

引用格式:朱琪,林果园. 基于改进随机森林算法的钓鱼网站检测方法研究 [J]. 微电子学与计算机, 2019, 36(4): 43-46.
[ZHU Q, LIN G Y. Research on detection methods of phishing websites based on improved random forest algorithm [J]. Microelectronics & Computer, 2019, 36(4): 43-46.]

基于改进随机森林算法的钓鱼网站检测方法研究

朱 琪^{1,2}, 林果园^{1,2,3}

(1 中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116; 2 矿山数字化教育部工程研究中心, 江苏 徐州 221116; 3 南京大学 计算机软件新技术国家重点实验室, 江苏 南京 210023)

摘 要: 为了更准确快捷的对钓鱼网站进行识别, 提出了一种基于改进随机森林算法的钓鱼网站检测方法. 该方法挖掘钓鱼网页特征之间潜在的关联规则, 并对数据集进行分区, 以此区分特征数据的重要程度并计算权重以及数据选取的比例, 选取数据后对数据空间进行相应的集合化与剪辑以此优化森林的建立, 并根据建立的森林达到对钓鱼网站检测识别的目的. 最终实验说明, 该方法对钓鱼网站的检测识别具有很好的效果和效率.

关键词: 钓鱼检测; 关联规则; 特征分区; 数据空间

中图分类号: TP182

文献标识码: A

文章编号: 1000-7180(2019)04-0043-04

Research on Detection Methods of Phishing Websites Based on Improved Random Forest Algorithm

ZHU Qi^{1,2}, LIN Guo-yuan^{1,2,3}

(1 School of computer science and technology, China University of Mining and Technology, XuZhou 221116, China;
2 Mine Digitization Engineering Research Center of the Ministry of Education, XuZhou 221116, China;
3 State Key Laboratory for Novel Software Technology, Nanjing University, NanJing 210023, China)

Abstract: In order to improve the efficiency of phishing detection, a new algorithm was proposed to improve the traditional random forest algorithm. Potential association rules between web features are mined and used to partition the data set, in order to distinguish the features of different structures and calculate the weight of different data space to determine the scale of the selection. After selection of data, training data sets need to be aggregated and clipped to optimize the establishment of forests. Websites are trained and predicted using voting in decision forest. Experiments result shows that the new algorithm has obvious advantages in efficiency and effectiveness compared with the other two algorithm.

Key words: fishing detection; association rules; feature partition; data space

1 引言

网络钓鱼网站更新速度极快, 存在周期较短, 这对反钓鱼工作提出挑战. 如何有效的从大量网站数据中分类识别出钓鱼网站是当今面临的棘手的问题. 钓鱼网站的检测主要分为对网页的 URL

检测和网站页面检测两个方面. 随着钓鱼网站和其反检测的发展, 单纯对 URL 的检测片面性和误判率都过高. 相反, 在对钓鱼网站的检测过程中, 更具说服力的是基于网站页面的检测, 网站页面特征是钓鱼网站的核心, 对钓鱼网站具有绝对的代表性和说明性. 以网站特征为依据对网站进行

检测分类时,需要对特征数据进行深度挖掘^[1]和分析.由于特征具有结构多样性的特点,在分类检测中需要保证数据处理的全面性,在数据选取时采用具有随机性的算法更具优势.由于特征结构的不同,对特征进行分层次判别时便需采用多层次的分类模型.与此同时,特征的维度多样性对分类器的要求甚是苛刻.因此,在处理钓鱼网站特征结构不同,维度复杂的问题上,随机森林算法^[2]便脱颖而出.随机森林算法的随机性在钓鱼网站检测时可以保证选取特征数据的全面性且此算法不拘泥于特征结构的统一性.因此本文引入关联规则,挖掘钓鱼网站的通性,在建立决策森林时具有高效率与高准确度,增加了森林投票的准确性,加快钓鱼网站的检测与识别.最后在具体的实验分析中证明了改进后的算法的有效性.

2 随机森林算法简介

随机森林是由一系列决策树 $\{H(X, \theta_k), k = 1, 2, \dots, n\}$ 组成的集成分类器,其中 $\{\theta_k\}$ 是服从独立同分布的随机向量, n 表示随机森林中决策树的个数,在给定自变量 X 下,每个决策树分类器通过投票来决定最优的分类结果.随机森林单棵树构造时,每个节点的划分以 m 个属性的 Gini 指数^[3]变化作为衡量标准,根据 Gini 指数计算该数据的增益 Gain 来得到其最佳的分割方式.对于每个属性值都是不同数字时,其划分采用 Gini 指数选择节点划分值,生成二叉树.

随机森林算法的基本步骤如下:

(1)从原训练集中,采用 Bootstrap^[4]方法进行重采样,随机构建 k 个训练集 $X_1, X_2, \dots, X_k$.

(2)对于每个训练样本,通过如下过程生成不剪枝 CART 树:首先假设训练样本的属性个数为 $M(m \leq M)$,从 M 个属性中随机抽取 m 个属性作为当前节点的分裂属性集.根据 Gini 指数与增益值 Gain 从 m 个属性中选出最好的分裂方式对该节点进行分裂.每棵树最大限度地生长,不做任何剪裁.

(3)生成的 k 棵树组成随机森林,用此随机森林进行分类应用,结果按树分类器的投票数而定.

3 改进的随机森林算法

目前,随机森林算法利用多棵树对样本进行训练并预测.为了提高森林的投票决策的准确率,在独立的特征数据中引入关联规则^[5]即提取出钓

鱼网站的潜在通性,由此特征数据被分为独立特征与关联特征组两类且将特征空间一分为二.其次,随机森林算法采用等概率随机选取数据的方法,但是关联特征组对钓鱼检测具有更高的决定度,由此数据集权重不同.面对不同决定度的分区与特征类型,在选取分裂数据时需要计算各结构的权重与选取比例.最后,建立随机森林之前,需要对选中的特征组进行集合化与剪辑,释放因关联规则而绑定的特征组,不根据关联规则而选用特定指数建立起公平的森林,同时为建立森林节省空间,提高效率.

3.1 提取关联规则与数据分区

特征作为网站的判别依据,综合提取 URL 特征^[6]与网站页面特征进行钓鱼检测^[7].为了加快检测速度,提取钓鱼网站的潜在通性即特征的关联规则,运用 Apriori 算法,根据支持度与信任度先从数据集集中找出所有的高频项目组,再由这些高频项目组中产生关联规则及其支持度记为(WAg).得出的关联规则即是关联特征组(记为 A),不存在关联规则的特征数据便为独立特征(记为 F).此时特征空间 S 内存在两种类型数据:独立特征^[8]与关联特征组.由此根据数据类型对特征数据空间 S 进行划分,将所有关联特征组划分为关联特征分区记为 S_A ,将独立特征数据划分为独立特征分区记为 S_F .在 S_A 中每个关联特征组作为随机森林选择的独立单位.

3.2 权重度量与选取比例

在原始的随机森林算法中,采用随机等概率选择输入数据.不同分区的数据类型^[9]与决定度不同,等概率选择的方法不再适用,数据选择便根据权重计算选取比例.

(1)单个特征权重:网站特征具有结构多样,维度不同的性质.在对特征重要性度量时,必须对特征进行分类评定.所有特征的独立权重具体如公式(1)所示:

$$W_i^F = \frac{N_i}{f} \quad i = 1, 2, \dots, y \quad (1)$$

式中, W_i^F 代表第 i 个特征的独立权重; f 代表数据集样本个数; y 代表特征类型数; N_i 代表第 i 类特征存在的样本个数.

(2)分区权重:数据空间被划分为两个分区:关联特征分区(S_A)与独立特征分区(S_F),在改进的算法中,需要同层次先后在两个分区中按照一定的比例进行选取特征数据.分别将 S_A 与 S_F 的权重记为

Q_A 与 Q_F , 具体如公式(2), (3) 所示:

$$Q_A = \sum_{i=1}^z W_i^A = \sum_{i=1}^z \left(\frac{1}{2} \left(\sum_{j=0}^n W_j^F + W_i^{Ag} \right) \right);$$

$$i = 1, 2, \dots, z \quad (2)$$

式中, W_i^A 代表第 i 个关联特征组的权重, 其是由分组的特征权重与分组的关联权重 (W^{Ag}) 所得; W_j^F 代表第 j 个特征的独立权重; W_i^{Ag} 代表第 i 个关联特征组的支持度即在大数据样本中此关联分组在所有关联分组中的决定性占比; z 表示 S_A 中关联特征组的个数; n 表示关联特征组中特征的个数。

$$Q_F = \sum_{i=1}^x W_i^F \quad i = 1, 2, \dots, x \quad (3)$$

式中, W_i^F 代表第 i 个特征的独立权重; x 表示 S_F 中独立特征的个数。

(3) 数据选取: 在对特征选取的过程中, 根据 Q_A 与 Q_F 分别计算对 S_A 与 S_F 的选取个数: δ_A 与 δ_F , 具体如公式(4), (5) 所示, 构造出训练特征集。

$$\delta_A = \text{ROD}(C_A \frac{Q_A}{Q_A + Q_F}) \quad (4)$$

$$\delta_F = \text{ROD}(C_F \frac{Q_F}{Q_A + Q_F}) \quad (5)$$

$$\text{式中 } \text{ROD}(\lambda) = \begin{cases} [\lambda], & \lambda \leq \text{mid}; \\ [\lambda], & \lambda > \text{mid}. \end{cases}$$

$$(\text{mid} = \frac{[\lambda] + [\lambda]}{2})$$

式中, λ 代表权重; δ_A 与 δ_F 分别代表 S_A 与 S_F 选取的数据个数; C_A 代表 S_A 的关联特征组个数; C_F 代表 S_F 的独立特征的个数; Q_A 与 Q_F 分别代表 S_A 与 S_F 的权重。

3.3 数据集的集合化与建立森林

建树的数据集从两个分区选取包含两种不同数据类型的特征, 其中改进的算法对于 S_A 的样本选取的皆是成组的关联特征数据, 与 S_F 中选取的数据形成建树训练特征集合。建树时需要对整个数据集进行集合化, 释放因关联规则束缚的特征, 在分裂建立决策树时, 不再根据关联规则绑定特征而根据特定指数进行选取结点并分裂。建立决策森林具体有三个步骤:

步骤 1: 将从 S_A 中选取的数据集 $D_A \{A_1, A_2, \dots, A_{\delta_A}\}$ 解除关联规则绑定为 $D_A \{F_{a1}, F_{a2}, \dots, F_{an}\}$ 并与从 S_F 中选取的 $D_F \{F_{f1}, F_{f2}, \dots, F_{f\delta_F}\}$ 合并组成训练特征集 $D_T \{D_A, D_F\}$ 。

步骤 2: 对于训练特征集 D_T , 根据 Gini 指数与增益值 Gain 从属性中选出最好的分裂方式对该节点进行分裂, 每次选取一个特征直至训练特征集为

空, 每棵树都尽最大程度的生长, 并且没有剪枝过程。

步骤 3: 重复步骤 1、2, 直至建立随机森林结束, 最终根据投票情况来确定网站属性。

3.4 算法描述

运用改进的随机森林算法进行钓鱼网站检测包含特征处理, 数据集处理以及训练数据三大部分。其中以提取网页的特征与关联规则为基础, 将具有关联规则的特征视为一个数据整体并将原始数据集分区, 并在选取数据时作为独立单位, 与选取的不同分区数据一起建立树森林。本文的算法运用大数据中钓鱼网站特征通性, 加快森林建立与提高决策效率。

算法 1 钓鱼网站特征检测算法

begin

获取网页特征, 并计算特征权重

$X_i = (F_1, F_2, \dots, F_f), i = 1, 2, \dots, y;$

获取关联规则并记录特征是否归属关联规则的状态

$\text{Apriori}(F_i), i = 1, 2, \dots, f;$

运用关联规则对特征空间进行划分

If(满足归属标记){

$\text{SpaceA}(F_i, \text{link}(F_i));$

}else{ $\text{SpaceF}(F_i);$ }

获取特征组的关联权重并根据公式(2), (3) 计算分区权重

While(不满足数据集特征上限)

按分区选取特征并进行集合化

$\text{Space}(F_i) = \text{Random}(S_A, S_F, Q_A, Q_F);$

根据集合化后的训练数据建树

$\text{RandomForest}(\text{Space}(F_i));$

end while

结果显示

end

4 实验结果与分析

4.1 实验数据

实验中的钓鱼网站取自 PhishTank^[10] 中 2014 年 5 月 1 日—2018 年 5 月 1 日间 6000 个的钓鱼网站 URL, 利用 htmlparser 爬取实验用例中的 URL 地址的网页, 然后利用钓鱼网站特征提取算法提取 23 种特征项, 将特征项全部设置为布尔类型。

4.2 关联规则的提取

关联特征组中仅含有 8 种特征如表 1 所示, 经

实验,得到 5 组关联特征组如表 2 所示.

表 1 特征说明

F	特征
1	URL 长度大于 23
2	未含有正常 ICP 号
3	未含有正常 copyright
4	含有文本输入
5	标签关键词
6	外部链接数量过多
7	链接与网页不同域
8	注册时间

表 2 关联特征组

A	关联特征组	占有率	支持率
1	(F2,F3)	0.234	0.344
2	(F3,F5)	0.157	0.23
3	(F4,F7)	0.109	0.16
4	(F5,F6)	0.102	0.15
5	(F1,F8)	0.079	0.116

表 2 列出了 5 组关联特征组对钓鱼网站检测的占有率与支持率,占有率即在所有实验样本中关联规则的存在比例,而支持率为在所有关联规则样本中该关联规则的占比. 根据结果表明 ICP 号与 copyright 这对关联特征适用于将近三分分子一的钓鱼网站,并有将近五分之一的钓鱼网站同时含有不正常 ICP 号并且包含网站敏感关键词. 这些关联规则透露钓鱼网站检测的重要着手点可以在于判断其是否含有正常的 ICP 号、copyright 与文本输入;同时页面的关键词特征,链接异常情况与注册时间的长短对钓鱼网站的判别起到相对关键性的作用. 在疑似钓鱼网站检测中,对以上的特征进行标记运用可以提升检测速度与效率. 正因如此,分区选取特征数据更有利于对钓鱼网站数据的分类. 由此表 2 得出的关联规则在对特征数据空间划分上起到决定作用,也是改进算法的基础依据.

4.3 钓鱼检测与性能对比

运用改进的随机森林算法,原始的随机森林算法与决策树分别对多组钓鱼网站数据集进行分类检测,在数据样本增加变化的情况下探究两种算法对钓鱼网站检测的准确率,如图 1 所示:

图中列出了不同样本数量下三种算法分类准确率,在样本数量变化的情况下,总体而言,算法整体趋势均在上升,并在大数据集的情况下改进的算法正确率趋于 96%,而原始的随机森林算法正确率趋

于 93%,由此得出改进的算法比原始算法在正确率上得到了提升.

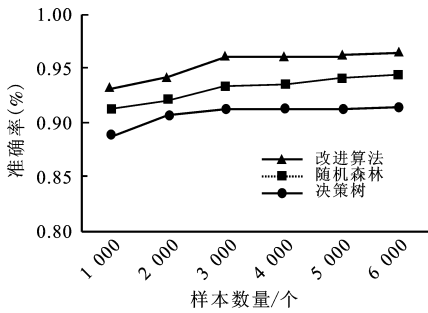


图 1 算法的准确率比较

为了验证算法的有效性和性能,将数据集平均分成 6 组,使用改进算法,决策树算法,随机森林算法分别对每组数据进行钓鱼网站检测,记录每组实验的准确率,误判率,漏判率,以及时间,并根据每种算法的检测结果计算平均值. 最终三种算法的性能比较如表 3 所示:

表 3 算法性能比较

	改进算法	随机森林	决策树
平均准确率	0.962	0.93	0.911
平均时间/秒	0.63	0.84	1.43
平均误判率	0.057	0.064	0.143
平均漏判率	0.044	0.047	0.036

表 3 详细的列出了三种算法的性能比较,根据实验结果,改进的算法在各方面都比原算法得到了提升,体现在执行时间的减少与对钓鱼检测正确率的提升,这样整体运算的效率提高了近 20%,准确率在原始随机森林算法的基础上提高了近 3%. 改进算法相对于原始算法的漏判率与误判率亦有所减少,这都主要基于对特征空间进行了科学划分与选取.

5 结束语

本文在传统随机森林算法的基础上,提出一种基于改进的随机森林算法的钓鱼网站检测方法. 该方法根据钓鱼网站特征潜在的关联规则,对特征空间分区并计算权重,以及对训练数据集进行剪辑,达到对钓鱼网站进行更准确地分类检测的目的. 经实验结果证实:关联规则的引入提高了钓鱼网站检测效率;关联组与数据选取方法的合理选择与运用对最终结果具有一定的影响. 综上所述改进的随机森林算法提高了钓鱼网站检测的效率与准确率.

5 结束语

学习和适应手写体数字的几何变换能使数字分类达到了更好的效果,使得识别精度更高.因此,本文通过引入形变卷积操作,对 VGG-16 网络进行改进,并使用 MNIST 数据集进行训练,从而获取能有效提取和分类手写体数字的形变卷积神经网络.通过在 MNIST 数据集上进行实验,并与现有的方法以及卷积神经网络进行比较.实验结果表明,基于形变卷积神经网络的手写体数字方法具有较高的识别精度以及收敛速度,也侧面证明了对几何变换的处理能有效的提高特征提取精度.同时,本文的网络结构具有普适性,能够很好的泛化到多任务中.

参考文献:

[1] 陈岩,李洋洋,余乐,等.基于卷积神经网络的手写体数字识别系统[J].微电子学与计算机,2018,35(2):71-74.

[2] KAREN SIONYAN, ANDREW ZISSERMAN. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, arXiv preprint arXiv:1409.1556.

[3] 赵朋成,冯玉田,涂云轩.基于高倍特征深度残差网络的手写数字识别[J].电子测量技术,2018,6(41):

86-89.

[4] YANG F, JIN L, YANG W, et al. Handwritten/Printed Receipt Classification Using Attention-Based Convolutional Neural Network [C]// International Conference on Frontiers in Handwriting Recognition. IEEE, 2017:384-389.

[5] DAI JIFENG, et al. Deformable Convolutional Networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:764-773.

[6] DENG J, DONG W, Socher R, et al. ImageNet: A large-scale hierarchical image database [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2009:248-255.

作者简介:

茹晓青 女,(1993-),硕士研究生.研究方向为图像处理、模式识别.

华国光 男,(1994-),硕士研究生.研究方向为计算机视觉、深度学习.

李丽宏(通讯作者) 女,(1974-),博士,副教授.研究方向为图像处理、计算机视觉、深度学习. E-mail:lilihgg@163.com.

李莉 女,(1984-),女,博士,讲师.研究方向为像处理、计算机视觉、模式识别.

(上接第 46 页)

参考文献:

[1] VERIKAS A, GELZINIS A, BACAUSKIENE M. Mining data with randomforests: a survey and results of new tests[J]. Pattern Recognition, 2011,44(2):330-349.

[2] AMARATUNGA D, CABRERA J, LEE Y. Enriched random forests[J]. Bioinformatics, 2008,24(18):2010-2014.

[3] MOORTHY K, MOHAMAD M S. Random forest for gene selection and microarray data classification[J]. Springer Berlin Heidelberg, 2012,7(3):142.

[4] 王燕,王兴芬,任俊玲.面向钓鱼网站敏感特征项选取的 IIGAIN 算法[J].计算机应用与软件,2016,33(4):297-301.

[5] HIPPI J, G NAKHAEIZADEH. Algorithms for association rule mining - a general survey and comparison [J]. Acm Sigkdd Explorations Newsletter, 2000,2(1):58-64.

[6] ZHANG W W, GONG J, LIU Q, et al. Lightweight

domain name detection algorithm based on morpheme features[J]. Ruan Jian XueBao/Journal of Software, 2016,27(9):2348-2364.

[7] 张茜,延志伟,李洪涛,等.网络钓鱼欺诈检测技术研究[J].网络与信息安全学报,2017,3(7):7-24.

[8] 王伟平,张兵.支持页面特征伪造识别的钓鱼网页检测方法[J].山东大学学报(理学版),2014,49(9):90-96.

[9] BERGHOLZ A, CHANG J H, PAASS G. Improved Phishing detection using model-based features [C]. Proceedings of the Nonference on Email and Anti-span(CEAS),2008.

[10] PhishTank. <http://www.phishtank.com>.

作者简介:

朱琪 男,(1994-),硕士研究生.研究方向为云计算与信息安全. E-mail:747116218@qq.com.

林果园 男,(1975-),博士,副教授.研究方向为网络空间安全、移动互联及其安全、云计算及其安全、信息系统及其安全.