中国科学: 信息科学 2018 年 第 48 卷 第 12 期: 1681–1696
SCIENTIA SINICA Informationis
论文

基于客户端的个性化邮件再过滤系统

徐丹丹, 陈松灿*

南京航空航天大学计算机科学与技术学院, 南京 211106
* 通信作者. E-mail: s.chen@nuaa.edu.cn

收稿日期: 2018–05–28; 接受日期: 2018–08–22; 网络出版日期: 2018–12–04

国家自然科学基金项目 (批准号: 61672281, 61472186) 资助

摘要 电子邮件是一种必不可少的通讯工具, 但是众多的垃圾邮件会严重影响用户的工作和生活, 甚至还会造成财产损失. 由于兴趣、爱好的不同, 用户对垃圾邮件的定义可能存在巨大差异, 因此实现个性化垃圾邮件过滤成为目前邮件过滤领域研究的重要课题. 当出现邮件错滤情况, 用户不得不手动修改, 这给用户体验带来了极大不便. 为了有效解决上述问题, 并实现个性化邮件过滤及错滤邮件自动修改等功能, 本文结合规则和统计方法提出了一种基于客户端的个性化邮件再过滤系统. 大部分现有的垃圾邮件过滤器仅对邮件数据流进行在线过滤, 而未考虑不同邮箱的邮件类先验存在差异和类不平衡问题, 本过滤系统首先对进入收件箱和垃圾箱的邮件进行分别处理, 然后基于多任务学习原理设计了两个互学习的过滤器分别用于收件箱和垃圾箱的邮件再过滤, 并对错滤邮件进行自动修改. 同时, 为保证在随时间变化的用户兴趣点和邮件数据分布情况下过滤器的性能, 设计了结合重要性加权的多窗口学习框架, 从而有效实现了过滤器的动态自适应. 最后, 在 TREC 2006c 和 2007p 数据集上验证了我们所提出的过滤器拥有显著的过滤效果.

关键词 垃圾邮件过滤, 个性化邮件再过滤, 自动修正, 重要性加权, 多任务学习, 多窗口学习框架

1 引言

电子邮件作为一种快速、有效、便捷的交换信息的方式, 成为人们生活、学习、工作必不可少的通讯工具. 人们对电子邮件的依赖也导致 “非法邮件” (即垃圾邮件) 的产生. 这些垃圾邮件通常指的是未经允许大量发送到用户邮箱的邮件. 一些早期的研究表明, 垃圾邮件已经占据所有邮件的 80% [1], 近期统计显示其比例可能达 90% [2], 不仅浪费网络传输带宽, 增加了互联网使用成本同时也影响到人们日常工作和生活. 近年来垃圾邮件过滤技术得到了快速的发展, 一定程度上缓解了非法邮件对用户的影响.

引用格式: 徐丹丹, 陈松灿. 基于客户端的个性化邮件再过滤系统. 中国科学: 信息科学, 2018, 48: 1681–1696, doi: 10.1360/N112018-00138
Xu D D, Chen S C. A personalized mail re-filtering system based on the client (in Chinese). Sci Sin Inform, 2018, 48: 1681–1696, doi: 10.1360/N112018-00138

过去几十年中, 已有众多有效的过滤器用于垃圾邮件过滤. 然而据调查 [3], 现有的主流邮件过滤系统仍存在错滤情况, 如垃圾箱中存入了正常邮件, 而收件箱收到了垃圾邮件. 这种邮件误判问题仍是邮件过滤领域未能解决的难题. 导致这种问题的主要原因可归结为如下 3 个方面. 首先, 垃圾邮件制造者为了躲避过滤器的检测, 不断地改变垃圾邮件的内容特征, 导致数据分布随时间发生改变. 其次, 垃圾邮件过滤可看作是面向文本二分类问题的一种, 但又不同于一般的文本分类, 因为垃圾邮件过滤存在很大程度上的个性化差异, 不同用户对同一封邮件可能有着截然不同的分类结果 [4], 全局统一的二值过滤标准不能满足所有用户对邮件的主观评判. 例如, 急需学习日语的张三对推广语言学习网站的相关邮件感兴趣, 将其作为正常邮件 (normal), 而李四则没有这样的需求, 所以同样内容的相关邮件对他而言就是垃圾邮件 (spam), 导致不同用户收到同样类型的邮件产生不同的类标记. 最后, 用户对于收到的邮件是否是垃圾邮件, 与其现阶段的兴趣点相关, 即不同时间段, 一个用户的兴趣点会发生变化, 对同类型的邮件根据主观因素会有个性化标记. 例如需要购买电脑的用户在购买之前, 将推销电脑的邮件归为正常邮件, 而当其购买电脑之后, 不再需要电脑促销相关的邮件, 认为这类邮件是垃圾邮件. 以上 3 个方面给垃圾邮件过滤带来很大挑战, 使得用户收件箱常会收到垃圾邮件, 正常邮件被放进垃圾箱, 所以解决以上 3 个问题成为过滤器设计的首要任务.

传统的服务器端垃圾邮件过滤系统主要通过将在基本语料库中学习到的过滤器直接应用于客户邮箱的方式实现垃圾邮件过滤. 通常该方式主要存在两方面不足. 首先, 这类过滤器学习的是垃圾邮件全局统一概念, 不能准确地反映出个体用户的兴趣特点, 存在很多误判情况, 必须手动修改才能校正, 严重影响用户体验. 这促进了个性化垃圾邮件过滤器的发展, 使其能够动态更新以识别和跟踪用户兴趣爱好的变化. 个性化过滤器根据用户的反馈信息分析用户当前的兴趣特点然后进行邮件过滤, 缓解传统方法严重的误判问题. 所以, 根据应用范围, 现有的垃圾邮件过滤器大致分为一般化过滤器和个性化过滤器.

邮件传输作为一种数据流形式, 也不可避免地存在概念漂移问题, 即用户兴趣、邮件和其标记的联合分布有可能随时间发生变化. 现有的一般化垃圾邮件过滤系统通常无法解决该问题.

不同于大部分现有的过滤器所处的场景, 首先邮件经现有成熟的一般化过滤系统初步过滤后, 我们对放进收件箱和垃圾箱的邮件分别再过滤, 根据用户的个性化特征, 检测是否存在误判邮件, 若存在, 则自动调整该邮件所处的邮箱. 由于分别过滤常会带来类不平衡问题, 所以我们基于多任务学习原理设计了两个互学习的过滤器分别用于收件箱和垃圾箱的邮件再过滤, 达到收件箱和垃圾箱两者的邮件特征共享, 缓解邮箱内类不平衡带来的过滤器性能下降问题.

本文提出一种基于客户端的个性化邮件再过滤系统. 我们的主要贡献如下:

(1) 为了解决统一二值分类标准不能满足用户对邮件主观判定问题, 我们设计了个性化过滤器来缓解严重的误判情况.

(2) 在系统设计过程中, 我们重点关注个体用户的兴趣点变化, 相比较全局过滤器, 再次误判的可能性降低. 为了跟踪用户不同时间段的兴趣点, 本系统设计多窗口框架, 并采用分析、比较不同窗口过滤器精度的方式, 为当前决策选择合适的过滤器.

(3) 由于垃圾邮件的特征分布会随着时间发生变化, 本文通过分析不同窗口的密度比检测特征分布是否发生变化, 并采用重要性加权方法实现过滤器的动态自适应.

(4) 我们结合规则和统计两个方法, 大大减少了整个过滤系统的运算量, 加快了过滤速度.

2 相关工作

垃圾邮件过滤技术根据已有垃圾邮件特征识别当前邮件是否正常, 正常则为 normal (标记为 0), 否则 spam (标记为 1). 垃圾邮件过滤基本过程请参考文献 [5] 中的图 1, 一旦被判定为 spam, 该邮件将被特殊处理, 取决于其技术是应用到客户端 (个性化) 还是服务器端 (一般化). 邮件过滤技术分为基于协议过滤和基于内容过滤两种 [6]. 基于协议过滤方法根据邮件发送认证协议识别邮件类型, 通常应用于服务器端, 例如反向地址解析、黑白名单和数字签名等. 基于内容过滤方法利用过滤器自动学习训练样本的特征实现对邮件的二分类, 这种方法经常使用机器学习算法提取垃圾邮件特征并建立模型, 通常应用于客户端过滤. 常用的算法包括朴素 Bayes [7~9]、支持向量机 [10,11]、决策树 [12,13]、神经网络等. 基于内容的过滤方法, 又细分成 3 种: 基于规则的分类 [12,13]、基于统计的分类 [7~11] 和基于规则和统计的分类 [6]. 本文提出的方法属于最后一种.

黑白名单, 分别是记录了已知的垃圾邮件发送者和可信任的邮件发送者的 IP 地址或者是邮件地址, 是常见的一般化过滤器之一. Spam Assassin 作为典型的公开过滤系统, 经过合适的参数调整, 能够连续保持优越的性能. 然而随着垃圾邮件发送者不断改进其邮件策略, 以上过滤器的性能通常会逐渐减低, 造成大量的错滤邮件情况发生. 文献 [14] 采用半监督分类器集成算法实现个性化垃圾邮件过滤器. 半监督学习基于有效的 SVM 和 NB 集成学习通用的标记邮件和个体用户无标记邮件以精确识别用户邮件. 文献 [2] 提出双层过滤结构: SMF (spam mail filter) 和 LMF (legitimate mail filter) 分别识别 spam 和 normal 邮件, 以解决伪垃圾邮件 (FP) 和伪正常邮件 (FN) 问题. 用户收件箱中正常邮件作为 LMF 训练样本, SMF 采用一般的过滤器, 个人邮件不包含在其训练样本中. 当前邮件先由 LMF 预测是否为 normal, 是则放进用户收件箱, 否则将可疑邮件传递给 SMF 进一步识别.

由于用户兴趣、邮件和其标记的联合分布有可能随时间发生变化, 这种模式称为概念漂移. 根据漂移的时长和频率分为: 突发性 (abrupt) 漂移、增量式 (incremental) 漂移、平缓式 (gradual) 式漂移和再现性 (recurring) 漂移 4 种 [15]. 目前概念漂移的解决方式根据是否基于检测器分为两种类型: 主动式和被动式 [16]. 主动式在检测器的辅助下检测是否有漂移出现, 然后决定是否需要模型更新和重置. 被动式通常被称为自适应分类器, 不断更新模型, 不管是否有漂移出现. 主动式能够有效检测突发性漂移, 而被动式主要针对平缓式漂移有较优效果.

为了适应动态环境, 过滤器须具有自适应能力. 文献 [6] 结合统计方法和规则方法提出一种新的个性化垃圾邮件过滤技术, 该方法首先抽取邮件的语言特征和行为特征构建多个基于规则的子过滤器, 然后采用 SVM 集成学习方法设计过滤器预测邮件标记. 然而该方法并没有考虑邮件的内容分布随着时间发生变化 (即概念漂移) 的情况. 早在 1999 年, 文献 [17] 就提出增量式 SVM 能够很好地解决概念漂移问题. 文献 [18] 又提出一种基于支持向量机的垃圾邮件在线识别新方法, 即结合传统增量学习及主动学习理论, 先通过随机选择代表样本寻找分类最不确定的样本进行人工标注: 接着引入用户兴趣度的概念, 提出了新的样本标注模型和算法性能评价标准; 最后结合 “轮盘赌” 方法将标注后样本加入训练样本集. 最近, 文献 [11] 基于增量式 SVM 提出新的过滤方法, 在引用增量式 SVM 之前先启发式更新属性集, 使得分类模型有效学习变化后的数据分布. 利用信息增益 (IG) 对重训练样本进行特征选择, 生成新属性集, 替换原属性集中部分 IG 分值低的属性. 为避免用户反馈, 由文献 [19] 提出的个性化过滤器用一种无须用户反馈的自动化方法实现. 该方法根据有标记训练样本建立 spam 和 normal 词汇的统计模型, 基于个体用户收件箱无标记样本更新该模型, 适应邮件和标签的联合分布随着不同用户, 不同时间发生变化的状态. 所给测试邮件由词汇的统计模型计算其 spam 和 normal 评分值.

被动式不断更新模型造成资源消耗较大, 通常在错分率保持稳定的情况下, 数据分布处于稳定的

状态, 无需进行模型更新. 所以, 主动式为了解决以上问题, 提供漂移检测器, 即触发器, 当漂移出现时进行模型更新或重置. 传统的漂移检测器是否被触发是基于分类器错分率 [20~22], 错分率高低直接反应出近阶段数据流是否平稳. 文献 [23] 提出双窗口方法, 采用朴素 Bayes (NB) 作为分类器. 通过对比反应性和稳定性窗口分类器 (R, S) 的错分率检测是否有突发性概念漂移产生, 由当前分类性能好的分类器决策样本标记, 同时 S 模型不断更新, 自适应渐变式概念漂移, R 用作漂移检测器, 在检测到漂移时, 用 R 置换 S. 文献 [24] 借助双窗口想法, 将框架应用到回归问题中, 利用在线极值学习机 (OS-ELM) 和极值学习机 (ELM) 作为长窗口模型 L 和短窗口模型 S. 但是错误率仅仅反应了当前性能状态, 不能反应信息系统的信息量. 所以文献 [25] 提出用信息熵度量系统的不确定性, 当两个数据块的熵值之差达到阈值, 则认为漂移出现, 更新分类器权重.

据我们所知, 针对邮件误判的 3 个主要问题, 现有的垃圾邮件过滤器只考虑部分方面, 而本文提出的 PRFC (a personalized email re-filtering system based on the client) 对于 3 个主要问题都给出相应的解决方法.

3 基于客户端的个性化邮件再过滤系统

PRFC 主要目的是对已经过主流邮箱过滤后的数据流进行再过滤, 使得邮件误判率降低, 并将误判邮件自动召回. 为了加快过滤速度, 本系统结合规则和统计两种方法完成邮件过滤. 算法 1 给出过滤过程的伪代码, 并在后面的小节中详细解释.

阐述系统之前, 给出如下符号定义. 数据流 $T=\{T_{\mathrm{i}},T_{\mathrm{j}}\}$ 分流为 T_{i} 和 T_{j}, 分别表示进入收件箱 (inbox) 和垃圾箱 (junk box) 的邮件数据流. 过滤 T_{i} 和 T_{j} 的过滤器分别为 Filter_inbox 和 Filter_junkbox, 并且 Filter_inbox 和 Filter_junkbox 的框架结构一致, 设计多窗口框架 —— 长窗口 (LW)、短窗口 (SW) 和目标窗口 (TW). 令每个数据实例为 (x^k,y), $k=s$, 代表 $x^k\in\mathbb{R}^{d_{\mathrm{s}}}$ 是主题 (subject) 的向量化表示, $k=b$ 代表 $x^k\in\mathbb{R}^{d_{\mathrm{b}}}$ 是正文 (body) 的向量化表示, d_{s} 代表邮件主题向量化维度大小, d_{b} 代表邮件体向量化维度大小, $y=\{0,1\}$ 表示样本标记, 0 代表正常邮件 (normal), 1 表示垃圾邮件 (spam). $P(\cdot)$ 则表示概率分布. 令可信任发件人的信息库为 $\mathrm{Database}_{\mathrm{sender}}:\{(\mathrm{SenderID}_j,\mathrm{num}_j)\}_{j=1}^{K}$, 集合中每个元素是二元组, 其中 $\mathrm{SenderID}_j$ 表示发件人地址, num_j 表示发件人和该用户邮件往来次数, 且 K 固定. E 表示集成算法中最大集成尺度.

3.1 基于规则过滤

过滤一封邮件需要进行邮件解析、分词、去除停用词等无效词汇、向量化表示以及过滤器预测标记, 其中包含较长的数据预处理阶段, 时间消耗较大. 本文过滤系统所设定的场景是再过滤, 因此部分邮件可以简单地从发件人地址或者邮件主题中是否含有 “Re:” 字段等规则就可以判定是否是正常邮件, 是则无需进行较复杂的数据处理和预测. 当邮箱收到一封邮件时, 首先解析邮件, 并抽取邮件主要部分: 发件人地址、邮件主题、邮件正文. 我们设定如下两条简单规则.

规则 1. 如果发件人地址出现在 $\mathrm{Database}_{\mathrm{sender}}$ 中, 则判定该邮件为正常邮件 (本文不考虑存在可信任发件人发送垃圾邮件的情况); 否则, 检查是否符合规则 2.

规则 2. 检查主题字段中是否有 “Re:” 或 “回复:” 关键词, 如果有则代表该邮件是一封回复件, 则判定该邮件为正常邮件; 否则, 由基于统计方法设计的过滤器判定邮件标记.

针对规则 1, 需要确保可信任发件人的可信度不低于可接受范围, 我们根据通讯频率和时长对信息库 $\mathrm{Database}_{\mathrm{sender}}$ 进行更新. 如果基于规则方法不能判定邮件标记, 则采用基于统计方法设计的过

算法 1 PRFC 实现

输入: 有真实标记样本 $\{\mathrm{SW}^i\}_{i=1}^{N_m}$, 无真实标记样本 $\{\mathrm{TW}^{(i)}\}_{i=1}^{N_m}$, 解析后测试邮件 email, LW 的起始位置 T_0 和当前位置 T_1, L 模型的可接受错滤率阈值 ρ, 预测标记的置信度阈值 ξ, 已初始化的过滤器 Filter_inbox 和 Filter_junkbox;

if 根据 email['From'] 或 email['Re'] 并基于规则能判定 $y = 0$ **then**
  return y;
else $\{\mathrm{email} \in T_{\mathrm{i}}\}$
  利用 Filter_inbox 过滤器;
  基于主题过滤: 向量化 email['Subject'] 为 x^s;
  $\mathrm{TW}^{(N_m+1)} \leftarrow x^s$,$\mathrm{SW}^{(N_m+1)} \leftarrow \mathrm{TW}^{(1)}$;
  $\mathrm{SW}^{(i)} \leftarrow \mathrm{SW}^{(i+1)}$,$\mathrm{TW}^{(i)} \leftarrow \mathrm{TW}^{(i+1)}, i = 1, \ldots, N_m$;
  if SW 中出现类不平衡 **then**
    MTFL 学得模型参数 w^i, w^j;
    更新 L;
  end if
  利用 $\mathrm{SW}^{(N_m)}$ 增量学习 L;
  更新 $\{\mathrm{SW}^{(i)}\}_{i=1}^{N_m}$ 权重 $\{\alpha_i\}_{i=1}^{N_m}$;
  if 检测到协变量漂移发生 **then**
    重新计算 $\{\alpha_i\}_{i=1}^{N_m}$;
  end if
  利用加权 $\{\mathrm{SW}^{(i)}\}_{i=1}^{N_m}$ 更新 S;
  if $\mathrm{Err}(L) > \mathrm{Err}(S)$ 且 $\mathrm{Err}(L) > \rho$ **then**
    $L \leftarrow S$;
    $T_0 = T_1 - N_m$;
  else
    $T_1 = T_1 + 1$;
  end if
  $L.\mathrm{predict}(x^s) \rightarrow [y, \mathrm{confidence}]$;
  if $\mathrm{confidence} > \xi$ **then**
    return y;
  else
    基于正文过滤: 向量化 email['Body'] 为 x^b;
    同理, 重复 6∼24;
    return y;
  end if
else $\{\mathrm{email} \in T_{\mathrm{j}}\}$
  利用 Filter_junkbox 过滤器;
  同理, 重复 5∼31;
end if

输出: email 的预测标记 y.

滤器判定.

3.2 基于统计过滤

利用 Word2Vector [26] 将主题内容向量化, 获得 $x^k \in R^{d_{\mathrm{s}}}$, 主题过滤器将其作为输入变量, 并预测其标记, 如果所预测的标记置信度低于设定的阈值, 则下一步向量化邮件正文为 $x^k \in R^{d_{\mathrm{b}}}$, 由正文过滤器作最后的判定.

在传统的监督学习中, 通常假设训练数据集和测试数据集服从相同的分布. 然而这种基本假设忽

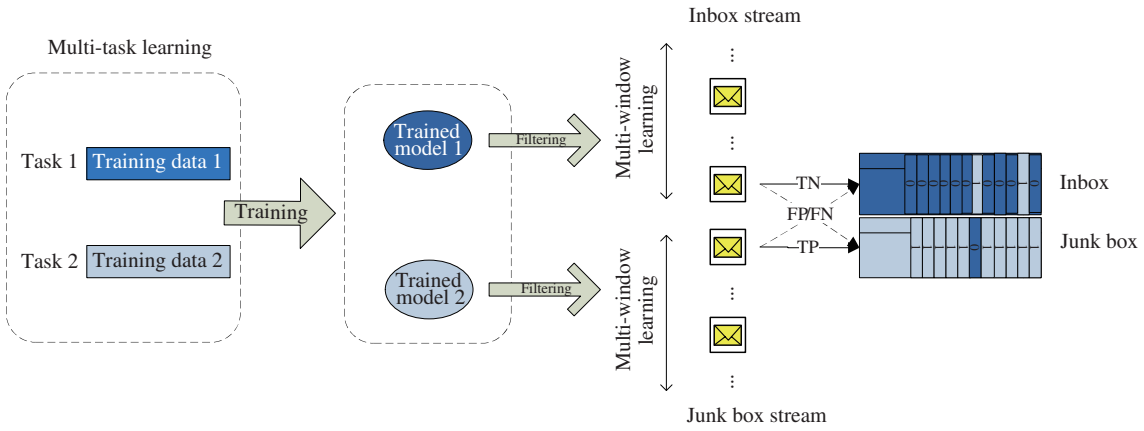

图 1 (网络版彩图) PRFC 系统框架图

Figure 1 (Color online) System framework of PRFC

略了真实数据的动态性, 因此严重影响了基于此假设的机器学习算法的性能. 在垃圾邮件过滤中, 文献 [27] 提出存在协变量漂移 (covariate shift), 即输入变量分布 $P(x^k)$ 在训练阶段不同于测试阶段, 而条件分布 $P(y|x^k)$ 的输出值则保持不变. 基于我们全面的观察, 并进一步对个体用户的邮箱系统数据流的动态性作出如下假定:

(1) 场景 1. 进入同一邮箱 (垃圾箱/收件箱) 的数据流, 不同时, 即

$$P_{t_1}(x^k) \neq P_{t_2}(x^k), \quad P_{t_1}(y|x^k) = P_{t_2}(y|x^k) \longrightarrow P_{t_1}(x^k, y) \neq P_{t_2}(x^k, y), \tag{1}$$

其中, $t_1 \neq t_2$ 表示不同时刻.

(2) 场景 2. 同一时刻, 进入不同邮箱 (垃圾箱和收件箱) 的数据流, 即

$$P_i(x^k) \neq P_j(x^k), \quad P_i(y|x^k) \neq P_j(y|x^k) \rightarrow P_i(x^k, y) \neq P_j(x^k, y), \tag{2}$$

$$P_i(y|x^k) \neq P_j(y|x^k) \Rightarrow \begin{cases} P_i(y=1|x^k) \leqslant P_j(y=1|x^k), \\ P_i(y=0|x^k) \geqslant P_j(y=0|x^k), \end{cases} \tag{3}$$

其中, $P_i(\cdot)$ 表示收件箱数据分布, $P_j(\cdot)$ 表示垃圾箱数据分布. 我们将以上两种场景的动态形式称为“广义虚漂移”. 而导致这样场景的其中一个原因是存在类不平衡, 即

$$P_i(y=0) \gg P_j(y=0), \quad P_i(y=1) \ll P_j(y=1). \tag{4}$$

在现有的垃圾邮件过滤器的设计中, 并没有对垃圾箱和收件箱中的邮件分开过滤, 而是当作一条动态数据流处理, 则丢失场景 2 信息. 为了能够精准地检测到误判邮件, 我们综合考虑两种场景的情况, 对即将进入收件箱的数据流和即将进入垃圾箱的数据流分别设计过滤器进行再过滤. 我们把这样的双数据流处理看作是两个任务, 并基于多任务学习学得两个过滤器, 每个过滤器结合重要性加权设计多窗口框架达到动态适应广义虚漂移的目的. 下面将对该方法进行详细阐述, 系统框架如图 1 所示.

3.2.1 多任务特征学习

为了解决广义虚漂移问题, 本文设计的邮件再过滤系统是对双数据流分别进行过滤, 如前文所述, 垃圾箱会进入极少量正常邮件, 收件箱也会存放少量垃圾邮件, 这样就导致这两条数据流存在类不平

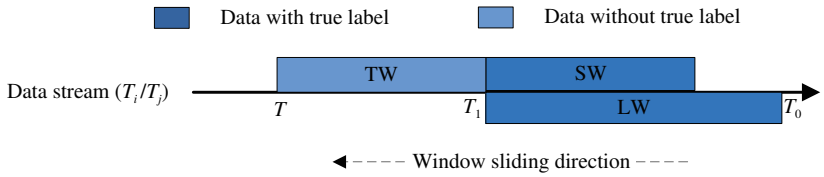

图 2 (网络版彩图) 多窗口框架示意图

Figure 2 (Color online) An illustration of multi-window framework

衡. 为了使得类不平衡环境下保证过滤器过滤精度在可接受范围内, 我们采用多任务特征学习 (multi-task feature learning), 让两个过滤器保持自身独有特征的情况下共享一部分共有特征, 克服单个过滤器对应类别训练样本少的问题, 以达到通过借鉴其他任务信息, 提升自身任务过滤效率的目的. 其中:

任务 1. 对即将投入收件箱的误判邮件自动召回到垃圾箱, 模型参数 $w^1 = w^i$;

任务 2. 对即将投入垃圾箱的误判邮件自动召回到收件箱, 模型参数 $w^2 = w^j$.

在模型构建过程中, 本文使用如下目标函数:

$$\min_{w^t} \sum_{t=1}^{T} \sum_{m=1}^{M_t} l(y_{tm}, (w^t)^T x_{tm}^k) + \frac{\lambda}{2} \sum_{t=1}^{T} \|w^t - \eta \times (w^1 + w^2)\|_2^2, \quad T = 2, \tag{5}$$

$$\min_{w^t} \sum_{t=1}^{T} \sum_{m=1}^{M_t} \|y_{tm} - (w^t)^T x_{tm}^k\|_2^2 + \frac{\lambda}{2} \sum_{t=1}^{T} \|w^t - \eta \times (w^1 + w^2)\|_2^2, \quad T = 2, \tag{6}$$

其中 $w^t \in R^{d_{\mathrm{s}}}$ 是 t 任务模型基于主题过滤参数, $w^t \in R^{d_{\mathrm{b}}}$ 为基于正文的过滤参数. 目标函数第 1 项是损失函数, 第 2 项是正则化项 (使得两个任务的决策面尽可能靠近, 达到共享一部分共同特征的目的), $\eta \geqslant 0$ 确定两个模型决策面靠近程度, $\lambda \geqslant 0$ 用于调整经验风险和正则化项的关系. w^1, w^2 指任务 1 和任务 2 互学习之前的模型参数.

3.2.2 结合重要性加权的多窗口框架

动态环境下, 用户兴趣点的转变和虚漂移成为个性化垃圾邮件过滤过程中的难题. 我们针对数据流 T_{i} 和 T_{j} 分别设计多窗口框架 (multi-window framework), 通过对不同窗口实施不同操作, 使得过滤器动态自适应. 本文实现 3 个滑动窗口 —— 长窗口 (LW)、短窗口 (SW) 和目标窗口 (TW), 其中, SW 和 TW 的大小相同且固定, 样本数为 N_m, 二者的样本信息被保留内存, 而 LW 内的样本将作为训练样本, 利用在线学习算法增量更新其模型 L, 所以无需连续保存 LW 中的样本. 我们假设进入长、短窗口的邮件会获得真实标记, 所以前两个窗口内的样本带有真实标记, 目标窗口内的样本有预测标记, 但是还未获得真实标记, 多窗口框架示意图如图 2 所示.

随着数据不断流动, 滑动窗口整体移动, 判断 SW 和 LW 中的模型 S 和 L 的样本预测结果是否正确, 并对比当前 S 和 L 的过滤性能, 以 Err(S) 表示模型 S 的错误率, 类似的, Err(L) 表示 L 的错误率. SW 内的样本代表的是最近短时期的邮件信息, 而 LW 中数据包含了近期长时间内的邮件信息. 当 Err(S) < Err(L) 时, 意味最近短时期的邮件信息更能代表当前的数据状态, 表明用户的兴趣点突然发生转变, 所以利用 S 模型置换 L 模型作为当前的决策模型; 当 Err(S) $\geqslant$ Err(L) 时, 说明当前数据状态比较稳定, LW 中训练样本量远大于 SW, 所以 L 模型过滤性能优于 S 模型. L 模型不断的更新, 能够适应用户兴趣点潜移默化的转变.

如场景 1 所描述, 发生协变量漂移也是邮件误判的主要原因之一. 目前为止, 相关研究人员提出许多方法应对协变量漂移问题. 文献 [28] 提出在 Hilbert 空间中通过对齐核矩阵来匹配漂移前后数据分布. 文献 [29] 提出了一个极大极小方法学习分类器. 最近, 利用重要性加权方法解决协变量问题避免了复杂的数据概率密度分布估计成为研究热点. 因此, 本文通过对 TW 和 SW 的数据进行密度比, 检测是否发生协变量漂移, 如有, 则对 SW 中的样本重新加权; 否则, 对 SW 中的样本权重在线更新.

重要性加权成为解决协变量漂移的主流方法, 该方法利用训练数据和测试数据的分布密度比 $\beta(x) = \frac{P_{\rm te}(x)}{P_{\rm tr}(x)}$ 定义学习过程中训练样本 x 的重要性权重, 然后将基于加权训练样本所学的学习器用于测试数据预测. 目前已有很多方法用来实现数据分布密度比估计, 例如矩匹配 [30]、基于 KL 散度的密度匹配算法 KLIEP [27, 31]、利用最小二乘方法拟合密度比 [32] 等. 本文采用与文献 [31] 类似的在线方法进行 SW 和 TW 窗口内数据密度比估计.

SW 和 TW 的数据分布分别为 $P_{\rm S}$ 和 $P_{\rm T}$, 样本量标记为 N_m, 则样本 x^k 密度比如下:

$$\beta(x^k) = \frac{P_{\rm T}(x^k)}{P_{\rm S}(x^k)}. \tag{7}$$

若 x 来自 SW, 则 $\beta(x)$ 值作为其权重用于过滤器训练. 为了避免对 $P_{\rm S}$ 和 $P_{\rm T}$ 进行密度估计, 算法直接估计重要性 $\beta(x^k)$, 模型如下:

$$\hat{\beta}(x^k) = \sum_{i=1}^{N_m} \alpha_i \varphi_i(x^k), \tag{8}$$

其中 $\{\alpha_i\}_{i=1}^{N_m}$ 是从样本数据中所学到的参数, 与 TW 中第 i 个样本 ${\rm TW}^{(i)}$ 相关, $\{\varphi_i\}_{i=1}^{N_m}$ 是基底函数, 且 $\varphi_i \geqslant 0$, 则 Gauss 核模型满足约束, 所以密度比估计表达式重写如下:

$$\hat{\beta}(x^k) = \sum_{i=1}^{N_m} \alpha_i K_\sigma(x^k, {\rm TW}^{(i)}), \tag{9}$$

$K_\sigma(\cdot,\cdot)$ 是带宽为 σ 的 Gauss 核, 即 $K_\sigma(x, x') = \exp\{-\frac{\|x-x'\|^2}{2\sigma^2}\}$, 实验通过交叉验证确定 σ 大小, 其中 $\|\cdot\|$ 表示向量的 2 范数.

参数学习. 由上可得, TW 的数据分布 $P_{\rm T}(x^k)$ 可估计为 $\hat{P}_{\rm T}(x^k) = \hat{\beta}(x^k)P_{\rm S}(x^k)$, 通过最小化真实分布 $P_{\rm T}(x^k)$ 和所估计的分布 $\hat{P}_{\rm T}(x^k)$ 的 KL 散度 (Kullback-Liebler divergence), 可学得参数 $\{\alpha_i\}_{i=1}^{N_m}$:

$${\rm KL}(P_{\rm T}(x^k) \parallel \hat{P}_{\rm T}(x^k)) = \int_D P_{\rm T}(x^k) \log \frac{P_{\rm T}(x^k)}{\hat{P}_{\rm T}(x^k)} {\rm d}x^k. \tag{10}$$

式 (10) 右边第一项与 $\{\alpha_i\}_{i=1}^{N_m}$ 无关, 所以目标函数转化为最大化第二项. 由于 $\hat{\beta}(x^k)$ 非负, 且 $\int_D \hat{P}_{\rm T}(x^k) {\rm d}x^k = 1$, 所以优化目标如下:

$$\begin{aligned} &\max_{\{\alpha_i\}_{i=1}^{N_m}} \frac{1}{N_m} \sum_{j=1}^{N_m} \log \sum_{i=1}^{N_m} \alpha_i K_\sigma({\rm TW}^{(j)}, {\rm TW}^{(i)}) \\ &{\rm s.t.} \quad \frac{1}{N_m} \sum_{j=1}^{N_m} \sum_{i=1}^{N_m} \alpha_i K_\sigma({\rm SW}^{(j)}, {\rm TW}^{(i)}) = 1, \quad \alpha_i \geqslant 0,\ i = 1, \ldots, N_m. \end{aligned} \tag{11}$$

基于梯度下降算法优化以上凸函数可以得到 $\{\alpha_i\}_{i=1}^{N_m}$, 然后根据式 (9) 为 SW 中每个样本加权, 以便后续模型学习和更新.

参数在线更新. 邮件过滤是在线应用, 所以我们要对过滤模型进行在线学习与更新. 考虑如下情况: 根据模型式 (11) 可知, 当 SW 进入一个新样本, 会改变其约束条件, 但是不会对优化目标产生直接影响: 而 TW 中出现新样本时, 会直接影响优化目标, 所以需要对 α 在满足约束的情况下进行更新.

基于文献 [33] 所提出的核方法在线学习技术对模型进行在线更新. 令 $E_i(\beta)$ 表示样本 $\mathrm{TW}^{(i)}$ 的经验误差, $E_i(\beta) = -\log\beta(\mathrm{TW}^{(i)})$. 由式 (11) 可以看出, 在约束条件下最小化 $\sum_{i=1}^{N_m} E_i(\beta)$, 就可算出 $\hat{\beta}$. 假设在再生核空间 H 中研究 β, 则有如下再生属性:

$$\langle \beta(\cdot), K(\cdot, x')\rangle_H = \beta(x'), \tag{12}$$

其中 $K(\cdot,\cdot)$ 为空间 H 中的再生核, 而 $\langle\cdot,\cdot\rangle_H$ 表示 H 中的内积运算. 令 $\tilde{E}_i(\beta)$ 表示正则化的经验误差:

$$\tilde{E}_i(\beta) = -\log\beta(\mathrm{TW}^{(i)}) + \frac{\lambda'}{2}\|\beta\|_H^2, \tag{13}$$

其中 $\lambda'(>0)$ 是正则化参数, $\|\beta\|_H$ 表示 H 空间范数运算. 核方法在线学习更新参数的基本思路是

$$\hat{\beta}' = \hat{\beta} - \eta'\partial_\beta \tilde{E}_{N_m+1}(\hat{\beta}), \tag{14}$$

其中 η' 表示学习率, ∂_β 是关于 β 的偏导. 利用基于 Gauss 核模型拟合的密度比 (式 (9)) 替换式 (14) 中的偏导, 上述更新规则可明确表示为

$$\hat{\beta}' = \hat{\beta} - \eta'\left(-\frac{K_\sigma(\cdot, \mathrm{TW}^{(N_m+1)})}{\hat{\beta}(\mathrm{TW}^{(N_m+1)})} + \lambda'\hat{\beta}\right). \tag{15}$$

所以根据式 (9), 参数 $\{\alpha_i\}_{i=1}^{N_m}$ 的在线更新如下:

$$\begin{cases} \hat{\alpha}_i' \leftarrow (1-\eta'\lambda')\hat{\alpha}_{i+1}, & i = 1,\ldots,N_{m-1}, \\ \hat{\alpha}_i' \leftarrow \dfrac{\eta'}{\hat{\beta}(\mathrm{TW}^{(N_{m+1})})}, & i = N_m. \end{cases} \tag{16}$$

漂移检测. 如上文所述, 数据分布 $P_{\mathrm{T}}(x)$ 可估计为 $\hat{P}_T(x) = \hat{\beta}(x)P_{\mathrm{S}}(x)$. 令 α_0 表示初始参数集, 而 α_t 表示 t 时刻参数集. $\hat{\beta}_0$ 和 $\hat{\alpha}_t$ 分别对应 α_0 和 α_t 的密度比. 利用以下的对数比率来度量测试阶段的数据分布和训练阶段的数据分布偏差:

$$S = \sum_{i=1}^{N_m} \ln\frac{P_{\mathrm{T}}(\mathrm{TW}^{(i)})}{\hat{\beta}_0 P_{\mathrm{S}}(\mathrm{TW}^{(i)})} = \sum_{i=1}^{N_m} \ln\frac{\hat{\beta}_t(\mathrm{TW}^{(i)})}{\hat{\beta}_0(\mathrm{TW}^{(i)})}, \tag{17}$$

如果 $S > \mu$, 则检测到漂移发生, 则重新计算 SW 中样本权重, 然后将其作为训练样本重新学习过滤器, 其中 μ 是自定义阈值. μ 的大小直接影响假警报 (false alarm) 发生的概率. 当 SW 或者 TW 中有新样本时, 更新参数 α.

4 实验与结果

为了仿真在线客户端邮箱系统, 我们将 TREC 2006c 和 TREC 2007p 数据集以 $N^i_{\mathrm{normal}}/N^i_{\mathrm{spam}} \approx 9$ 比例构造收件箱数据流, 以 $N^j_{\mathrm{normal}}/N^j_{\mathrm{spam}} \approx 17$ 比例仿真垃圾箱数据流, 其中 N^i_{normal}, N^i_{spam} 为收件箱内正常邮件和垃圾邮件数, 同理 N^j_{normal}, N^j_{spam} 表示垃圾箱中正常邮件和垃圾邮件数量. 同时, 为了突出个性化邮箱设定, 我们改变部分特定类型邮件原标记, 验证 PRFC 有效性.

表 1 实验数据集

Table 1 Experimental corpuses

Corpus	Normal	Spam	Total
TREC 2006c	21766	42854	64620
TREC 2007p	25220	50199	75419

4.1 实验数据

本文实验数据采用来自 Text Retrieval Conference (TREC) 的公开中文数据集 TREC 2006c 和英文数据集 TREC 2007p, 表 1 列出数据集的详细情况[1)]. 为了实现对邮件数据的预处理, 我们先利用 Python 中 email 库将每一封测试邮件解析为邮件发件人 ID、收件人 ID、主题和正文 4 部分. 同时, 在必要的情况下使用结巴分词去除主题和正文中的停用词、空格符、邮件附件等信息. 随着深度学习的发展以及 RNN, CNN 的陆续出现, 特征向量的构建将会由网络自动完成, 因此我们只要将分词后的邮件主题和正文的文本向量表示输入到网络中就能够自动完成特征的构建. 为了应对邮件主题和正文中有新特征出现的情况, 我们对向量化技术所依赖的特定语料库进行不断的更新, 提高过滤器的识别率.

4.2 实验评估方法

作为实用的过滤系统, 用户的满意度应该是唯一的评估标准. 因此该评估标准需将系统应用到个体用户邮箱, 由于用户的主观性和系统的客观性无法进行自动对比, 所以很难实现对用户满意度的评估. 我们通过改变数据集特定类型邮件标记, 仿真个性化邮箱数据实现对 PRFC 的评估.

对于用户而言, 相比较垃圾邮件被误判为正常邮件, 更不能接受正常邮件被放进垃圾箱, 所以使用 FPR (假正例率) 作为过滤系统的评估标准之一:

$$\mathrm{FPR} = \frac{N_{\mathrm{normal}\to\mathrm{spam}}}{N_{\mathrm{normal}\to\mathrm{normal}} + N_{\mathrm{normal}\to\mathrm{spam}}}.$$

同时, 为了检验本系统对类不平衡问题的鲁棒性, 引用两个不平衡分类问题的评价指标 G-mean 值以及 $F1$ 值:

$$\mathrm{G\text{-}mean} = \sqrt{\frac{N_{\mathrm{normal}\to\mathrm{normal}}}{N_{\mathrm{normal}\to\mathrm{normal}} + N_{\mathrm{normal}\to\mathrm{spam}}} \times \frac{N_{\mathrm{spam}\to\mathrm{spam}}}{N_{\mathrm{spam}\to\mathrm{normal}} + N_{\mathrm{spam}\to\mathrm{spam}}}},$$

$$F1 = 2 \times \frac{\mathrm{recall} \times \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}},$$

$$\mathrm{recall} = \frac{N_{\mathrm{spam}\to\mathrm{spam}}}{N_{\mathrm{spam}\to\mathrm{spam}} + N_{\mathrm{spam}\to\mathrm{normal}}}, \quad \mathrm{precision} = \frac{N_{\mathrm{spam}\to\mathrm{spam}}}{N_{\mathrm{spam}\to\mathrm{spam}} + N_{\mathrm{normal}\to\mathrm{spam}}},$$

其中 $N_{\mathrm{normal}\to\mathrm{normal}}$ 和 $N_{\mathrm{spam}\to\mathrm{spam}}$ 表示过滤器正确识别正常邮件和垃圾邮件的数量; $N_{\mathrm{normal}\to\mathrm{spam}}$ 是正常邮件被误判为垃圾邮件的数量; 类似地, $N_{\mathrm{spam}\to\mathrm{normal}}$ 是指垃圾邮件被误判为正常邮件的数量. 同时以精确度 Accuracy 评价过滤器性能.

根据 Text Retrieval Conference 统一评估垃圾邮件过滤器过滤性能, 指出采用指标 (1−ROCA)% 和 lam%. normal% 是正常邮件误判率, spam% 是垃圾邮件误判率. 为了保证 normal% 和 spam% 值能

1) 数据来源于 https://trec.nist.gov/data/spam.html.

同时很小, 所以用 (1−ROCA)% 和 lam% 作为过滤器评估指标. (1−ROCA)% 是以 spam% 为纵坐标, normal% 为横坐标的 ROC 曲线以上的面积大小; lam% 是逻辑平均误判率, 定义如下:

$$\text{lam\%} = \text{logit}^{-1}(\text{logit}(\text{normal\%})/2 + \text{logit}(\text{spam\%})/2),$$

$$\text{其中 } \text{logit}(x) = \log\frac{x}{1-x}, \quad \text{logit}^{-1}(x) = \frac{e^x}{1+e^x},$$

当 (1−ROCA)% 和 lam% 的值越小, 表示过滤器性能越好.

4.3 实验结果和分析

我们使用开发工具 Python3.5 实现本文所提出的过滤系统 PRFC. 过滤系统中 LW 和 SW 对应的模型 L 和 S 都采用集成算法实现, 以 SVM 作为基学习器. 为了获取最优参数, 我们随机从 TREC 2006c 选取 10000 个样本数据, 按照上文比例构造类不平衡的收件箱数据流和垃圾箱数据流.

4.3.1 参数敏感度分析

我们通过实验分析对邮件数据流 T_{i} 和 T_{j} 进行过滤的过滤器 Filter_inbox 和 Filter_ junkbox 的过滤性能对参数的敏感度. 为了观察扩大过滤器对参数的敏感度, 我们对过滤器评价指标结果作如下变换再绘制图 3:

$$\text{G-mean}' = (\text{G-mean} - 0.9) \times 10, \quad \text{Accuracy}' = (\text{Accuracy} - 0.9) \times 10.$$

从图 3(a) 中可以看出, 多任务学习算法中的参数 $\eta = 0.64$ 时两个过滤器的不平衡分类问题的评价指标都达到较好程度; 如图 3(b) 所示, Filter_inbox 在集成尺度 $E_i = 3$, 而 Filter_junkbox 在 $E_j = 6$ 时达到最高准确率; 图 3(c) 可以得到滑动窗 SW 和 TW 的大小 N_m, 权衡运行时间和过滤精确度, 确定过滤器 Filter_inbox 的 $N_m^i = 50$, Filter_junkbox 的 $N_m^j = 100$; 协变量漂移检测中阈值 μ 的大小直接决定是否会有假警报发生, 造成非必要的模型更新, 影响过滤性能, 如图 3(d) 表明, Filter_inbox 部分在 $\mu_i = 6$ 时, 过滤精准率最高, 同理, Filter_junkbox 中取 $\mu_j = 2$. 同时, 在下面的实验中, 我们设置参数 $\lambda = 0.01$, 并且设置重要性加权中的参数 $\lambda_i' = \lambda_j' = 0.01$, $\eta_i' = \eta_j' = 1$. 同时设置所有 L 模型的可接受错误率 $\rho = 0.1$, 基于主题预测标记的置信度阈值 $\xi = 0.9$.

4.3.2 过滤比例分析

本文过滤系统 PRFC 结合规则和统计方法识别进入特定用户邮箱的邮件是否是垃圾邮件. 如上文所述, 基于规则的方法包括检测发件人是否可信任、邮件主题内容开头是否有 “Re:” 或 “回复:” 字段; 基于统计方法包括根据向量化的主题判别是否为垃圾邮件以及基于向量化的正文内容识别邮件性质. 所以我们将整个系统分成 3 个部分: 基于规则过滤、基于主题过滤和基于正文过滤. 这 3 个部分形成一个多层次过滤器, 如果基于规则过滤不能确定测试邮件的类别, 则由基于主题过滤器识别; 同理, 如果基于主题过滤也不能明确该邮件是否是垃圾邮件, 则由基于正文内容的过滤器来判断. 通过统计不同过滤部分所识别的邮件数量, 分析每个部分的过滤性能.

将 PRFC 应用于中文数据集 TREC 2006c 上, 累计处理 T_{i} 和 T_{j} 的过滤器对应部分识别邮件的数据量, 同时计算每个部分的数据量占总样本数比例, 如图 4 所示. 从图中可以看出, PRFC 第 2 层 (基于主题过滤) 是整个过滤系统的主力, 所识别的邮件比例是 74.13%, 远大于第 1 层 (基于规则过滤) 和第 3 层 (基于正文过滤). 第 1 层的识别比例仅是 18.87% 的原因之一是可信任发件人信息集合 D_{sender}

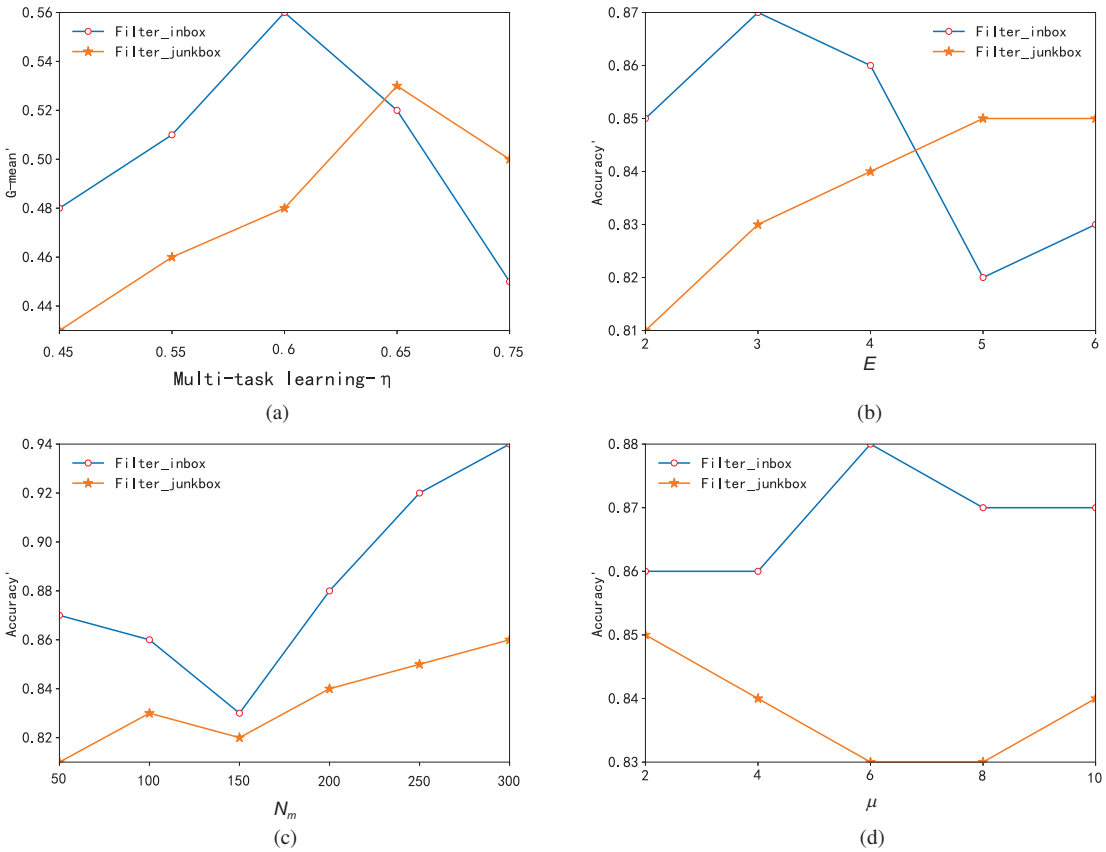

图 3　(网络版彩图) PRFC 应用于 TREC 2006c 数据集上的参数敏感度

Figure 3　(Color online) Parameter sensitivity of PRFC on TREC 2006c. (a) Sensitivity to the Multi-task learning parameter; (b) sensitivity to the ensemble size; (c) sensitivity to the sliding window size; (d) sensitivity to the threshold of drift detection

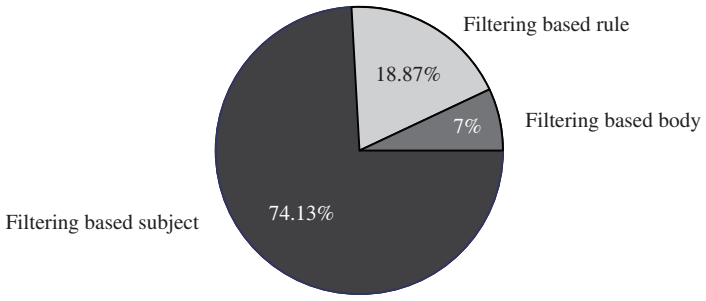

图 4　PRFC 基于不同部分的过滤比例

Figure 4　The proportion of filtration based on PRFC different parts

的最大容量是不变的, 这需要对其进行不断更新. 同时, 当用户的邮箱内有很少回复邮件时, 根据主题是否有 “Re:” 和 “回复:” 只能判定少量邮件为正常邮件. 第 2 和第 3 层模型结构类似, 但是通常情况下, 根据邮件主题就能判断是否是垃圾邮件.

表 2 多任务学习过滤器和单任务过滤器性能对比结果

Table 2 Multi-task vs. single task [a)]

Method	G-mean (↑)	$F1$ (↑)	(1−ROCA)% (↓)	Accuracy (↑)
Multi-task	0.9895	0.9921	0.0104	0.9896
Single task	0.9703	0.9775	0.0296	0.9705

a) "↑" means that the higher the value, the better the performance; "↓" means that the smaller the value, the better the performance.

表 3 在 TREC 2006c 和 TREC 2007p 数据集上评估不同算法的结果

Table 3 Evaluating different algorithms on TREC 2006c and TREC 2007p [a)]

Corpus	TREC 2006c				TREC 2007p			
Evaluation criteria	Accuracy (↑)	FPR (↓)	(1−ROCA)% (↓)	lam% (↓)	Accuracy (↑)	FPR (↓)	(1−ROCA)% (↓)	lam% (↓)
DISvm [34]	0.9594	0.0107	0.0383	2.73	0.9658	0.0087	0.0321	2.10
ROSVM [35]	0.9935	0.0036	0.0094	0.34	0.9848	0.0060	0.0108	0.86
MLC [36]	**0.9992**	0.0021	**0.0004**	**0.08**	0.9855	0.0056	0.0096	0.64
PRFC	0.9984	**0.0013**	0.0025	0.12	**0.9865**	**0.0053**	**0.0068**	**0.45**

a) "↑" means that the higher the value, the better the performance; "↓" means that the smaller the value, the better the performance. Bold fonts indicate the best results.

4.3.3 多任务与单任务性能对比分析

本实验从 TREC 2006c 数据集抽取一半的数据集用来对比本文基于多任务学习的过滤器和单任务情况下过滤器的过滤性能. 这里基于多任务过滤和单任务过滤之间的区别即是否将数据流 $T_{\rm i}$ 和 $T_{\rm j}$ 分开过滤. 如果将 $T_{\rm i}$ 和 $T_{\rm j}$ 看作是一条数据流识别, 则是单任务过滤; 反之则是多任务过滤.

实验从类不平衡分类问题评价指标 G-mean 和 $F1$ 以及 ROC 曲线以下面积 ROCA、精确度 Accuracy 4 个方面对比分流情况下多任务学习过滤器和分流前单任务过滤器的过滤性能. 如表 2 所示, 表明我们提出的基于多任务学习的过滤系统能够有效提高过滤性能, 并改善误判情况.

4.3.4 不同算法性能对比分析

将我们提出的 PRFC 与以下 3 种过滤方法通过实验进行性能对比.

(1) Disc.Inf.Sp.SVM(DISvm). 文献 [34] 为了应对文本分类过程中数据分布变化导致的概念漂移问题, 提出将数据特征空间基于判别信息量降至二维, 形成判别信息空间. 并在该空间学得 SVM 模型, 指出该模型对噪声和漂移敏感并用于文本分类.

(2) ROSVM. 文献 [35] 提出松弛的在线支持向量机 (relaxed online SVM, ROSVM) 模型, 该方法通过松弛约束条件, 在低成本的情况下显著加快过滤器训练速度, 并采用典型的在线学习方法 Online SVM 作为过滤器识别邮件类别.

(3) MLC. 文献 [36] 在 ROSVM 的基础上提出基于误判和低确定性 (misclassification and low-certainty, MLC) 的主动学习方法, 即选择被误判的邮件和不确定预测结果是否正确的邮件作为训练数据集, 降低训练成本. MLC 将文献 [35] 提出的 ROSVM 作为垃圾邮件过滤器.

我们在数据集 TREC 2006c 和 TREC 2007p 上验证 PREC 的可行性, 结果如表 3 所示. 从表中可以看出我们提出的过滤系统性能一定程度上优于 DISvm 和 ROSVM 算法. 虽然 DISvm 也能够缓

解数据概念漂移导致的过滤器性能下降问题, 但是邮件过滤作为在线应用需要对模型在必要情况下进行更新, 从而保证过滤精度在可接受范围内, 所以 DISvm 作为离线模型性能没有 PRFC 显著. PRFC 是针对在线应用, 没有必要第一时间获得邮件真实标记也能适应动态环境, 而 ROSVM 在标记延迟情况下并没有考虑广义虚漂移对过滤性能的影响. 由于 PRFC 基于多任务学习设计过滤器, 在一定程度上缓解类不平衡导致的误判, 所以在评价指标 FPR 上优于以上 3 种方法. 但是在 TREC 2006c 为数据集的实验中, PRFC 在其他指标上略差于 MLC, 表明我们提出的方法在中文邮件过滤中, 特征和训练样本的选择问题上需进一步改善. 从该实验中可以得出, PRFC 在中文邮件和英文邮件数据集上都能获得较好的过滤性能.

5 总结

我们结合规则和统计方法提出一种基于客户端的个性化邮件再过滤系统 (PRFC). 不同于一般的文本分类, PRFC 是对邮箱系统过滤后的双数据流进行再次过滤, 根据用户的个性化特征, 检测是否存在误判邮件, 并对误判邮件自动调整, 达到降低误判率的目的. 我们借助多任务学习使得对双数据流过滤的两个任务能够一起学习, 分别过滤. 同时, 本文提出垃圾邮件过滤存在 "广义虚漂移" 问题, 并在基于主题和正文过滤部分结合重要性加权方法设计多窗口框架以适应动态坏境下 "广义虚漂移" 问题. 实验结果表明, PRFC 在中文数据集和英文数据集上都能获得较好的过滤性能, 同时相对比与其他算法, PRFC 易于实现, 且动态环境下自适应能力强. 最后, 本文提出的 PRFC 也可以适用于垃圾短信过滤, 考虑用户兴趣度, 达到个性化过滤目的.

后期工作将考虑在结合个体用户兴趣度的同时, 基于社交网络方法共享其他用户对邮件过滤器过滤结果的反馈意见, 进一步避免误判情况的出现, 当然, 这需要考虑对其他用户邮件内容的隐私保护等问题.

参考文献

1 Messaging Anti-Abuse Working Group. MAAWG email metrics program. First Quarter 2006 Report. 2006. http://www.maawg.org/about/FINAL_1Q2006_Metrics_Report.pdf

2 Teng W L, Teng W C. A personalized spam filtering approach utilizing two separately trained filters. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Washington: IEEE Computer Society, 2008. 125–131

3 Lin H Z, Wang J L, Wu J P, et al. Effect of cold-rolling cladding on microstructure and properties of composite aluminum alloy foil. J Commun, 2017, 34: 121–132 [林海卓, 王继龙, 吴建平, 等. 高校误判拉圾邮件自动召回系统的研究与实现. 通信学报, 2017, 34: 121–132]

4 Huang G W, Liu Y X, Chen Z. Personalized spam filtering method based on users' feedback. Electron Design Eng, 2014, 22: 53–56 [黄国伟, 刘云霞, 陈志. 基于用户反馈的个性化垃圾邮件过滤方法. 电子设计工程, 2014, 22: 53–56]

5 Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering. Expert Syst Appl, 2009, 36: 10206–10222

6 Liu W Y, Wang T. Ensemble learning and active learning based personal spam email filtering. Comput Eng Sci, 2011, 33: 34–41 [刘伍颖, 王挺. 集成学习和主动学习相结合的个性化垃圾邮件过滤. 计算机工程与科学, 2011, 33: 34–41]

7 Clark J, Koprinska I, Poon J. Linger-a smart personal assistant for e-mail classification. In: Proceedings of the 13th International Conference on Artificial Neural Networks (ICANN'03), 2003. 274–277

8 Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail. In: Proceedings of AAAI Workshop on Learning for Text Categorization, 1998. 62: 98–105

9 Graham P. Better Bayesian filtering. 2003. http://www.paulgraham.com/better.html

10 Amayri O, Bouguila N. A study of spam filtering using support vector machines. Artif Intell Rev, 2010, 34: 73–108
11 Sanghani G, Kotecha K. Personalized spam filtering using incremental training of support vector machine. In: Proceedings of Conference on Computing, Analytics and Security Trends (CAST), Pune, 2016. 323–328
12 Yeh C Y, Wu C H, Doong S H. Effective spam classification based on meta-heuristics. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, 2005. 4: 3872–3877
13 Toolan F, Carthy J. Feature selection for spam and phishing detection. In: Proceedings of Conference on eCrime Researchers Summit (eCrime), Dallas, 2010. 1–12
14 Cheng V, Li C H. Personalized spam filtering with semi-supervised classifier ensemble. In: Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence. Washington: IEEE Computer Society, 2006. 195–201
15 Gomes H M, Barddal J P, Enembreck F, et al. A survey on ensemble learning for data stream classification. ACM Comput Surv, 2017, 50: 23
16 Wang S, Minku L L, Yao X. A systematic study of online class imbalance learning with concept drift. IEEE Trans Neural Netw Learning Syst, 2018, 29: 4802–4821
17 Syed N A, Liu H, Sung K K. Handling concept drifts in incremental learning with support vector machines. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1999. 317–321
18 Wang Y W, Liu Y N, Feng L Z, et al. A novel online spam identification method based on user interest degree. J South China Univ Tech (Nat Sci Ed), 2014, 7: 21–27 [王友卫, 刘元宁, 凤丽洲, 等. 基于用户兴趣度的垃圾邮件在线识别新方法. 华南理工大学学报 (自然科学版), 2014, 7: 21–27]
19 Junejo K N, Karim A. Robust personalizable spam filtering via local and global discrimination modeling. Knowl Inf Syst, 2013, 34: 299–334
20 Cohen L, Avrahami-Bakish G, Last M, et al. Real-time data mining of non-stationary data streams from sensor networks. Inf Fusion, 2008, 9: 344–353
21 Gama J, Medas P, Castillo G, et al. Learning with drift detection. In: Proceedings of Conference on Brazilian Symposium on Artificial Intelligence. Berlin: Springer, 2004. 286–295
22 Harel M, Mannor S, El-Yaniv R, et al. Concept drift detection through resampling. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014. 1009–1017
23 Bach S H, Maloof M A. Paired learners for concept drift. In: Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, 2008. 23–32
24 Xu Y, Xu R, Yan W, et al. Concept drift learning with alternating learners. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), Anchorage, 2017. 2104–2111
25 Wang J, Xu S, Duan B, et al. An ensemble classification algorithm based on information entropy for data streams. 2017. ArXiv: 1708.03496
26 Mandelbaum A, Shalev A. Word embeddings and their use in sentence classification tasks. 2016. ArXiv: 1610.08229
27 Sugiyama M, Nakajima S, Kashima H, et al. Direct importance estimation with model selection and its application to covariate shift adaptation. In: Proceedings of Conference on Advances in Neural Information Processing Systems, Vancouver, 2008. 1433–1440
28 Zhang K, Zheng V, Wang Q, et al. Covariate shift in hilbert space: a solution via sorrogate kernels. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, 2013. 388–395
29 Liu A, Ziebart B. Robust classification under sample selection bias. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, 2014. 37–45
30 Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data. In: Proceedings of Conference on Advances in Neural Information Processing Systems, Vancouver, 2007. 601–608
31 Kawahara Y, Sugiyama M. Sequential change-point detection based on direct density-ratio estimation. Statistical Analy Data Min, 2012, 5: 114–127
32 Kanamori T, Hido S, Sugiyama M. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In: Proceedings of Conference on Advances in Neural Information Processing Systems, Vancouver, 2009. 809–816
33 Kivinen J, Smola A J, Williamson R C. Online learning with kernels. IEEE Trans Signal Process, 2004, 52: 2165–2176
34 Junejo K N. Distribution shift resilient discrimination information space for SVM classification. In: Proceedings of the

8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, 2017. 378–383

35 Han Y, He X, Yang M, et al. Chinese spam filter based on relaxed online support vector machine. In: Proceedings of Conference on Asian Language Processing (IALP), Harbin, 2010. 185–188

36 Sun G, Li S, Chen T, et al. Active learning method for Chinese spam filtering. Int J Performability Eng, 2017, 17: 511

A personalized mail re-filtering system based on the client

Dandan XU & Songcan CHEN*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106*, China*

* Corresponding author. E-mail: s.chen@nuaa.edu.cn

Abstract Email is an essential communication tool, but a large number of spam emails can seriously affect the work and life of users and can even cause property damage. Due to different interests and hobbies, there may be huge differences in the definition of spam by users; the realization of personalized spam filtering has become an important issue in the field of spam filtering. When emails are misjudged, the user has to manually modify it, which brings great inconvenience to the user experience. In order to effectively solve the above problems and realize the functions of personalized email filtering and automatic correction of mis-filtered emails, this paper combined with rules and statistical methods presents a personalized email re-filtering system based on the client (PRFC) and implements the automatic modification of the mis-filtered emails. A large part of existing spam filters do not consider the difference between class prior probability and class imbalance problem; they only filter the mail online. Firstly, the proposed filter system processes the mails entering the inbox and the garbage and then designs two mutually learned filters based on the multi-task learning principle to be used for the automatic modification of the mis-filtered emails in inbox and garbage. To ensure the performance of the filter based on the interests of users and data distribution of mails varying with time, a multi-window learning framework that combines important weights to effectively implement the dynamic adaptation of the filter was designed. Finally, our proposed filtering system on the TREC 2006c and 2007p data sets that gets a significant filtering efficiency was verified.

Keywords spam filtering, personalized mail re-filtering, automatic correction, importance weights, multi-task learning, multi-window learning framework

Dandan XU was born in 1993. She obtained her BSc degree in Computer Science and Technology from Jinling Institute of Technology. Currently, she is a master's candidate at Nanjing University of Aeronautics and Astronautics. Her main research interests include machine learning and pattern recognition.

Songcan CHEN was born in 1962. He is a professor and Ph.D. supervisor of pattern recognition and artificial intelligence. He is a senior member of CCF. His main research interests include pattern recognition, machine learning, and neural computing.