文章编号:1007-5321(2019)06-0111-07 **DOI**:10.13190/j.jbupt.2019-147

混合采样与遗传算法相结合的垃圾网页检测

刘 寒

(1.北京邮电大学 软件学院, 北京 100876; 2.北京邮电大学 可信分布式计算与服务教育部重点实验室, 北京 100876)

摘要: 垃圾网页检测存在数据不平衡、特征空间维度较高的问题,为此,提出一种基于随机混合采样和遗传算法的集成分类算法. 首先,使用随机混合采样技术,通过随机抽样,减少多数类样本数量,用少数类样本合成过采样技术方法生成少数类样本,获得多个平衡的训练数据子集;然后使用改进的遗传算法对训练数据集进行降维,得到多个具有最优特征的训练数据子集;使用极端梯度算法(XGBoost)作为分类器,训练多个平衡数据子集,用简单投票法对多个分类器进行集成,得到新的分类器;最后对测试集进行预测,得到最终预测结果. 实验结果表明,提出算法的分类结果与 XGBoost 的结果相比,准确率提高了约 19.25%,且减少了建立学习模型的时间,提高了分类性能,是一种较好的分类算法.

关 键 词: 垃圾网页检测; 混合采样; 集成分类; 遗传算法; 极端梯度算法

中图分类号: TP181 **文献标志码**: A

Spam Web Detection Based on Hybrid-Sampling and Genetic Algorithm

LIU Han

(1. School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Key Laboratory of Trustworthy Distributed Computing and Service (Beijing University of Posts and Telecommunications), Ministry of Education, Beijing 100876, China)

Abstract: Spam web detection is often troubled by the problem of unbalanced data and high feature space dimension. In order to solve these two problems, the ensemble classification algorithm based on random hybrid-sampling and genetic algorithm was proposed. Firstly, a number of balanced training data subsets is obtained by reducing the number of majority samples through random sampling and generating minority samples by synthetic minority over-sampling technique (SMOTE) method. Then, the improved genetic algorithm is used to reduce the dimension of training data set to obtain multiple subsets of training data with optimal feature. Extreme gradient boosting (XGBoost) is also used as the classifier to train multiple balanced data subsets, and so a new classifier is obtained by ensemble multiple classifiers with simple voting method. Finally, the test set is predicted and the final prediction is obtained. Experiments show that, compared with XGBoost, the proposed algorithm improves the accuracy by about 19.25%, reduces the time to build the learning model, and improves the classification performance.

Key words: spam web detection; hybrid-sampling; ensemble classification; genetic algorithm; extreme gradient boosting

收稿日期: 2019-11-22

基金项目: 国家重点研发计划项目(2017YFC1307705)

作者简介: 刘 寒(1997—), 女, 硕士生, E-mail: liu_han@bupt.edu.cn.

随着互联网的快速成长,搜索引擎也逐渐发展起来,如今人人都需要使用搜索网站来找到自己需要的信息. 但是有研究发现, 85% 的用户只会去查看第一页的内容,甚至只是查看搜索引擎中排名前十的网页[1-2]. 可以看出,排名靠前的网站就会拥有更大的访问量. 因此有些网站的所有者就会试图以不正当的手段和技术欺骗搜索引擎,从而提高网站排名,使他们的网站排名高于他们应得的,这些网页被称为垃圾网页.

目前垃圾网页分为链接垃圾网页、内容垃圾网页[3]和伪装垃圾网页3种. 垃圾网页降低了搜索引擎的搜索结果质量,也浪费了用户的时间,因此,有很多人研究检测垃圾网页的方法[4]. 目前检测垃圾网页主要有2种方法:第1种是基于web图,根据页面有效性估计网页信任程度,从而识别出垃圾网页[5-6];第2种方法是基于网页内容和组件,属于分类和监督学习问题[7-9],一方面人们强调从网页中提取特征[10-11];另一方面,用一些性能较优的分类算法进行垃圾网页检测,属于数据不平衡分类中的典型问题. Singh T 等[12]提出了一个基于模糊逻辑的框架来检测 web 垃圾页面. Fdez-Glez J 等[13]提出了一个使用增量学习检测 web 垃圾网页的框架. 此外,神经网络、深度学习、支持向量机等方法也被用来检测垃圾网页[14-15].

在实际应用中,垃圾网页检测通常都会遇到类似数据量大、数据含不平衡类、高计算成本和内存消耗等问题. 本研究将从不平衡数据和特征选择两方面入手研究:一方面通过重采样技术获得平衡数据集;另一方面通过降低特征空间的维度提高分类器的准确率.

1 相关工作

1.1 不平衡数据

在不平衡数据集中,每一类样本数据的分布都是不平衡的. 在垃圾网页分类中,样本较多的一类是正常网页,较少的一类是垃圾网页. 通常人们更关心的是样本少的那一类,例如垃圾网页检测. 如果在分类过程中使用典型常用的分类器,分类结果往往会更加倾向于样本数多的一类,即少数类样本不容易被识别出来. 因此要改变这种情况,就要设法使数据集样本均衡,在这基础之上对其分类. 也有人根据其数据集的特征专门设计出适用于不平衡数据集的分类算法.

1.2 重采样

重采样指的是可以得到平衡数据集,使得样本分布均匀的一种技术,它可以用来减缓在模型训练过程中样本的不平衡问题,从而提高分类性能,且重采样是独立于所选分类器的,更加通用. 重采样技术主要分为过采样技术和欠采样技术[16]. 过采样技术指的是通过在数据集中增添少数类样本的方式来平衡数据集,其中最典型的方法就是简单地复制少数类样本. 也有人提出少数类样本合成过采样技术(SMOTE, synthetic minority over-sampling technique)[17],它创造出了新的少数类样本,从而扩充了少数类样本的数量,这种方法如今也被很多人运用. 欠采样指的是通过删掉部分多数类样本来达到平衡数据集的目的. 其中最早被提出来的就是随机欠采样技术(RUS, random under-sampling),在很多用来消除多数类样本的例子中都应用到了 RUS. 混合方法是过抽样技术与欠采样的结合.

1.3 特征选择和提取

在数据集不平衡的情景下,少数类样本很容易作为噪音被丢弃,但是如果删除了样本空间中不相关的特征,就会降低这一风险. 通常,特征选择的目标是从整个特征空间中选择 k 个特征的子集,优化分类器的结果.

处理特征的第2种方法就是特征提取,将数据转为更低维的空间. 特征选择与特征提取的区别在于特征提取是从原始特征使用功能映射创建新特征,特征选择是返回原始特征的子集.

1.4 分类算法

人们尝试构建一种分类算法,使其适用于不平衡数据集,比传统的分类器效果更好. 通常有2个方向,分别是集成方法和对算法分类器的改进. 本研究使用的是集成方法.

集成分类器就是一个多分类器系统,它将性能较优的几个基分类器进行组合,从而提高其分类效果[18],现在已成为解决类不平衡问题的常用方法. 现在常用的方法有袋装算法(Bagging)、提升算法(Boosting)和混合集成(Bagging 和 Boosting 混合). 集成一般分为两类,分别为基于迭代的集成和基于并行的集成.

1) 基于迭代的集成. Boosting 是集成学习中最常见且有效的方法. 第一个被提出使用的 Boosting 算法就是自适应增强算法(Adaboost). Adaboost 在使用时能够将更高的权重赋予给被分配到正确类的

样本,使得之后的分类器更多关注于学习分类失败的这些样本. 其他典型的迭代集成方法包括梯度增强决策树,基于进化算法的集成算法.

2）基于并行的集成. 基于并行的集成能够并行训练每个基分类器,主要有基于 Bagging 的集成、基于重采样的集成和基于特征选择的集成. 其中,Bagging 算法代表就是随机森林算法. 并行集成具有节省时间和易于开发的优点,因此常被用来解决实际问题.

在实现迭代或是并行集成模型时,需要一个基分类器,它可以是支持向量机和决策树等任何经典模型. 每种分类器都有自己的优缺点,要根据实际需求进行选择.

1.5　评价指标

对不平衡数据的处理在数据挖掘中一直是有难度的问题,尤其是以垃圾网页检测为首的二元分类问题. 由于样本分布不均,分类过程中很容易将样本归为多数类,但是目标样本往往属于少数类. 因此选择能够体现出分类效果好坏的评价指标尤为重要.

为了对数据集分类效果的好坏有一致的评价标准,本研究采用 3 个评价指标,分别是准确率、F_1 测度和 A_{UC} 值. 由于垃圾网页检测属于二元分类,其中,垃圾网页属于目标类. 对于二元分类来说,样本分为正例与负例,构成混淆矩阵如表 1 所示. 其中 T_P 为被正确分类的正例数,F_P 为被错分为正例的负例数,F_N 为被错分为负例的正例数,T_N 为被正确分类的负例数.

表 1　混淆矩阵

预测样本	真正的正例	真正的负例
正例	T_P	F_P
负例	F_N	T_N

F_1 测度值全面反映了整体的分类性能,而且对不平衡数据集的分类也有效果,其计算如式(1)所示. 准确率即为被正确分类的样本数占所有样本数的比值,其计算如式(2)所示. 人们经常用 A_{UC} 值来评价二元分类模型训练效果的好坏[19],A_{UC} 评估的是分类器避免错误分类的能力,其计算如式(3)所示,A_{UC} 值越大,分类效果越好.

$$F_1 = \frac{2T_P}{2T_P + F_N + F_P} \tag{1}$$

$$A_{\text{ccuracy}} = \frac{T_P + T_N}{T_P + F_P + T_N + T_N} \tag{2}$$

$$A_{UC} = \frac{1}{2}\left(\frac{T_P}{T_P + F_N} + \frac{T_N}{T_N + F_P}\right) \tag{3}$$

2　算法分析

2.1　随机混合采样

提出一种随机的欠采样与过采样结合的混合方法(RUOS, random under-over-sampling)对样本进行采样,得到一个平衡数据集,具体算法如下.

输入数据集,包含正常网页样本集 N 和垃圾网页样本集 S;

输出多个平衡样本子集 $D_i(i = 1,2,\cdots,k)$.

1）计算:

s = 少数类样本个数

n = 多数类样本个数

z = 设定的 n 与 s 的比值

$$k = \text{round}\left(\frac{n}{zs}\right)$$

其中 round 为四舍五入操作.

2）将 N 随机平均地分为 k 个样本子集 $N_1', N_2', \cdots, N_k'$,其中 N_i' 的样本个数约等于 z 倍的 s;

3）分别将 N_i' 与 S 组合为一个新的不平衡数据集,且 N_i' 的样本数与 S 的样本数比值为 $z:1$;

4）分别将 N_i' 与 S 组成的不平衡样本集通过 SMOTE 方法生成一个新的平衡数据集 D_i',打乱数据集内样本顺序;

5）返回 D_i'.

基于上述的 RUOS 方法得到了 k 个平衡数据集,然后就可以训练出多个分类器,使用简单投票法对多个分类器进行集成,得到一个新的集成分类器.

2.2　基于改进的遗传算法的特征选择

在遗传算法(GA,genetic algorithm)[20]中,子代为最优特征子集,遗传产生子代的过程就是搜索最优特征子集的过程. 每个个体由二进制位串表示,“1”表示选中该特征,“0”则未选中该特征,位串长度即为特征的总数. 首先对种群进行初始化,种群包含若干个体,若没有达到最大迭代次数就进入迭代. 在迭代过程中,先计算出种群中个体的适应值,根据适应值的大小来判断是否进入交配池形成父代子群. 再根据交叉、变异概率进行交叉、变异操作,使种群保证其多样性. 最终产生优异的子代.

具体算法如下.

1) 种群初始化:随机生成一个种群,这个种群中含 m 个个体. 产生方式如下.

$$R = \text{round}(\text{rand}(m,g))$$

其中:R 为种群,m 为个体,g 为特征的个数,$\text{rand}(m,g)$ 返回一个 m 行 g 列的矩阵,其中每个元素取值的范围都在[0,1]区间中,使得得到的矩阵所有元素为0或1,从而返回 m 个个体,每个个体由 g 位的二进制位串组成.

2) 个体评价:遗传的过程就是迭代产生最优特征子集的过程,在本研究中使用 F_1 值来衡量分类器的效果,也代表种群中个体的适应值. 其中,封装的分类器为基于混合采样的 XGBoost 算法.

个体适应值算法如下.

输入:训练集 B,二进制位串 Q,整数 n 代表 n 折交叉验证.

输出:F_1 值.

① 根据 Q 表示的特征子集对训练集样本 B 投影,得到新的样本集 B',并且划分为多数类样本集 M' 与少数类样本集 S'.

② 分别将 M' 与 S' 平均分为 n 等份,得到 $M'_1, M'_2, \cdots, M'_n$ 与 $S'_1, S'_2, \cdots, S'_n$,分别合并 $S'_i, M'_i (i=1,2,\cdots,n)$,得到 n 个新的数据集 N:$N_1, \cdots, N_n$.

③ 对每一个 N_i 进行操作,将 N_i 作为测试集,N 中其他的数据集作为训练集,对训练集进行混合采样,使用 XGBoost 进行训练,对 N_i 进行测试,得到 N_i 的分类结果.

④ 将每一个 N_i 的结果合并,得到整个样本集 B' 的分类结果,然后计算出 F_1 值.

经过上述计算得到每个位串的 F_1 值,即个体适应值.

3) 迭代设置:设种群的最大迭代次数为 $g_{\max}$,且令 $g=1$;

4) 个体选择:所有个体按照适应值进行降序排序,选择前 a 个个体进入交配池形成父代,且每个个体进入交配池的数目按照下式计算.

$$l_i = \alpha(a+1-i)$$

l_i 表示第 i 个个体进入交配池的数目,α 为乘数因子. 进入交配池的个体总数目为 $\sum l_i$,它们作为父代来产生新的个体.

5) 交叉算子:设置交叉概率,依据交叉概率进行父代个体交叉操作,且保证交叉后得到的个体不在父代种群中;

6) 变异算子:为了使种群保证其多样性,设置了变异概率,依据变异概率对个体进行随机变异操作,变异即二进制位串中由"0"变为"1"或者由"1"变为"0",且保证变异后得到的个体不在父代种群中;

7) 父代种群生成新的子代,令 $g=g+1$,跳到步骤4)进行新一轮迭代,直到 $g=g_{\max}$.

迭代后得到的子代即为最优特征子集.

2.3 极端梯度算法

本实验通过比较最终使用的是基于迭代的集成方法——极端梯度算法(XGBoost)[21],它是梯度下降树的高效实现,相比之下,它生成决策树时考虑了树的复杂度,准确性更高,在满足相同训练效果的情况下,XGBoost 所需迭代次数更少,并且它可以多线程同时进行,运行速度更快.

XGBoost 在目标函数中加上了正则化项,如式(4)所示,损失函数不仅用到了一阶导数,还使用了二阶导数,如式(5)所示.

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4}$$

$$\text{where } \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2$$

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

$$g_i = \partial_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)}) \tag{5}$$

定义 $I_j = \{ i | q(x_i) = j \}$ 为叶节点 j 的实例集,然后以扩展 Ω 的方式将损失函数改写为

$$\tilde{L}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_j \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_j + \lambda \right) \omega_j^2 \right] + \gamma T \tag{6}$$

叶节点 j 的最优权值 ω_j^* 计算式为

$$\omega_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{7}$$

并且通过式(8)计算相应的最优值.

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{8}$$

式(9)可以作为评分函数来衡量一个树结构的质量,类似于评价决策树的结构分数,只针对更大范围的目标函数推导出来.

通常不可能枚举所有的决策树结构,因此采用

贪婪算法，从单个叶节点开始，迭代地向树添加分支，假设 I_L 和 I_R 分别是拆分后左右节点的实例集，令 $I=I_L\cup I_R$，则分割后的损失减少量如式(9)．在实际中，这个公式通常用于评估分割后的候选项．

$$L_{\text{split}}=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}g_i\right)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{\left(\sum_{i\in I_R}g_i\right)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{\left(\sum_{i\in I}g_i\right)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{9}$$

3　实验

3.1　数据集

笔者采用的数据集为 WEBSPAM-UK2006[22]，这是由垃圾网页检测挑战赛和对抗性信息检索 Web 讨论组于 2006 年收集的公开资料集．该数据集数据丰富，已被多个研究团体采纳．垃圾网页数和正常网页数的比值表明，数据集是不平衡的．在这些数据中，以四组形式提供了 277 个特征，分别是基于内容的特征、基于链接的特征、转换后的基于链接的特征和基于邻接图的特征．

基于内容的特性主要关注的是 web 页面的内容，包括页面上的字数、标题中的字数、平均字数、压缩率和页面的熵等特性．在该数据集中，提取了 96 个基于内容的特征．基于链接的特性主要关注的是 web 页面中的链接，包括诸如页面输出链接的数量、进入页面链接的数量以及到内部页面的输出链接数量与总输出链接的比例等特性．在这个数据集中，提取了 41 个基于链接的特征．转换后的基于链接的功能包括简单的数值转换和基于链接的功能的组合．在这些特征中，可以看出基于链接的特征的对数和它们之间的比值．在这个数据集中，提取到 138 个特征．基于邻接图的特征是学习 Stacked 链接图得到的，共提取到 2 个特征．

表 2　WEBSPAM-UK2006

数据集类别	标注主机数	垃圾网页数	正常网页数
训练数据集	6 492	632	5 860
测试数据集	1 937	613	1 324
合计	8 429	1 245	7 184

3.2　不同分类算法的比较

为了比较几种分类器对垃圾网页检测数据集分类效果的优劣，本实验中使用不同的分类器进行实验，分别有决策树、随机森林、Adaboost 和 XGBoost．此时的数据集是不平衡数据集，交叉验证是 5 折，迭代次数是 1 000．从表 3 中可以看出，不论是 F_1 值，准确度还是 A_{UC} 值，XGBoost 都比其他分类器的评估指标值高，效果更好，因此接下来的实验使用 XGBoost 作为分类器．

表 3　不同分类器对应的评估指标

分类器	F_1 值	A_{UC} 值	准确率
决策树	0.729 4	0.758 9	0.698 5
随机森林	0.701 3	0.762 1	0.682 5
Adaboost	0.750 1	0.791 8	0.723 8
XGBoost	0.774 0	0.806 4	0.745 0

3.3　基于遗传算法的垃圾网页检测

本实验采用基于随机混合采样和遗传算法的集成分类方法，其中分类器为 XGBoost，交叉验证为 5 折，迭代次数是 1 000．混合采样中，z 值取 2．遗传算法中，依次选用 3～20 个最优个体，多次实验发现最优个体数为 9，即最后得到的最优特征子集为 9 个．实验结果如表 4 所示，从结果中可以看出，基于随机混合采样与遗传算法的分类器分类效果最佳，准确率与 XGBoost 算法相比提高了 19.25%，说明了算法的有效性，通过随机采样与遗传迭代，能够获得平衡且特征数较少的数据集，从而提高分类器性能．表 5 为 2007 年垃圾网页挑战赛中优胜团队的结果，可以看出，在 F_1 值这个评价指标上笔者提出的算法是优于其他队伍的，但是在 A_{UC} 这个指标上较

表 4　不同分类器对应的评估指标

方法	F_1 值	A_{UC} 值	准确率
RUOS + XGBoost	0.907 5	0.862 6	0.875 1
GA + XGBoost	0.774 2	0.806 0	0.745 0
RUOS + GA + XGBoost	0.917 5	0.877 7	0.888 4

表 5　2007 年垃圾网页挑战赛中各优胜团队的分类结果

团队	F_1 值	A_{UC} 值
Benczur 等，匈牙利科学院	0.91	0.93
Filoche 等，法国电信	0.88	0.93
Geng 等，中国科学院	0.87	0.93
Abou-Assaleh 等，Genie Knows 公司	0.81	0.80
Fetterly 等，微软搜索实验室	0.79	—
Cormack，滑铁卢大学	0.67	0.96

低,然而 A_{UC} 值最高的 Cormack 团队的 F_1 值仅为 0.67,说明其分类效果并不好.

Singh S 等[23]提出了基于关联的特征选择和粒子群优化策略的分类器,其实验的数据集分别为基于内容、链接、内容和链接、完整内容和转换链接的特征集,并不是将所有的特征放在一起进行选择,其 F_1 值并不如 RUOS + GA + XGBoost 高. Scarselli F 等[24]提出了包含概率映射图自组织映射和图神经网络的图层叠架构技术. 由于数据集具有不平衡的特点,样本很容易被分到多数类样本中,准确率并不是评价分类器好坏的最优指标,而其 F_1 值明显低于 RUOS + GA + XGBoost,因此所提出的算法在 F_1 值这个评价指标上提升显著,而准确率却较低. 说明所提算法还有改善的空间,下一步将进一步研究特征工程,将特征工程与遗传算法相结合,提高分类器的准确率.

表6 与其他方法的比较

方法	F_1 值	A_{UC} 值	准确率
CFS-PSO-Adaboost(B + D)[23]	0.726	–	–
CFS-PSO-MLP Classifier(B + D)[23]	0.739	–	–
Autoassociator + GNN(1)[24]	0.417 3	0.807 0	0.910 4
FNN,PM-G + GNN(3) + GNN(1)[24]	0.632 4	0.936 2	0.929 4
RUOS + GA + XGBoost	0.917 5	0.877 7	0.888 4

4 结束语

垃圾网页检测是搜索引擎中的一大难题,具有数据不平衡、特征维度高的特点,为此提出一种基于随机混合采样和遗传算法的极端梯度集成分类算法. 实验结果表明,这种算法能够获得平衡且特征空间维度更低的数据集,建立更简单符合逻辑的模型,并减少了建立学习模型的时间,提高了分类性能,是一种较好的分类算法. 在未来的工作中,会进一步研究特征工程,将其与遗传算法结合进行迭代,提高算法准确率. 由于使用的是现有数据集,怎样将其投入实际应用,实现实时检测也是今后研究的目标. 此外,还可以考虑将该算法应用到其他有类似问题的领域中.

参考文献:

[1] Silverstein C, Marais H, Henzinger M, et al. Analysis of a very large web search engine query log[J]. ACM SIGIR Forum, 1999, 33(1): 6-12.

[2] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析[J]. 中文信息学报, 2007, 21(1): 109-114.
Yu Huijia, Liu Yiqun, Zhang Min, et al. Research in search engine user behavior based on log analysis[J]. Journal of Chinese Information Processing, 2007, 21(1): 109-114.

[3] Prieto V M, Álvarez M, Cacheda F. SAAD, a content based web spam analyzer and detector[J]. Journal of Systems and Software, 2013, 86(11): 2906-2918.

[4] Castillo C, Donato D, Gionis A, et al. Know your neighbors[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'07. New York: ACM Press, 2007: 423-430.

[5] Yu Mei, Zhang Jie, Wang Jianrong, et al. The research of spam web page detection method based on web page differentiation and concrete cluster centers[M]// Wireless Algorithms, Systems, and Applications. Cham: Springer International Publishing, 2018: 820-826.

[6] Whang J, Jung Y, Dhillon I, et al. Fast asynchronous anti-trust rank for web spam detection[C]// WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2). Los Angeles: [s. n.], 2018: 1-4.

[7] Oskuie M D, Razavi S N. A survey of web Spam detection techniques[J]. International Journal of Computer Applications Technology and Research, 2014, 3(3): 180-185.

[8] Goh K L, Singh A K. Comprehensive literature review on machine learning structures for web spam classification[J]. Procedia Computer Science, 2015, 70: 434-441.

[9] Lingala T, Saritha G. Towards evaluating web spam threats and countermeasures[J]. Intl J Innov Adv Comput Sci, 2018, 7(3): 71-80.

[10] Wan Jing, Liu Mufan, Yi Junkai, et al. Detecting spam webpages through topic and semantics analysis[C]// 2015 Global Summit on Computer & Information Technology (GSCIT). New York: IEEE Press, 2015: 1-7.

[11] Mamun M S I, Rathore M A, Lashkari A H, et al. Detecting malicious URLs using lexical analysis[M]// Network and System Security. Cham: Springer International Publishing, 2016: 467-482.

[12] Singh T, Kumari M, Mahajan S. Feature oriented fuzzy logic based web spam detection[J]. Journal of Information and Optimization Sciences, 2017, 38(6): 999-1015.

[13] Fdez-Glez J, Ruano-Ordas D, Méndez J R, et al. A dy-

namic model for integrating simple web spam classification techniques[J]. Expert Systems with Applications, 2015, 42(21): 7969-7978.

[14] Silva R M, Almeida T A, Yamakami A. Towards web spam filtering using a classifier based on the minimum description length principle[C]//2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). New York: IEEE Press, 2016: 470-475.

[15] Li Yuancheng, Nie Xiangqian, Huang Rong. Web spam classification method based on deep belief networks[J]. Expert Systems With Applications, 2018, 96: 261-270.

[16] Barandela R, Valdovinos R M, Sánchez J S, et al. The imbalanced training sample problem: under or over sampling? [M]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004: 806-814.

[17] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[18] Guo Haixiang, Li Yijing, Shang J, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert Systems With Applications, 2017, 73: 220-239.

[19] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[20] Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2000, 33(9): 1455-1465.

[21] Chen Tianqi, Guestrin C. XGBoost[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'16. New York: ACM Press, 2016: 785-794.

[22] Castillo C, Donato D, Becchetti L, et al. A reference collection for web spam[J]. ACM SIGIR Forum, 2006, 40(2): 11-24.

[23] Singh S, Singh A K. Web-spam features selection using CFS-PSO[J]. Procedia Computer Science, 2018, 125(125): 568-575.

[24] Scarselli F, Tsoi A C, Hagenbuchner M, et al. Solving graph data issues using a layered architecture approach with applications to web spam detection[J]. Neural Networks, 2013, 48: 78-90.