本分布对分类结果的影响

为解决上面问题,本文考虑了训练集中样本分布特点,提出了一种基于样本密度及改进角度差异的样本分类确定性评价新方法.若 A_i 中垃圾邮件、合法邮件集合分别为 A_{si}、A_{hi},针对增量集 S_i 中样本 s_{ij}($j=1,\cdots,|S_i|$,$|S_i|$为 S_i 中样本总数),本文样本分类确定性评价过程描述如下:

Step1:计算训练集 A_{si}、A_{hi}的样本密度 ρ_{si}、ρ_{hi}:

$$\rho_{si} = \frac{1}{|A_{si}|}\sum_{x\in A_{si}}\frac{1}{q}\sum_{x_j\in Neighbour(x,q,A_{si})} dis(x,x_j). \quad (2)$$

$$\rho_{hi} = \frac{1}{|A_{hi}|}\sum_{x\in A_{hi}}\frac{1}{q}\sum_{x_j\in Neighbour(x,q,A_{hi})} dis(x,x_j). \quad (3)$$

式(2)、(3)中,$dis(x,x_j)$表示邮件 x 与邮件 x_j间的距离,使用欧式距离表示;$Neighour(x,q,A_{si})$、$Neighour(x,q,A_{hi})$分别表示样本集 A_{si}、A_{hi}中与 x 距离最近的前 q 个样本(本文取 $q=5$).

Step2:选择 $|S_i|\alpha$(本文取 $\alpha=0.05$)封邮件(记为 S_{ri})推荐给用户进行类别标注,具体过程如下:

2.1 初始化.

令: $S_{ri}=$ null,临时样本集合 $R_i=$ null.

令:变量 $Count=|S_i|\alpha$.

2.2 选择第一个待推荐样本.

2.2.1 计算 S_i 中每个样本 s_{ij}与 A_{si}(A_{hi})之间的距离 d_{si}(d_{hi})

$$d_{si}(s_{ij}) = \frac{1}{q}\sum_{s\in Neighbour(s_{ij},q,A_{si})} dis(s_{ij},s) \quad (4)$$

$$d_{hi}(s_{ij}) = \frac{1}{q}\sum_{s\in Neighbour(s_{ij},q,A_{hi})} dis(s_{ij},s) \quad (5)$$

2.2.2 按照式(6)计算 S_i 中每个样本 s_{ij}对应 p 值:

$$p(s_{ij}) = \left|\frac{d_{si}(s_{ij})}{\rho_{si}} - \frac{d_{hi}(s_{ij})}{\rho_{hi}}\right|. \quad (6)$$

2.2.3 将具有最小 p 值的样本加入样本集 R_i,并将此样本从样本集 S_i 中移除.

2.3 选择剩余的待推荐样本.循环执行下面过程,直到条件 $|R_i|>Count$ 成立:

2.3.1 针对 S_i 中每个样本 s_{ij},按式(7)计算 s_{ij}分类确定性大小 $F(s_{ij})$:

$$F(s_{ij}) = \frac{p(s_{ij})\pi}{\min\limits_{s_v\in R_i}(\arccos(s_{ij},s_v))}. \quad (7)$$

其中，$\| s_{ij} \| = \sqrt{s_{ij} \cdot s_{ij}}$，$\| s_v \| = \sqrt{s_v \cdot s_v}$，

$$\cos(s_{ij}, s_v) = \frac{s_{ij} \cdot s_v}{\| s_{ij} \| \| s_v \|}.$$

2.3.2 将具有最小 F 值的样本加入样本集 R_i，并将此样本从样本集 S_i 中移除.

由式(6)知，本文综合考虑了样本与它在不同类别中最近邻样本的平均距离与不同样本集的样本密度特点，因此相对于 MS 方法更适用于与图 2 中 x 样本类似的情况；式(7)借助反余弦函数，通过计算向量间实际夹角量化两个样本之间的相似性，避免了传统 AD 方法单纯使用 $|\cos(s_{ij}, s_v)|$ 表示样本间角度差异导致无法处理 $\cos(s_{ij}, s_v) < 0$ 的情况.

2.5 更新用户兴趣集

在本文样本分类过程中，S_i 中样本 s_{ij} 可能满足以下 4 种情况：

(1) $\frac{d_{si}(s_{ij})}{\rho_{si}} > \frac{d_{hi}(s_{ij})}{\rho_{hi}}$ 且 s_{ij} 被分类为合法邮件.

(2) $\frac{d_{si}(s_{ij})}{\rho_{si}} < \frac{d_{hi}(s_{ij})}{\rho_{hi}}$ 且 s_{ij} 被分类为合法邮件.

(3) $\frac{d_{si}(s_{ij})}{\rho_{si}} > \frac{d_{hi}(s_{ij})}{\rho_{hi}}$ 且 s_{ij} 被分类为垃圾邮件.

(4) $\frac{d_{si}(s_{ij})}{\rho_{si}} < \frac{d_{hi}(s_{ij})}{\rho_{hi}}$ 且 s_{ij} 被分类为垃圾邮件.

若 s_{ij} 满足情况(2)、(3)，则认为 s_{ij} 分类错误，否则认为 s_{ij} 分类正确. 本文将 S_i 中所有样本按照分类确定性从大到小排序，选择 S_i 中前 $\beta|S_i|$（本文取 $\beta = 0.05$）封分类错误的样本记作 S_{fi}，前 $\beta|S_i|$ 封分类正确的样本记作 S_{ti}. 先将 S_{fi} 集合中样本类别标签进行反转，接着提取 S_{fi}、S_{ti} 及用户标注集合 S_{ri} 中每个样本 x 的内容主题词集合 $\{ct_{xk}\}(0 \leqslant k < 5)$，最后，按照下面的过程更新正负兴趣集 I_P、I_N 及训练样本集：

如果：x 标签为合法邮件

则：$I_P \leftarrow \{ct_{xk}\} \cup I_P$，$I_N \leftarrow I_N - \{ct_{xk}\}(0 \leqslant k < 5)$

如果：x 标签为垃圾邮件

则：$I_N \leftarrow \{ct_{xk}\} \cup I_N$，$I_P \leftarrow I_P - \{ct_{xk}\}(0 \leqslant k < 5)$

2.6 更新训练样本集

本文将 S_i 中 S_{fi}、S_{ri} 及 S_{ri} 集合加入到训练集 A_i 生成更新后训练集 A'_i. 为避免新训练集中的那些代表用户早期兴趣的支持向量样本对新样本识别产生影响，本文综合考虑了样本与分类超平面距离及样本加入训练集时间，选择那些离分类超平面距离最远且最先被加入训练集的样本进行淘汰. 给定训练集中样本 x，本文定义了 x 价值评价函数 $V(x)$：

$$V(x) = \left(\frac{t(x) \times |f|_{\max}}{t_{\text{current}} \times |f(x)|} \right)^{\varphi} \tag{8}$$

其中，$t(x)$ 为样本 x 加入训练集时对应增量集序号，t_{current} 为当前增量集序号；$|f(x)|$ 为 x 到分类超平面的距离，即 SVM 决策函数值[8]，$|f|_{\max}$ 为训练集中所有样本到分类超平面距离的最大值；φ 为权重系数，取 $\varphi = 1$. 当 A'_i 中样本数目 $|A'_i|$ 大于某阈值 $SN_{\max}$ 时，按照式(8)选择其中具有最小 V 值的 $|A'_i| - SN_{\max}$ 个样本进行移除即得到更新后训练样本集 A_{i+1}.

3 实验结果与分析

3.1 实验条件

使用样本集 TREC2007[14]（包含 50199 封垃圾邮件，25220 封合法邮件）作为实验数据集. 设置特征提取对应特征向量维数 $n_f = 600$. 无特殊说明情况下，测试样本集数目 $n_t = 5$，每个测试集中样本数目 $n_{ts} = 100$. 设置 SMO 分类器相关参数如下：惩罚因子：$C = 1.0$，容忍极限值 tol = 0.001，核函数 = RBFKernal. 为仿真在线邮件增量学习过程，在每次增量学习、测试之前由用户自行标注每个增量集、测试集中样本类别，在此基础上保证初始训练集 A_0、增量集 $S_i(0 \leqslant i < n_s)$、测试集 $T_j(0 \leqslant j < n_t)$ 满足图 3 中时间先后关系. 另外，使用垃圾邮件识别准确率（Spam Precision，SP）和垃圾邮件召回率（Spam Recall，SR）衡量垃圾邮件识别算法精度[12].

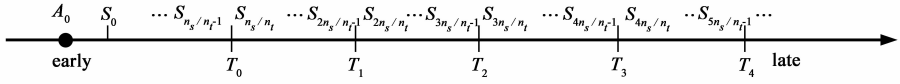


图3 初始训练集 A_0、增量样本集 S_i 及测试集 T_j 选择方法

3.2 算法耗时分析

若当前训练集样本数为 n_a，支持向量数目为 n_s，特征维数为 n_f，增量集 S_i 样本数目为 n_i，本文分别从以下两个方面进行算法耗时分析.

3.2.1 样本训练时间复杂度分析

由文献[15]知，在样本数为 n 的训练集上，传统 SVM 分类方法训练时间复杂度为 $T_{\text{T-SVM}} = O(n^3)$. 若当前增量集中 t 个样本即将被加入训练集，则针对本文训练时间复杂度 $T_{\text{T-ISVM}}$ 讨论如下：

情况 1：$n_a + t < SN_{\max}$：$T_{\text{T-ISVM}}$ 与 $T_{\text{T-SVM}}$ 相等，即：

$$T_{\text{T-ISVM}} = T_{\text{T-SVM}} = O((n_a + t)^3) \tag{9}$$

情况 2：$n_a + n_i > n_a + t \geq SN_{\max}$：此时，样本训练分两个过程执行：

Step1：先按照式(8)选择训练集中 $n_a + t\text{-}SN_{\max}$ 个样本进行淘汰，对应时间复杂度 $T_{\text{T-ISVM1}}$ 为：

$$T_{\text{T-ISVM1}} = O(2(n_a + t)n_s n_f) + O(n_a + t)$$

$$+ O((n_a + t)\log_2(n_a + t)) \quad (10)$$

Step2:对 $SN_{\max}$个样本进行 SVM 训练.时间复杂度为:

$$T_{\text{T-ISVM2}} = O(SN_{\max}{}^3) \quad (11)$$

对于传统增量学习方法,n 将无限制增加,而对文本而言,$n_a \leqslant SN_{\max}$,$n_s < n_a + t$;又由:$SN_{\max} < n$、$n_i < n$,因此:

$$\begin{aligned} T_t(n) &= T_{\text{T-ISVM1}} + T_{\text{T-ISVM2}} \\ &= O(2(n_a + t)n_{sv}n_f) + O(n_a + t) \\ &+ O((n_a + t)\log_2(n_a + t)) + O(SN_{\max}{}^3) \\ &< 2O((SN_{\max} + n_i)^2 n_f) + O(SN_{\max} + n_i) \\ &+ O((SN_{\max} + n_i)\log_2(SN_{\max} + n_i)) + O(SN_{\max}{}^3). \end{aligned} \quad (12)$$

针对特定 n_i,$T_t(n) = O(1) < O(n^3)$.可见,本文方法样本训练时间复杂度明显低于传统 SVM 的增量学习方法.

3.2.2 样本分类耗时分析

若两个实数比较所需时间为 t_1,相加所需时间为 t_2,相减所需时间为 t_3,相乘所需时间为 t_4.依据文献[8]知,针对样本 x 的类别决策函数为:

$$\operatorname{sgn}(f(x)) = \operatorname{sgn}\left(\sum_{i=1}^{n_a} \alpha_i^* y_i K(x, x_i) + b^*\right)$$

$$b^* = y_j - \sum_{i=1}^{n_a} \alpha_i^* y_i K(x, x_j), 0 < \alpha_j < C \quad (13)$$

其中,$K(x, x_i)$为核函数,α_i^* 为样本 x_i 对应的拉格朗日乘子,C 为惩罚因子,y_i 为样本 x_i 类别.参考线性核函数公式[6],可估计在 RBF 核函数情况下式(13)耗时 $T_{\text{C-SVM}}$为:

$$T_{\text{C-SVM}} = 2n_s(n_f + 2)t_4 + \delta \quad (14)$$

其中,$\delta > 0$.进一步地,记用户正、负兴趣集合 I_p、I_n 中内容主题词数目分别为 n_p、n_n.由 2.3 节 Step1 知,本文单纯使用兴趣集进行样本分类耗时 $T_{\text{C-I}}$为:

$$T_{\text{C-I}} = 5(n_p + n_n)t_1 + 8t_2 + 2t_3 + 10t_4 \quad (15)$$

或者:

$$T_{\text{C-I}} = 5(n_p + n_n)t_1 + 8t_2 + 4t_3 + 10t_4 \quad (16)$$

由于两个实数比较、相加、相减耗时相对于乘法耗时可忽略不计,即 $t_1 \ll t_4$,$t_2 \ll t_4$,$t_3 \ll t_4$.于是:

$$T_{\text{C-I}} \approx 10t_4 \quad (17)$$

假设增量集 S_i 中由兴趣分类的样本数占 S_i 中总样本数的比例为 $m(0 \leqslant m \leqslant 1)$,因此本文分类方法对 S_i 中每个样本类别检测平均耗时为:

$$\begin{aligned} T_{\text{C-ISVM}} &= T_{\text{C-I}}m + (T_{\text{C-I}} + T_{\text{C-SVM}})(1 - m) \\ &= T_{\text{C-I}} + T_{\text{C-SVM}}(1 - m) \end{aligned} \quad (18)$$

即:当 m 满足下面式子时,$T_{\text{C-ISVM}} < T_{\text{C-SVM}}$成立:

$$m > \frac{T_{\text{C-I}}}{T_{\text{C-SVM}}} = \frac{10t_4}{2n_s(n_f + 2)t_4 + \delta} \quad (19)$$

一般而言,$\dfrac{10t_4}{2n_s(n_f + 2)t_4 + \delta}$为一个极小值,因此式(19)极易被满足.进一步地,m 越大,参与兴趣分类过程的样本就越多,$T_{\text{C-ISVM}}$与 $T_{\text{C-SVM}}$之间差距也就越明显.

3.3 实验结果

为验证本文方法的有效性,将它与几种典型方法进行对比.这几种方法包括:Hu 方法[8]、Leng 方法[9]、Chen 方法[10]与 Liu 方法[1].其中,使用 SMO 作为 Liu 方法中的域分类器,并取 Chen 方法中自学习阈值为 0.7.公平起见,保证上述方法在每次增量学习过程后加入训练集样本数与本文一致.

3.3.1 阈值 $SN_{\max}$的选择

给定$|A_0| = 100$、200、300、400、500,$|S_i| = 100$、200、300、400、500 $(0 \leqslant i < n_s)$,$n_s = 100$.首先定义算法平均召回率 sr'、平均准确率 sp' 如下:

$$sr' = \frac{1}{25n_t}\sum_{|A_0|}\sum_{|S_i|}\sum_{k=0}^{n_t} SR(|A_0|, |S_i|, T_k) \quad (20)$$

$$sp' = \frac{1}{25n_t}\sum_{|A_0|}\sum_{|S_i|}\sum_{k=0}^{n_t} SP(|A_0|, |S_i|, T_k) \quad (21)$$

其中,$SR(|A_0|, |S_i|, T_k)$、$SP(|A_0|, |S_i|, T_k)$分别表示当初始训练集样本数为$|A_0|$,增量样本集样本数为$|S_i|$时,使用本文方法针对测试集 $T_k(0 \leqslant k < n_t)$所得召回率、准确率.

为了获取最优 $SN_{\max}$值,本文在 $SN_{\max} \in [1000, 5000]$范围内进行统计实验,图 4 给出了本文方法在不同 $SN_{\max}$值下所得 sr'、sp' 值.由图知,当 $SN_{\max} = 2000$ 时,算法所得结果较其他情况明显偏高.为进一步比较准确地确定出一个有效的 $SN_{\max}$取值范围,本文接着在[1500,2500]区间内按步长 100 对 $SN_{\max}$取值并再次进行统计实验.结果显示,当 $SN_{\max} = 1900$ 时 sp'、sr' 均取最大值 0.978.因此,这里取 $SN_{\max} = 1900$ 作为最优 $SN_{\max}$值.

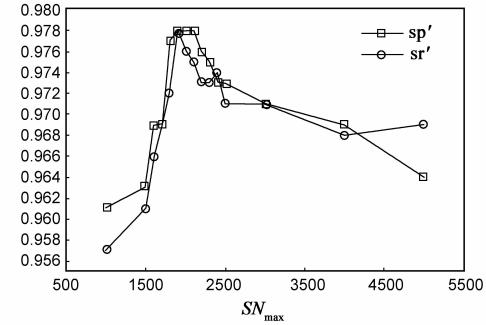
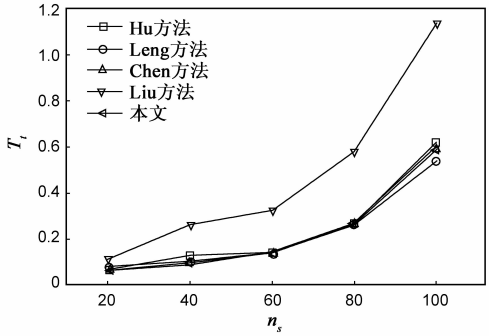


图4 阈值$SN_{\max}$选择

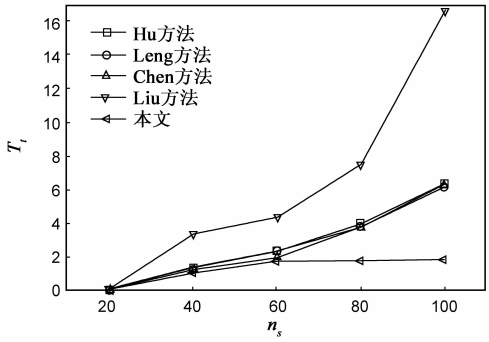
3.3.2 算法耗时比较

令初始训练集样本数$|A_0| = 200$,定义样本训练平

均耗时 T_t 及样本分类平均耗时 T_c 如下：



(*a*) $|S_i|$=50



(*b*) $|S_i|$=300

图5 $|S_i|$不同时算法所得T_t值比较

$$T_t = \frac{1}{n_s}\sum_{i=0}^{n_s-1} T\_Tra(A_i) \tag{22}$$

$$T_c = \frac{1}{n_t \times n_{ts}}\sum_{i=0}^{n_t-1}\sum_{j=0}^{n_{ts}-1} T\_Cla(t_{ij}) \tag{23}$$

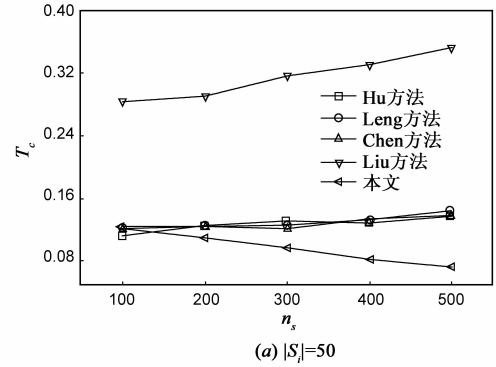
式(22)、(23)中，$T_Tra(A_i)$为对训练集 A_i 进行 SVM 训练耗时，$T_Cla(t_{ij})$为对测试集 $T_i(0\leqslant i < n_t)$中第 j 个样本 t_{ij}分类耗时.

图 5 给出了当 $n_s\in[0,100]$，$|S_i|$分别取 50、300 时，不同算法对应 T_t 值.由图知，Liu 方法所得 T_t 值较其他三种方法明显偏高，原因在于该方法在每次训练过程均需对每个子分类器执行 SVM 训练，因此其训练耗时较单个 SVM 分类器明显.观察图 5(a)发现，当$|S_i|=50$时，由于本文样本集规模最大时对应样本数量仍小于$SN_{\max}$，故其表现与 Hu 方法、Leng 方法及 Chen 方法近似；观察图 5(b)发现，除 Liu 方法外所有方法在 $n_s\in[20,60]$区间上表现一致，但当 $n_s>60$(对应本文训练集样本数量大于 $SN_{\max}$)时，Hu 方法、Leng 方法、Chen 方法对应 T_c值增长迅速，而本文方法 T_c却收敛于一个稳定值(约 1.8)，这说明：本文样本淘汰策略能有效控制训练样本集规模，对于降低在线邮件训练耗时十分有效.

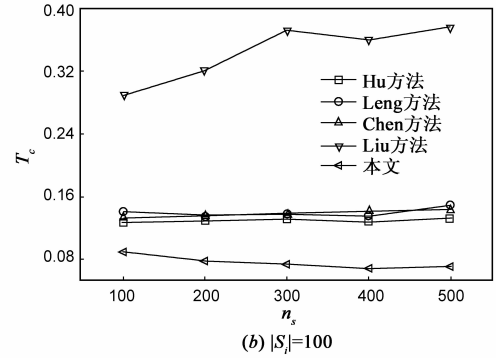
当 $n_s\in[0,500]$，$|S_i|$取不同值时，使用不同算法分别计算相应 T_c 值，结果如图 6 所示.由图 6(a)知，由于 Hu 方法、Leng 方法、Chen 方法均直接使用 SVM 进行样本分类，故这些方法所得 T_c 值近似且基本稳定不变；Liu 方法每次分类需综合各个域分类器的结果来获得样本类别，故算法耗时较其他方法偏大；本文在样本分类过程使用兴趣分类，避免直接对样本执行耗时较大的 SVM 分类过程，有效降低了样本分类时间.随着 n_s 增大，本文方法中用户兴趣集内容逐渐丰富，使得本文相对于其他算法优势越来越明显.由图 6(b)知，当 $n_s\geqslant 400$ 时，本文所得 T_c值最小且趋于稳定，原因在于：当用户正负兴趣集规模增长到一定程度后，样本分类任务直接由兴趣分类过程完成，故能有效避免传统 SVM 分类带来的复杂计算.

3.3.3 算法精度比较

令邮件初始训练集样本数为$|A_0|$ = 100、200、300、400、500；增量集 S_i样本数为$|S_i|$ = 100、200、300、400、500.分别当增量集数量 n_s = 20、40、60、80、100 时对 $T_k(1<k\leqslant n_t)$进行测试.图 7(a)、7(b)分别显示了 10 个用户对应于不同 n_s所得 sr 及 sp 均值(分别记为 sr_a、sp_a).由图知，与 Leng 方法与 Chen 方法相比，结合 AD 进行待标注样本选择的 Hu 方法所得结果偏高，原因在于 Hu 方法选择的待标注样本完全由专家标注且其中所含冗余样本较少，故使得分类器识别精度提升明显；随着 n_s 的增加，本文所得 sr_a、sp_a与 Liu 近似且均明显高于其他方法，进一步验证了本文在保证邮件识别精度方面的有效性.

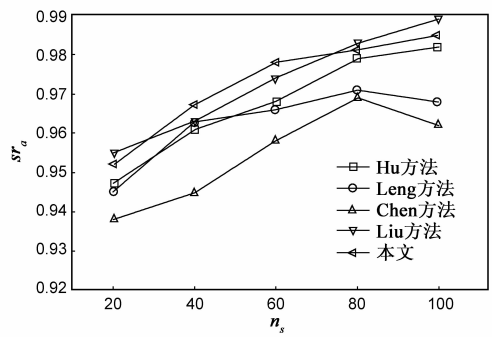


(*a*) $|S_i|$=50

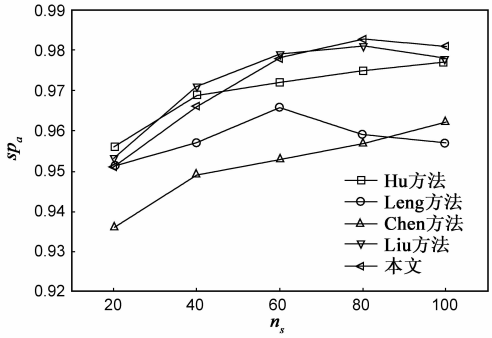


(*b*) $|S_i|$=100

图6 $|S_i|$不同时算法所得T_c值比较



(a) sr_a



(b) sp_a

图7　不同算法所得sr_a、sp_a值比较

3.3.4　样本标注负担比较

令初始邮件训练集样本数 $|A_0| = 300$;增量集 S_i 样本数 $|S_i| = 300$;测试集数目 $n_t = 10$;测试集样本数为 $n_{ts} = 200$. 当 $n_s \in [0,100]$ 时,统计不同方法对应 sr、sp 最大值(分别记为 sr_M、sp_M),结果见表 1. 可见,表中 sr_M、sp_M 最小值分别为 0.964、0.963. 进一步地,表 1 还给出了每种方法所得 sr、sp 首次不小于 0.96 时该方法提交给用户进行标注的样本总数目 n_r、n_p. 显然,本文方法所得 n_r、n_p 均最小,这主要是由于本文所提出的样本分类确定性评价新方法保证推荐给用户的样本更具代表性,故使得算法精度提升幅度明显;并且,使用由分类器自动识别出的确定分类错误的样本参与重复训练,在进一步提升算法识别能力的同时亦能有效降低人工标注负担.

表 1　不同算法对应的 sr_M、sp_M 值及样本标注负担比较

方法	sr_M	sp_M	n_r	n_p
Hu 方法	0.973	0.971	762	638
Leng 方法	0.969	0.965	987	1056
Chen 方法	0.964	0.963	796	625
Liu 方法	0.979	0.971	823	839
本文	0.976	0.975	423	476

4　结论

本文提出了一种在线垃圾邮件快速识别新方法. 主要贡献如下:①引入了用户正负兴趣集的概念,结合了正负兴趣集与 SVM 分类器对在线邮件进行分类;②提出了样本分类确定性评价新函数,在主动学习过程中将用户标注后样本及分类最确定性样本加入样本集进行重复训练. ③为避免训练集中早期样本对分类结果的影响,定义了样本价值评价新函数. 实验证明,本文邮件识别、训练速度较快,与一些代表性算法相比,本文能以较小的用户标注负担获得较高的邮件识别精度. 未来的研究工作主要从以下两个方面进行:①寻找更加高效的特征提取方式,保证在不影响算法分类精度的前提下降低特征提取计算耗时;②结合本体理论抽象邮件内容主题词,避免邮件正负兴趣集中语义相近的不同内容主题词给邮件识别过程带来歧义.

参考文献

[1] Liu W Y, Wang T. Online active multi-field learning for efficient email spam filtering[J]. Knowledge and Information Systems, 2012, 33(1): 117 – 136.

[2] Bertini J R, Zhao L, Lopes A A. An incremental learning algorithm based on the K-associated graph for non-stationary data classification[J]. Information Sciences, 2013, 246: 52 – 68.

[3] Costa J, Silva C, Antunes M, Ribeiro B. Customized crowds and active learning to improve classification[J]. Expert System with Applications, 2013, 40(18): 7212 – 7219.

[4] Syed N A, Liu H, Huan S, et al. Handling concept drifts in incremental learning with support vector machines[A]. Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence[C]. Stockholm, Sweden, 1999. 317 – 321.

[5] Wu C M, Wang X D, Bai D Y, et al. Fast incremental learning algorithm of SVM on KKT conditions[A]. Sixth International Conference on Fuzzy Systems and Knowledge Discovery[C]. Tianjin, China: IEEE Press, 2009. 551 – 554.

[6] Amayri O, Bouguila N. A study of spam filtering using support vector machines[J]. Artificial Intelligence Review, 2010, 34(1): 73 – 108.

[7] Tong S, Chang E. Support vector machine active learning for image retrieval[A]. Proceedings of the 9th ACM International Conference on Multimedia[C]. New York, USA: ACM, 2001. 107 – 118.

[8] Hu L S, Lu S X, Wang X Z. A new and informative active learning approach for support vector machine[J]. Information Sciences, 2013, 244: 142 – 160.

[9] Leng Y, Xu X Y, Qi G H. Combining active learning and semi-supervised learning to construct SVM classifier[J]. Knowledge-Based Systems, 2013, 44(5): 121 – 131.

[10] 陈荣,曹永锋,孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报, 2011, 37(8): 954 – 962.

Chen Rong, Cao Yong-feng, Sun Hong. Multi-class image classification with active learning and semi-supervised learning [J]. Acta Automatica Sinica, 2011, 37(8): 954 – 962. (in Chinese)

[11] Ali Haji N, Ibrahim N S. Porter stemming algorithm for semantic checking[A]. ICCIT 2012[C]. Chittagong University, Chittagong, 2012. 253 – 258.

[12] Yang J M, Liu Y N, Zhu X D, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization[J]. Information Processing & Management, 2012, 48 (4): 741 – 754.

[13] Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines[R]. Microsoft Research, 1998-04-21.

[14] Cormack G V. TREC 2007 spam track overview[A]. Proceedings of the 16th Text Retrieval Conference[C]. National Institute of Standards and Technology, Special Publication: 2007. 500.

[15] 丁文军,薛安荣.基于 SVM 的 Web 文本快速增量分类算法[J].计算机应用研究,2012,29(4):1275 – 1278.
Ding Wen-jun, Xue An-rong. Fast incremental learning SVM for web text classification[J]. Application Research of Computers, 2012, 29(4): 1275 – 1278. (in Chinese)

作者简介

王友卫 男. 1987 年 5 月出生,山东临沂人,吉林大学计算机科学与技术学院博士研究生,从事垃圾邮件处理、数字水印技术和 PDM 方面的有关研究.
E-mail: wyw4966198@126.com

刘元宁 男. 1962 年 9 月出生,辽宁抚顺人,博士,现为吉林大学计算机科学与技术学院教授、博士生导师,从事生物信息学、模式识别、图像处理、垃圾邮件行为识别和 PDM 研究.
E-mail: lyn@jlu.edu.cn