基于主动学习和否定选择的垃圾邮件分类算法

胡小娟[1],刘 磊[1],邱宁佳[2]
(1. 吉林大学计算机科学与技术学院,吉林长春 130012;2. 长春理工大学计算机科学技术学院,吉林长春 130022)

摘 要: 针对现在网络上泛滥的垃圾邮件问题,本文结合主动学习方法和否定选择算法提出了一种二类文本分类方法:主动否定学习算法. 根据用户少量标注建立双向兴趣集,利用否定选择算法的自体异常检测机制改善主动学习中的采样策略,并将双向兴趣集作为检测器,新增样本集作为自体集,对两者进行异常匹配. 本文算法与在线垃圾邮件快速识别方法、增强差异性的半监督协同分类算法、垃圾邮件过滤方法、基于人工高免疫的多层垃圾邮件过滤算法和在线主动多领域学习方法在六个常用邮件语料集上进行了分析比较,结果表明本文算法具有较高的准确率、召回率、分类精度,和较低的用户标注负担. 使用用户个性喜好转换为双向兴趣特征的方式有助于提高算法的分类能力;利用异常检测匹配选取未知类别特征的方式,有效地降低了用户标注负担.

关键词: 文本分类; 垃圾邮件检测; 主动学习; 否定选择; 双向用户兴趣集
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)01-0203-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.01.028

A Novel Spam Categorization Algorithm Based on Active Learning Method and Negative Selection Algorithm

HU Xiao-juan[1], LIU Lei[1], QIU Ning-jia[2]
(1. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, Jilin* 130022, *China*)

Abstract: A two-class text categorization method, active learning negative selection text categorization (ALNSTC) algorithm, based on active learning (AL) method and negative selection (NS) algorithm, is proposed for the problem of spam proliferation. The positive user interest set and the negative user interest set are established according to a small number of labeled samples. And the sampling engine (SE) of AL method is improved by the autologous anomaly detection mechanism of the NS algorithm. The two-way user interest sets are used as detectors, and a new sample set is employed as a self-set. The above two sets are matched with Hamming match rules. The classification process of each sample set is able to update the two user interest sets. The proposed algorithm is carried out with a full-scale test on six common spam corpus, which are selected as experimental material, and analyzed and compared with other five state-of-the-art spam classification methods, which are quick online spam identification (QOSI) method, semi-supervised collaboration classification algorithm with enhanced difference (DSCC), dynamic web spam filtering (WSF2) method, multilevel spam filtering algorithm based on artificial immunity (MSFA-AI), and integrated multi-field learning (MFL) method, in different evaluation metrics, such as precision, recall, ROC curve, categorization running time and the labeled number of spam. The results show that the proposed method has better precision rate, recall rate, classification accuracy, and can reduce the artificial labeled number of spam samples. It is advantageous to enhance the classification capacity of the algorithm that the user preferences are converted into positive and negative user interest sets. In addition, the user labeled number is reduced when unknown category features are obtained by the exception detection mechanism.

收稿日期:2016-10-24;修回日期:2017-05-17;责任编辑:覃怀银
基金项目:吉林省自然科学基金(No. 20150101054JC);吉林省博士后科研资助项目(No. 40301919);吉林省科技发展计划重点科技攻关项目(No. 20150204036GX);中国博士后科学基金(No. 2016M591482)

Key words: text categorization; spam detection; active learning; negative selection; two-way user interest set

1 引言

随着互联网的发展,邮件、微信、QQ 等网络通信设施已成为人们平时交流的必备方式. 而种类繁多的垃圾邮件和信息却时时困扰用户,如何高效检测出这些垃圾信息成为研究热点. 目前垃圾邮件识别的研究现状是:(1)由于专家标注的经济代价太大,且无法对大规模问题进行有效标注,无标记样本数据数量巨大且容易获取[1];(2)现有解决方法中的传统机器学习算法,尤其是有监督学习算法,必须大量标记样本数据,否则泛化性能较低[2];(3)对于垃圾邮件过滤问题,用户的个人喜好对分类结果影响较大[3];(4)在线进行人工样本标注时,专家无法直接选择最佳标注时机. 在这种情况下,主动学习(Active Learning, AL)方法成为解决上述问题的主流技术.

AL 方法主要分为学习引擎(Learning Engine, LE)和采样引擎(Sampling Engine, SE)两部分,LE 部分是在有标记的样本集上循环训练,当达到一定训练精度后输出. SE 部分则是对未标记样本进行选样,提交给专家进行人工标注[4]. Liu 基于在线学习、多领域学习和 AL 方法提出一种新的在线主动多领域学习方法 MFL,降低了垃圾邮件过滤中的人工标记负担和空间存储成本[5]. Benevenuto 针对 YouTube 中的垃圾邮件发送者和接收者信息,利用 AL 方法提取重要度最大的子集对 YouTube 进行检测[6]. Feng 基于 AL 方法提出了一种基于边缘密度的不确定性评估方法,在保证准确率的基础上降低了分类的耗费时间[7].

否定选择(Negative Selection, NS)算法模拟了免疫系统识别自体和非自体细胞的否定选择过程,首先随机产生检测器,删除那些检测到自体的检测器,保留检测到非自体的检测器,进而完成自体与非自体数据的分类[8]. 其缺点是采样不当时会对分类结果产生影响,且各检测器的覆盖空间有交集,会出现整体覆盖率较低的问题;优点是无需先验知识,只需利用有限数量的自体便能检测出无限数量的非自体. Ismaila 利用粒子群优化方法改善 NS 算法中的随机检测器生成机制,提出 NSA-PSO 模型,可应用于 CS 端的垃圾邮件检测[9],另将局部选择差分进化方法和 NS 算法相结合,用于提高垃圾邮件过滤的准确率[10].

综上,目前现存的方法中有基于 AL 方法改善邮件分类的,也有基于 NS 算法改进邮件分类的,但都只是在准确率、或召回率,或耗费时间,或标注负担等单方面上有所改善,本文结合 AL 方法和 NS 算法,结合双向兴趣集,以及关键特征选择方法,致力于解决邮件分类中准确率与召回率的提高和耗费时间与人工标注负担的降低等问题,提出一种新的二类文本分类算法:主动否定学习算法(Active Learning Negative Selection Text Categorization, ALNSTC). 利用基于二项假设的关键特征选择算法[11](Bi-Test)先对样本的原特征集进行筛选,提高所选关键特征的重要程度,减少计算耗费和用户的标注负担,提高用户标注的价值,减少耗时,解决 NS 算法中分类精度不高的问题. 利用根据用户标注的少量样本建立正向和负向兴趣集的方式,从两端向中间覆盖,提高覆盖率,解决检测器覆盖率低的问题. 将学习引擎引入到检测器生成机制中,将学习生成的双向兴趣集作为检测器,使得邮件分类能够符合用户的个性化喜好,同时,也使得检测器生成机制具有自学习性. 由于双向兴趣集和异常检测匹配规则的使用,使得未知类别的关键特征数量减少,进而减少人工标注负担,弥补 AL 方法需要大量人工标注的不足问题.

2 ALNSTC 基本思想

2.1 准备工作

在进行邮件分类前,首先对邮件进行预处理操作. 鉴于邮件本身的隐私性和特殊性,对邮件中的附件、标签、停用词等进行预处理后,进行分词和还原词根处理,再对其进行编码,将邮件文本中的文字转换为数字代码的格式,规避用户隐私泄露.

将经过预处理的邮件文本作为样本,每个样本经分词后形成一组原特征,组成原特征集. 为减少计算负担,本文采用文献[11]中的 Bi-Test 方法对每个原特征集进行关键特征筛选,用其代替样本进行分类操作,有效降低了特征空间的维度.

2.2 建立用户兴趣集

将已标注的少量邮件作为本文算法的初始训练集 S_0,其中既包含合法邮件,也包含垃圾邮件. 对 S_0 经过一系列预处理后变成 S_0',对 S_0 进行关键特征选择,得关键特征集 FS_0,$FS_0=\{FS_{01},FS_{02},\cdots,FS_{0K}\}$,$FS_{0i}$ 代表某一邮件的关键特征集. 合法邮件的关键特征集组成用户正向兴趣集 P,垃圾邮件的关键特征集组成用户负向兴趣集 N,且 $P\cap N=\varnothing$. 双向用户兴趣集的建立过程如图 1 所示. 通过 S_0 创建用户兴趣集 P 和 N 的详细算法如算法 1 所示.

算法 1 建立双向用户兴趣集

输入:原始特征集 S_0'

输出:正向用户兴趣集 P 和负向用户兴趣集 N

初始化集合 $FS_0 = P = N = \varnothing$
对 S_0' 进行分类
计算 S_0' 的关键特征集合 FS_0 = Bi-Test(S_0')
For each FS_{0j} in FS_0
  If FS_{0j}所属邮件是合法邮件
    $P = P \cup FS_{0j}$
  Else If FS_{0j}所属邮件是垃圾邮件
    $N = N \cup FS_{0j}$
  End
End
If $P \cap N \neq \varnothing$
    $P = P - P \cap N$
    $N = N - P \cap N$
返回 P 和 N

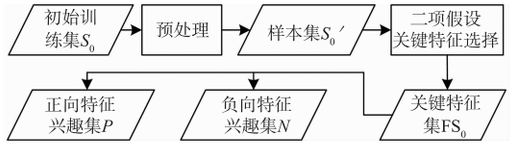


图1　生成用户兴趣集

算法 1 的计算复杂度为:$|FS_0| \times \text{Max}(|FS_{0j}|)$,其中$|FS_0|$为初始样本个数,$FS_{0j}$为关键特征个数最多的样本,且经由 Bi-Test 方法知 $\text{Max}(|FS_{0j}|) \leqslant 5$.

2.3　主动否定学习算法

根据邮件在线处理和实时处理的需求,将两个兴趣集作为检测器,新增样本集作为自体集,对两者进行异常检测匹配. 设 $FNew_i$ 为新增样本集 New_i 经过 Bi-Test 方法筛选后的关键特征集. 选用文献[12]中基于海明距离的相似度评估方法:

$$\text{Hamming similarity} = \sum_{i=1}^{M} \overline{(A_i \oplus B_i)}, A, B \in \{0,1\}^M \tag{1}$$

作为本文异常检测的匹配规则. 式(1)中 A、B 分别代表有限长度的符号串,M 代表检测器中符号串的总数. 将图 1 中获得的正向兴趣特征集 P 和负向兴趣特征集 N 作为检测器,每次从 $FNew_i$ 中抽取一个特征 f_j,根据设定的匹配规则与检测器进行匹配,当结果为“匹配”时代表该特征存在于 P 或 N 中,即可判定该特征所属样本类别为合法邮件或垃圾邮件,将此特征保存在 P_i 或 N_i 中. 当结果为“不匹配”时,表明该特征不存在于 P 或 N 中,将其放入未知类别特征集 NP_i 或 NN_i 中,等待下一步的不确定性鉴定.

设 XN_i 为最具有标注价值的关键特征集,$XN_i = NP_i \cap NN_i$,将 XN_i 推荐给用户进行标注. 由于用户是对邮件进行标注,所以需要将 XN_i 还原为邮件,用户标注完成后再将邮件还原为 XN_i. 将用户标注后的 XN_i 划分为正向兴趣子集 XNP_i 和负向兴趣子集 XNN_i,若 $XNP_i \cap XNN_i \neq \varnothing$,则对 XNP_i 作如下处理:$XNP_i = XNP_i - XNP_i \cap XNN_i$. 利用兴趣子集 XNP_i 和 XNN_i 更新用户兴趣集 P 和 N,根据分类后的 $P_i \cup XNP_i$ 和 $N_i \cup XNN_i$ 对样本集中的样本类别进行自动标注,进而对邮件进行自动标注,然后等待新邮件集 New_{i+1} 的到来,具体分类匹配流程如图 2 所示.

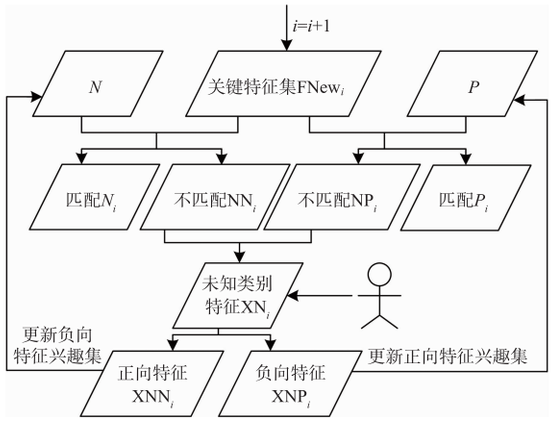


图2　新样本集的分类匹配过程

未知类别关键特征集 XN_i 经用户标注后分成新的兴趣子集 XNP_i 和 XNN_i,由于用户的动态需求,用户个性喜好会有所变化,在将新的兴趣子集并入双向用户兴趣集前,要先进行过期兴趣特征的淘汰. 首先,检测 $XNN_i \cap P$ 是否为空集,若不为空,交集中的特征即为需要淘汰的过期兴趣特征,将此交集从 P 中删除;其次,检测 $XNP_i \cap N$ 是否为空集,若不为空,将此交集从 N 中移除;最后,将 XNP_i 并入 P,将 XNN_i 并入 N,就可以得到更新后的用户兴趣集 P 和 N,其中 $P = P \cup XNP_i$,$N = N \cup XNN_i$.

新增样本集 New_i 中合法邮件的关键特征集 P_i 和垃圾邮件的关键特征集 N_i 是与已存兴趣集 P 和 N 匹配后所得,P_i 和 N_i 与新增样本集的关键特征集 $FNew_i$ 中未知类别的关键特征集 XN_i 之间的关系可证明如下. 已知 $FNew_i = P_i \cup NP_i = N_i \cup NN_i$,$XN_i = NN_i \cap NP_i$,则可推得:

$$\begin{aligned} &P_i \cup N_i \cup XN_i \\ &= P_i \cup N_i \cup (NP_i \cap NN_i) \\ &= P_i \cup [(N_i \cup NP_i) \cap (N_i \cup NN_i)] \\ &= P_i \cup (N_i \cup NP_i) \\ &= FNew_i \end{aligned}$$

又知 $XN_i = XNN_i \cup XNP_i$,因此亦可得 $FNew_i = P_i \cup N_i \cup XNP_i \cup XNN_i$,可推算出 $P_i \cup XNP_i$ 所对应的邮件样本集为 New_i 中的合法邮件集 H_i,以及 $N_i \cup XNN_i$ 所对应的邮件样本集为 New_i 中的垃圾邮件集 S_i.

由此可设计出针对新增邮件集 New_i 的分类算法如算法 2 所示.

算法 2 基于 AL 算法和 NS 算法的垃圾邮件分类(ALNSTC)

输入:增量样本集 New_i,正向用户兴趣集 P,负向用户兴趣集 N
输出:合法邮件集 H_i,垃圾邮件集 S_i,正向兴趣集 P,负向兴趣集 N
1. 初始化集合 $FNew_i = \varnothing$
2. $FNew_i$ = Bi-Test(New_i)
3. P_i = MatchRules($FNew_i$, P)
4. For each $f_j \in P_i$
5. 自动标记 f_j 所属邮件为合法邮件,并将该邮件放入 H_i 中
6. End
7. $NP_i = FNew_i - P_i$
8. N_i = MatchRules($FNew_i$, N)
9. For each $f_j \in N_i$
10. 自动标记 f_j 所属邮件为垃圾邮件,并将该邮件放入 S_i 中
11. End
12. $NN_i = FNew_i - N_i$
13. $XN_i = NP_i \cap NN_i$
14. 将关键特征集 XN_i 还原为邮件集合 $XFNew_i$
15. 对 $XFNew_i$ 进行人工垃圾邮件标注,将标注后的垃圾邮件的关键特征放入 XNN_i 中
16. $XNP_i = XN_i - XNN_i$
17. For each $f_j \in XNN_i$
18. 将 f_j 所属邮件放入 S_i 中
19. If $f_j \in P$
20. 将 P 中的 f_j 删除
21. End
22. End
23. For each $f_j \in XNP_i$
24. 将 f_j 所属邮件放入 H_i 中
25. If $f_j \in N$
26. 将 N 中的 f_j 删除
27. End
28. End
29. $P = P \cup XNP_i$
30. $N = N \cup XNN_i$
31. 返回 H_i, S_i, P, N

ALNSTC 算法特征选择的计算复杂度为 $O(|New_i|)$,$|New_i|$ 为新增样本集中的特征数量;分类匹配的计算复杂度为 $O((|P| + |N|) \times |FNew_i|)$,其中 $|P|$、$|N|$ 和 $|FNew_i|$ 分别表示集合 P、N 和 $FNew_i$ 中的特征总数,且因 $FNew_i$ 是关键特征集,$|New_i| >> |FNew_i|$,相较于计算复杂度为 $O(|S| \times \log(|S|)) + O(|S|)$ 的传统特征选择($|S|$ 为样本集的原特征数量[13]),本文分类算法计算复杂度 $O(|New_i|)$ 能有效减少 CPU 处理时间.

3 实验结果及分析

3.1 数据集

本文选择 PU1、PU3、PUA、PU2、Lingspam 和 Spambase 六个常用邮件语料数据集作为实验素材,其中前三个数据集中两种类别的邮件数量相当,后三者中邮件数量分布不均衡,具体如表 1 所示.

表 1 垃圾邮件语料数据集

Dataset	Ham (%)	Spam (%)	SUM
PU1	618(56.2%)	481(43.8%)	1099
PU3	2111(51.0%)	2028(49.0%)	4139
PUA	571(50.0%)	571(50.0%)	1142
PU2	577(80.0%)	144(20.0%)	721
Lingspam	2412(83.4%)	481(16.6%)	2893
Spambase	2788(60.6%)	1813(39.4%)	4601
SUM	9077	5518	14595

3.2 评价标准

为了更好的评价 ALNSTC 算法的分类性能,采用准确率(precision),召回率(recall),ROC(Receiver Operating Characteristic)曲线,算法分类平均耗费时间,用户标注负担作为评价标准. 本文利用非参数法,通过不同的阈值变化,得到对应分类算法的敏感度(sensitivity)和特异性值(specificity),将(1-特异性)值和敏感度值分别设定为横坐标和纵坐标,用平滑曲线连接各点得到 ROC 曲线. 选取 ROC 曲线中的曲线下面积(Area Under the Curve,AUC)作为 ALNSTC 算法的评估指标.

3.3 准确率和召回率分析

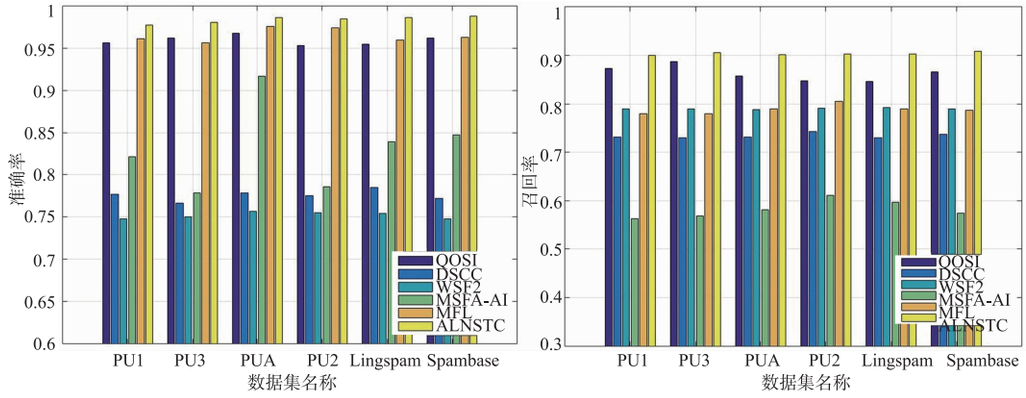
选择 QOSI 方法[3]、DSCC 算法[14]、WSF2 方法[15]、MSFA-AI 算法[16]、MFL 方法[5] 作为 ALNSTC 算法的参照算法. 将表 1 中的六个数据集进行预处理后平均分成十份,采用十折交叉验证方法分别进行实验,所得实验结果的准确率和召回率如图 3 所示.

从图 3 可看出,ALNSTC 算法的准确率和召回率均高于其他参照算法. MFL 方法是先利用信息文档的结构特性进行分域,后将多个域的结果组合,以提高分类性能;QOSI 方法则是利用样本分类确定性评价函数和样本价值评价函数提高分类能力;ALNSTC 算法是将合法邮件的关键特征和垃圾邮件的关键特征转换为正向兴趣集 P 和负向兴趣集 N,代表用户个性喜好,使得新增样本集 $FNew_i$ 与兴趣集进行匹配时获得的特征相似度更加准确. 且每次处理 $FNew_i$ 都会更新双向用户兴趣集,使得双向兴趣集更有时效性,更贴合用户当时的个性需求,因此每次匹配准确率更高. 而利用特征相似度评估方法对 $FNew_i$ 中任意特征 f_j 进行评估的方法,大大降低了将垃圾邮件错认为合法邮件的概率,尤其在高特征相似度和高准确率的保证下,因而召回率能够有效提高.

3.4 AUC 分析

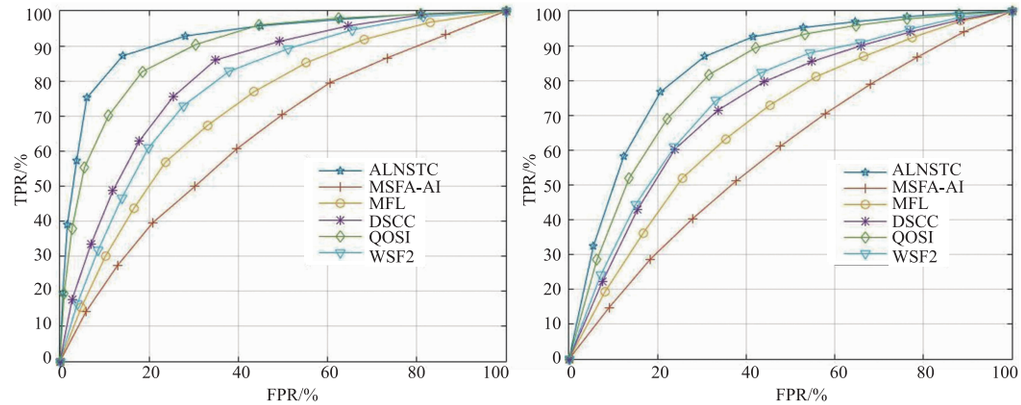
由于 ALNSTC 算法可将垃圾邮件和合法邮件中的

关键特征按照特征相似度做两端化处理,并能够依据正、负向用户兴趣集,对关键特征集进行分类,选取相似度最靠近中间部分的关键特征推荐给用户进行标注. 为了更准确的分析各算法的实际运行情况,选取具有代表性的数据集 PU3 和 Lingspam 作为实验数据集. 两个数据集中样本数量较大,PU3 正负数据量相当,Lingspam 正负样本分布最为不平衡. 阈值选择方法选择取百分比法,增量为 10%. 分别比较六个算法在两个数据集上的 10 次运行情况,取横坐标 1-Specificity = FPR,纵坐标 Sensitivity = TPR,制作成 ROC 曲线如图 4 所示.



（a）准确率比较　　（b）召回率比较

图3　各算法的分类结果比较



（a）准确集PU3　　（b）准确集Lingspan

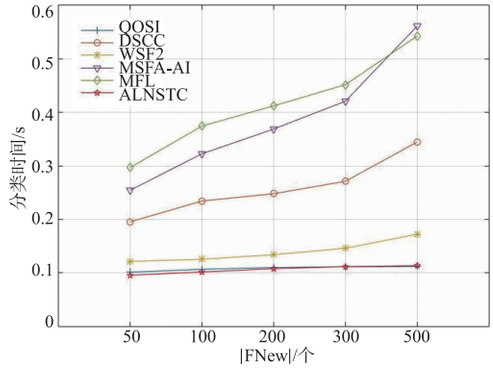
图4　各算法在不同数据集上的ROC曲线

从图 4 中可看出,与各算法在 PU3 上的 AUC 相比较,在 Lingspam 上的 AUC 都有所减少,这主要因为 Lingspam 数据分布不平衡的原因. ALNSTC 算法的 AUC 都略高于其他参照方法的 AUC,表明 ALNSTC 算法在两个数据集上都具有较高的分类精度. 相较于其他垃圾邮件检测中只针对垃圾邮件中的关键特征进行分析,本文算法能够同时对合法邮件和垃圾邮件的关键特征进行相似度评估,且分正向和负向两方面进行匹配,因此在进行分类时上有双重保障,因而有较高的分类精度.

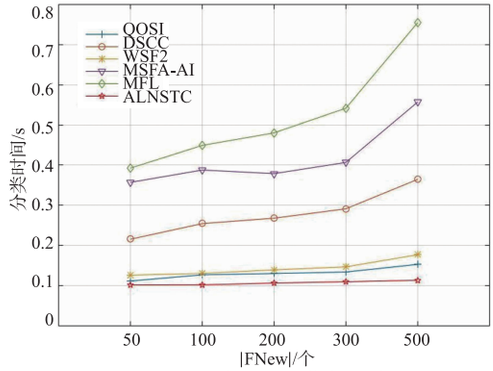
3.5　分类耗时分析

新增样本集 New_i 的数量分别取 50、200、500,截取关键特征集 $FNew_i$ 中的特征数量为 50、100、200、300、500 时的各算法所用时间. 经十折交叉验证方法后的平均分类时间如图 5 所示,其中 $|FNew|$ 表示关键特征数量.

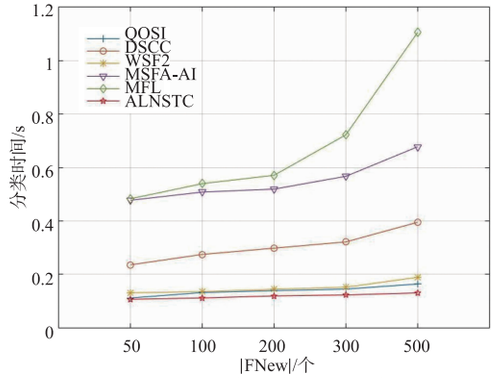
可以看出,随着 New_i 的不断增大,所需的分类耗费时间都有所增加. 当 $|New_i| = 50$ 时,ALNSTC 算法与 QOSI 方法的平均分类时间相差无几,而其他参照算法所用时间略长(如图 5(*a*)所示). 随着新 New_i 的不断出现,双向用户兴趣集愈加庞大,用户兴趣集的变化相对减小,且由于 $|FNew_i| << |New_i|$,因此所用分类时间变化不大,较其他参照算法的耗时有明显优势(如图 5(*b*)和图 5(*c*)所示). 由此可见,由于 ALNSTC 算法的低计算复杂度,因而能与同类分类算法在分类耗时上相持平,甚至低于同类算法的耗时. MFL 方法选择特定属性域文档分割策略,因为需要对每个域中的文档进行统计、计算和表示,每个域分类器的时间开销较大. MS-



(a) 当|New_i|取值为50时



(b) 当|New_i|取值为200时



(c) 当|New_i|取值为500时

图5 各算法在不同|New_i|下的分类时间

FA-AI 算法中因激活阈值设为 0.6,四个检测器间采用“AND”组合关系,为尽量提高算法准确率,因而时间耗费是六种方法中最大的.

3.6 用户标注负担分析

为了进行标准评估,设定固定准确率值,记录各个算法在这一准确率下需要人工标注的样本数量. 新增样本集 New_i 中的样本数量分别取值为 50、200、500. 各算法在六个数据集上采用十折交叉验证方法,取六十个结果中用户标注样本数量的平均值,如表 2 所示. ALNSTC 算法实际标注的样本数量均少于其他方法. 算法中根据异常检测机制,所有不匹配的情况组成了未知类别特征集 XN_i,而前期关键特征选择使得特征维度骤降,因而 $|XN_i|$ 很小,需要用户标注的邮件也相应很少,因此 ALNSTC 算法能够有效减少用户标注负担. DSCC 算法在此环节表现最差是因为本文实验中选取文献[14]中的样本标注 20% 作为实验条件.

表 2 各算法在不同 $|New_i|$ 下的用户标注数量

$\|New_i\|$	QOSI	DSCC	WSF2	MSFA-AI	MFL	ALNSTC
50	5.8	9.9	7.1	7.8	9.6	3.7
200	22.2	40.3	26.5	32.7	37.8	11.4
500	57.4	95.9	67.1	79	93.4	36.5

4 总结

本文针对二类文本分类问题,提出了一种新的主动否定学习算法,用于解决在线垃圾邮件分类. 采用准确率、召回率、ROC 曲线、耗费时间和用户标注数目作为 ALNSTC 算法的评价标准,与其他同类方法进行了实验分析,主要贡献总结如下:

(1)将用户个性喜好转换成正、负向用户兴趣集,对新增样本集中的关键特征分别与正、负向兴趣集中的关键特征进行相似度评估,通过评估确定特征的类别.

(2)将双向用户兴趣集作为检测器,新增样本的关键特征集作为自体集,通过 NS 算法中的异常检测机制,对两者进行异常检测匹配,结果为匹配时,算法自动对特征进行分类,结果为不匹配时,算法收集为未知类别特征,推荐给用户进行标注.

(3)通过在六个常用语料集上与其他五种参照算法进行比较,分析可得本文算法具有较高的分类精度以及较低的用户标注负担.

本文算法在常用语料集上能够表现出的良好分类性能,下一步将使用 ALNSTC 算法在海量垃圾邮件语料集上进行进一步的实验分析,验证其在各类数据集上的耗时、用户标注和分类能力.

参考文献

[1] 郭虎升,王文剑. 基于主动学习的模式类别挖掘模型[J]. 计算机研究与发展,2014,51(10):2148-2159.
Guo Hu-sheng, Wang Wen-jian. A pattern class mining model based on active learning[J]. Journal of Computer Research and Development, 2014, 51(10): 2148-2159. (in Chinese)

[2] Balcan M F, Blum A. A discriminative model for semi-supervised learning[J]. Journal of the ACM, 2010, 57(3): 1-46.

[3] 王友卫,刘元宁,凤丽洲,朱晓冬. 基于用户兴趣集的在线垃圾邮件快速识别新方法[J]. 电子学报,2015,43

(10):1963 – 1970.
Wang You-wei, Liu Yuan-ning Feng Li-zhou, Zhu Xiao-dong. A novel quick online spam identification method based on user interest set[J]. Acta Electronica Sinica, 2015,43(10):1963 – 1970. (in Chinese)

[4] 吴伟宁,刘扬,郭茂祖,刘晓燕. 基于采样策略的主动学习算法研究进展[J]. 计算机研究与发展,2012,49(6):1162 – 1173.
Wu Wei-ning, Liu Yang, Guo Mao-zu, Liu Xiao-yan. Advances in active learning algorithm based on sampling strategy[J]. Journal of Computer Research and Development, 2012,49(6):1162 – 1173. (in Chinese)

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

[6] Benevenuto F, Rodrigues T, Veloso A, Almeida J, Goncalves M, Almeida V. Practical detection of spammers and content promoters in online video sharing systems[J]. IEEE Transactions on Systems Man & Cybernetics-Part B (Cybernetics),2012,42(3):688 – 701.

[7] Feng L Z, Wang Y W, Zuo W L. Quick online spam classification method based on active and incremental learning [J]. Journal of Intelligent & Fuzzy Systems,2016,30(1): 17 – 27.

[8] 金章赞,廖明宏,肖刚. 否定选择算法综述[J]. 通信学报,2013,34(1):159 – 170.
Jin Zhang-zan, Liao Ming-hong, Xiao Gang. Survey of negative selection algorithms[J]. Journal on Communications, 2013,34(1):159 – 170. (in Chinese)

[9] Idris I, Selamat A, Nguyen N T, Omatu S, Krejcar O, Kuca K, Penhaker M. A combined negative selection algorithm-particle swarm optimization for an email spam detection system[J]. Engineering Applications of Artificial Intelligence,2015,39:33 – 44.

[10] Idris I, Selamat A, Omatu S. Hybrid email spam detection model with negative selection algorithm and differential evolution[J]. Engineering Applications of Artificial Intelligence,2014,28:97 – 110.

[11] Yang J M, Liu Y N, Liu Z, Zhu X D, Zhang X X. A new feature selection algorithm based on binomial hypothesis testing for spam filtering[J]. Knowledge-Based Systems, 2011,24(6):904 – 914.

[12] Harmer P K, Williams P D, Gunsch G H, Lamont G B. Artificial immune system architecture for computer security applications[J]. IEEE Transactions on Evolutionary Computation,2002,6(3):252 – 280.

[13] Yang J M, Liu Y N, Zhu X D, Liu Z, Zhang X X. 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.

[14] 于重重,商利利,谭励,涂序彦,杨扬,王竞燕. 一种增强差异性的半监督协同分类算法[J]. 电子学报,2013,41(1):35 – 41.
Yu Chong-chong, Shang Li-li, Tan Li, Tu Xu-yan, Yang Yang, Wang Jing-yan. A semi-supervised collaboration classification algorithm with enhanced difference[J]. Acta Electronica Sinica,2013,41(1):35 – 41. (in Chinese)

[15] Fdez-Glez J, Ruano-Ordas D, Mendez J R, Fdez-Riverola F, Laza R, Pavon R. A dynamic model for integrating simple web spam classification techniques[J]. Expert Systems with Applications,2015,42(21):7969 – 7978.

[16] 张泽明,罗文坚,王煦法. 一种基于人工免疫的多层垃圾邮件过滤算法[J]. 电子学报,2006,34(9):1616 – 1620.
Zhang Ze-ming, Luo Wen-jian, Wang Xu-fa. A multilevel spam filtering algorithm based on artificial immunity[J]. Acta Electronica Sinica, 2006, 34(9): 1616 – 1620. (in Chinese)

作者简介

胡小娟　女,1985 年生于山东淄博,现为吉林大学计算机科学与技术学院博士研究生. 主要研究方向为数据挖掘和机器学习.
E-mail:shdhuxiaojuan@163.com

刘　磊　男,1960 年生于吉林长春. 现为吉林大学教授、博士生导师. 主要研究方向为软件形式化、数据挖掘和机器学习.
E-mail:liulei@jlu.edu.cn

邱宁佳(通讯作者)　男,1984 年生于河南南阳. 现为长春理工大学讲师. 主要研究方向为数据挖掘和机器学习.
E-mail:qiunj@cust.edu.cn